Jay L. Devore
Kenneth N. Berk
Matthew A. Carlton

Modern Mathematical Statistics with Applications

Third Edition

Springer

Springer Texts in Statistics

Springer Texts in Statistics (STS) includes advanced textbooks from 3rd- to 4th-year undergraduate courses to 1st- to 2nd-year graduate courses. Exercise sets should be included. The series editors are currently Genevera I. Allen, Richard D. De Veaux, and Rebecca Nugent. Stephen Fienberg, George Casella, and Ingram Olkin were editors of the series for many years.

More information about this series at http://www.springer.com/series/417

Jay L. Devore · Kenneth N. Berk ·
Matthew A. Carlton

Modern Mathematical Statistics with Applications

Third Edition

Springer

Jay L. Devore
Department of Statistics (Emeritus)
California Polytechnic State University
San Luis Obispo, CA, USA

Kenneth N. Berk
Department of Mathematics (Emeritus)
Illinois State University
Normal, IL, USA

Matthew A. Carlton
Department of Statistics
California Polytechnic State University
San Luis Obispo, CA, USA

Preface

Purpose

Our objective is to provide a postcalculus introduction to the discipline of statistics that

- Has mathematical integrity and contains some underlying theory.
- Shows students a broad range of applications involving real data.
- Is up to date in its selection of topics.
- Illustrates the importance of statistical software.
- Is accessible to a wide audience, including mathematics and statistics majors (yes, there are quite a few of the latter these days, thanks to the proliferation of "big data"), prospective engineers and scientists, and those business and social science majors interested in the quantitative aspects of their disciplines.

A number of currently available mathematical statistics texts are heavily oriented toward a rigorous mathematical development of probability and statistics, with much emphasis on theorems, proofs, and derivations. The focus is more on mathematics than on statistical practice. Even when applied material is included, the scenarios are often contrived (many examples and exercises involving dice, coins, cards, widgets, or a comparison of treatment A to treatment B).

Our exposition is an attempt to provide a reasonable balance between mathematical rigor and statistical practice. We believe that showing students the applicability of statistics to real-world problems is extremely effective in inspiring them to pursue further coursework and even career opportunities in statistics. Opportunities for exposure to mathematical foundations will follow in due course. In our view, it is more important for students coming out of this course to be able to carry out and interpret the results of a two-sample t test or simple regression analysis, and appreciate how these are based on underlying theory, than to manipulate joint moment generating functions or discourse on various modes of convergence.

Content and Mathematical Level

The book certainly does include core material in probability (Chap. 2), random variables and their distributions (Chaps. 3–5), and sampling theory (Chap. 6). But our desire to balance theory with application/data analysis is reflected in the way the book starts out, with a chapter on descriptive and exploratory statistical techniques rather than an immediate foray into the axioms of probability and their consequences. After the distributional infrastructure is in place, the remaining statistical chapters cover the basics of inference. In addition to introducing core ideas from estimation and hypothesis testing (Chaps. 7–10), there is emphasis on checking assumptions and examining the data prior to formal analysis. Modern topics such as bootstrapping, permutation tests, residual analysis, and logistic regression are included. Our treatment of regression, analysis of variance, and categorical data analysis (Chaps. 11–13) is definitely more oriented to dealing with real data than with theoretical properties of models. We also show many examples of output from commonly used statistical software packages, something noticeably absent in most other books pitched at this audience and level.

The challenge for students at this level should lie with mastery of statistical concepts as well as with mathematical wizardry. Consequently, the mathematical prerequisites and demands are reasonably modest. Mathematical sophistication and quantitative reasoning ability are certainly important, especially as they facilitate dealing with symbolic notation and manipulation. Students with a solid grounding in univariate calculus and some exposure to multivariate calculus should feel comfortable with what we are asking of them. The few sections where matrix algebra appears (transformations in Chap. 5 and the matrix approach to regression in the last section of Chap. 12) can easily be deemphasized or skipped entirely. Proofs and derivations are included where appropriate, but we think it likely that obtaining a conceptual understanding of the statistical enterprise will be the major challenge for readers.

Recommended Coverage

There should be more than enough material in our book for a year-long course. Those wanting to emphasize some of the more theoretical aspects of the subject (e.g., moment generating functions, conditional expectation, transformations, order statistics, sufficiency) should plan to spend correspondingly less time on inferential methodology in the latter part of the book. We have opted not to mark certain sections as optional, preferring instead to rely on the experience and tastes of individual instructors in deciding what should be presented. We would also like to think that students could be asked to read an occasional subsection or even section on their own and then work exercises to demonstrate understanding, so that not everything would need to be presented in class. Remember that there is never enough time in a course of any duration to teach students all that we'd like them to know!

Revisions for This Edition

- Many of the examples have been updated and/or replaced, especially those containing real data or references to applications published in various journals. The same is true of the roughly 1300 exercises in the book.
- The exposition has been refined and polished throughout to improve accessibility and eliminate unnecessary material and verbiage. For example, the categorical data chapter (Chap. 13) has been streamlined by discarding some of the methodology involving tests when parameters must be estimated.
- A section on simulation has been added to each of the chapters on probability, discrete distributions, and continuous distributions.
- The material in the chapter on joint distributions (Chap. 5) has been reorganized. There is now a separate section on linear combinations and their properties, and also one on the bivariate normal distribution.
- The material in the chapter on statistics and their sampling distributions (Chap. 6) has also been reorganized. In particular, there is now a separate section on the chi-squared, t, and F distributions prior to the one containing derivations of sampling distributions of statistics based on a normal random sample.
- The chapters on one-sample confidence intervals (Chap. 8) and hypothesis tests (Chap. 9) place more emphasis on t procedures and less on large-sample z procedures. This is also true of inferences based on two samples in Chap. 10.
- Chap. 9 now contains a subsection on using the bootstrap to test hypotheses.
- The material on multiple regression models containing quadratic, interaction, and indicator variables has been separated into its own section. And there is now a separate expanded section on logistic regression.
- The nonparametric and Bayesian material that previously comprised a single chapter has been separated into two chapters, and material has been added to each. For example, there is now a section on nonparametric inferences about population quantiles.

Acknowledgements

A Final Thought

It is our hope that students completing a course taught from this book will feel as passionate about the subject of statistics as we still do after so many years in the profession. Only teachers can really appreciate how gratifying it is to hear from a student after he or she has completed a course that the experience had a positive impact and maybe even affected a career choice.

Los Osos, CA, USA	Jay L. Devore
Normal, IL, USA	Kenneth N. Berk
San Luis Obispo, CA, USA	Matthew A. Carlton

Contents

Introduction

Statistical concepts and methods are not only useful but indeed often indispensable in understanding the world around us. They provide ways of gaining new insights into the behavior of many phenomena that you will encounter in your chosen field of specialization.

The discipline of statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation. Without uncertainty or variation, there would be little need for statistical methods or statisticians. If the yield of a crop was the same in every field, if all individuals reacted the same way to a drug, if everyone gave the same response to an opinion survey, and so on, then a single observation would reveal all desired information.

Section 1.1 establishes some key statistics vocabulary and gives a broad overview of how statistical studies are conducted. The rest of this chapter is dedicated to graphical and numerical methods for summarizing data.

1.1 The Language of Statistics

We are constantly exposed to collections of facts, or **data**, both in our professional capacities and in everyday activities. The discipline of statistics provides methods for organizing and summarizing data and for drawing conclusions based on information contained in the data.

An investigation will typically focus on a well-defined collection of objects constituting a **population** of interest. In one study, the population might consist of all multivitamin capsules produced by a certain manufacturer in a particular week. Another investigation might involve the population of all individuals who received a B.S. in statistics or mathematics during the most recent academic year. When desired information is available for *all* objects in the population, we have what is called a **census**. Constraints on time, money, and other scarce resources usually make a census impractical or infeasible. Instead, a subset of the population—a **sample**—is selected in some prescribed manner. Thus we might obtain a sample of pills from a particular production run as a basis for investigating whether pills are conforming to manufacturing specifications, or we might select a sample of last year's graduates to obtain feedback about the quality of the curriculum.

We are usually interested only in certain characteristics of the objects in a population: the amount of vitamin C in the pill, the sex of a student, the age of a vehicle, and so on. A characteristic may be **categorical**, such as sex or college major, or it may be **quantitative** in nature. In the former case, the *value* of the characteristic is a category (e.g., female or economics), whereas in the latter case, the

value is a number (e.g., age = 5.1 years or vitamin C content = 65 mg). A **variable** is any characteristic whose value may change from one object to another in the population. We shall initially denote variables by lowercase letters from the end of our alphabet. Examples include

x = brand of computer owned by a student
y = number of items purchased by a customer at a grocery store
z = braking distance of an automobile under specified conditions

Data comes from making observations either on a single variable or simultaneously on two or more variables. A **univariate** data set consists of observations on a single variable. For example, we might consider the type of computer, laptop (L) or desktop (D), for ten recent purchases, resulting in the categorical data set

<div align="center">D   L   L   L   D   L   L   D   L   L</div>

The following sample of lifetimes (hours) of cell phone batteries under continuous use is a quantitative univariate data set:

<div align="center">10.6   10.1   11.2   9.0   10.8   9.5   8.8   11.5</div>

We have **bivariate** data when observations are made on each of two variables. Our data set might consist of a (height, weight) pair for each basketball player on a team, with the first observation as (72, 168), the second as (75, 212), and so on. If a kinesiologist determines the values of x = recuperation time from an injury and y = type of injury, the resulting data set is bivariate with one variable quantitative and the other categorical. **Multivariate** data arises when observations are made on more than two variables. For example, a research physician might determine the systolic blood pressure, diastolic blood pressure, and serum cholesterol level for each patient participating in a study. Each observation would be a triple of numbers, such as (120, 80, 146). In many multivariate data sets, some variables are quantitative and others are categorical. Thus the annual automobile issue of *Consumer Reports* gives values of such variables as type of vehicle (small, sporty, compact, midsize, large), city fuel efficiency (mpg), highway fuel efficiency (mpg), drive train type (rear wheel, front wheel, four wheel), and so on.

Branches of Statistics

An investigator who has collected data may wish simply to summarize and describe important features of the data. This entails using methods from **descriptive statistics**. Some of these methods are graphical in nature; the constructions of histograms, boxplots, and scatterplots are primary examples. Other descriptive methods involve calculation of numerical summary measures, such as means, standard deviations, and correlation coefficients. The wide availability of statistical computer software packages has made these tasks much easier to carry out than they used to be. Computers are much more efficient than human beings at calculation and the creation of pictures (once they have received appropriate instructions from the user!). This means that the investigator doesn't have to expend much effort on "grunt work" and will have more time to study the data and extract important messages. Throughout this book, we will present output from various packages such as R, SAS, and Minitab.

Example 1.1 Charity is a big business in the USA. The website charitynavigator.com gives information on roughly 5500 charitable organizations, and there are many smaller charities that fly below the navigator's radar. Some charities operate very efficiently, with fundraising and administrative expenses that are only a small percentage of total expenses, whereas others spend a high percentage of what they take in on such activities. Here is data on fundraising expenses, as a percentage of total expenditures, for a random sample of 60 charities:

6.1	12.6	34.7	1.6	18.8	2.2	3.0	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4.0	21.0	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12.0	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6.0	48.0	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12.0	4.7	14.7	6.4	17.0	2.5	16.2

Without any organization, it is difficult to get a sense of the data's most prominent features: what a typical (i.e., representative) value might be, whether values are highly concentrated about a typical value or quite dispersed, whether there are any gaps in the data, what fraction of the values are less than 20%, and so on. Figure 1.1 shows a *histogram*. In Section 1.2 we will discuss construction and interpretation of this graph. For the moment, we hope you see how it describes the way the percentages are distributed over the range of possible values from 0 to 100. Of the 60 charities, 36 use less than 10% on fundraising, and 18 use between 10% and 20%. Thus 54 out of the 60 charities in the sample, or 90%, spend less than 20% of money collected on fundraising. How much is too much? There is a delicate balance: most charities must spend money to raise money, but then money spent on fundraising is not available to help beneficiaries of the charity. Perhaps each individual giver should draw his or her own line in the sand.
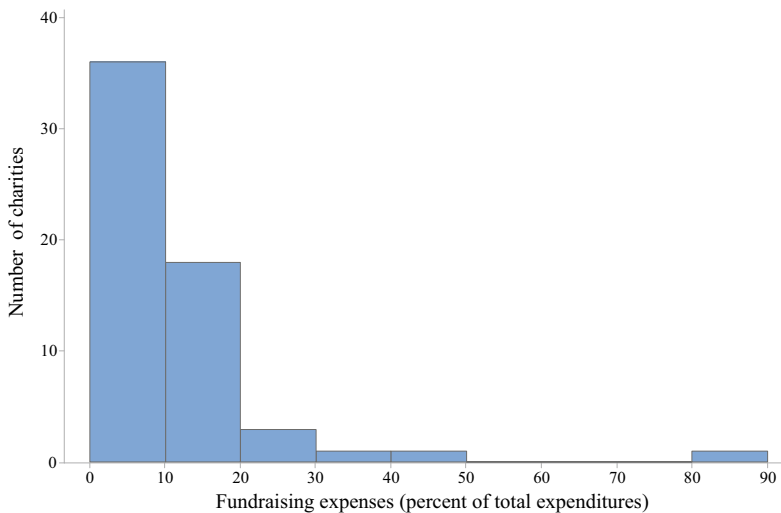
Figure 1.1 A histogram for the charity fundraising data of Example 1.1   ∎

Having obtained a sample from a population, an investigator would frequently like to use sample information to draw some type of conclusion (make an *inference* of some sort) about the population. That is, the sample is typically a means to an end rather than an end in itself. Techniques for generalizing from a sample to a population in a precise and objective way are gathered within the branch of our discipline called **inferential statistics**.

Example 1.2 The authors of the article "Fire Safety of Glued-Laminated Timber Beams in Bending" (*J. of Structural Engr.* 2017) conducted an experiment to test the fire resistance properties of wood pieces connected at corners by sawtooth-shaped "fingers" along with various types of commercial adhesive. The beams were all exposed to the same fire and load conditions. The accompanying data on fire resistance time (min) for a sample of timber beams bonded with polyurethane adhesive appeared in the article:

$$47.0 \quad 53.0 \quad 52.5 \quad 52.0 \quad 47.5 \quad 56.5 \quad 45.0 \quad 43.5 \quad 48.0 \quad 48.0$$
$$41.0 \quad 34.0 \quad 36.5 \quad 49.0 \quad 47.5 \quad 34.0 \quad 34.0 \quad 36.0 \quad 42.0$$

Suppose we want an *estimate* of the true average fire resistance time under these conditions. (Conceptualizing a population of all such beams with polyurethane bonding under these experimental conditions, we are trying to estimate the population mean.) It can be shown that, with a high degree of confidence, the population mean fire resistance time is between 41.2 and 48.0 min; this is called a *confidence interval* or an *interval estimate*. On the other hand, this data can also be used to predict the fire resistance time of a *single* timber beam under these conditions. With a high degree of certainty, the fire resistance time of a single such beam will exceed 29.4 min; the number 29.4 is called a *lower prediction bound*. ∎

Probability Versus Statistics

The main focus of this book is on presenting and illustrating methods of inferential statistics that are useful in research. The most important types of inferential procedures—point estimation, hypothesis testing, and estimation by confidence intervals—are introduced in Chapters 7–9 and then used in more complicated settings in Chapters 10–15. The remainder of this chapter presents methods from descriptive statistics that are most used in the development of inference.

Chapters 2–6 present material from the discipline of probability. This material ultimately forms a bridge between the descriptive and inferential techniques. Mastery of probability leads to a better understanding of how inferential procedures are developed and used, how statistical conclusions can be translated into everyday language and interpreted, and when and where pitfalls can occur in applying the methods. Probability and statistics both deal with questions involving populations and samples, but do so in an "inverse manner" to each other.

In probability, properties of the population under study are assumed known (e.g., in a numerical population, some specified distribution of the population values may be assumed), and questions regarding a sample taken from the population are posed and answered. In statistics, characteristics of a sample are available to the experimenter, and this information enables the experimenter to draw conclusions about the population. The relationship between the two disciplines can be summarized by saying that probability reasons from the population to the sample (deductive reasoning), whereas inferential statistics reasons from the sample to the population (inductive reasoning). This is illustrated in Figure 1.2.
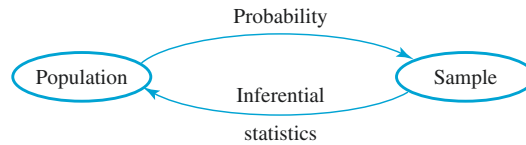
Figure 1.2   The relationship between probability and inferential statistics

Before we can understand what a particular sample can tell us about the population, we should first understand the uncertainty associated with taking a sample from a given population. This is why we study probability before statistics.

As an example of the contrasting focus of probability and inferential statistics, consider drivers' use of seatbelts in automobiles. According to the article "Somehow, Way Too Many Americans Still Aren't Wearing Seatbelts" (www.wired.com, Sept. 2016), data collected by observers from the National Highway Traffic Safety Administration indicates that 88.5% of drivers and front seat passengers buckle up. But this percentage varies considerably by location. In the 34 states in which a driver can be pulled over and cited for nonusage, 91.2% wore their seatbelts in 2015. By contrast, in the 15 states where a citation can be given only if a driver is pulled over for another infraction and the one state where usage is not mandatory (New Hampshire), usage drops to 78.6%.

In a probability context, we might assume that 85% of all drivers in a particular metropolitan area regularly use seatbelts (an assumption about the population) and then ask, "How likely is it that a sample of 100 drivers will include at most 70 who regularly use their seatbelt?" or "How many drivers in a sample of size 100 can we expect to regularly use their seatbelt?" On the other hand, in inferential statistics, sample information is available, e.g., a sample of 100 drivers from this area reveals that 80 regularly use their seatbelts. We might then ask, "Does this provide strong evidence for concluding that less than 90% of all drivers in this area are regular seatbelt users?" In this latter scenario, sample information will be employed to answer a question about the structure of the entire population from which the sample was selected.

Next, consider a study involving a sample of 25 patients to investigate the efficacy of a new minimally invasive method for rotator cuff surgery. The amount of time that each individual subsequently spends in physical therapy is then determined. The resulting sample of 25 PT times is from a population that does not actually exist. Instead it is convenient to think of the population as consisting of all possible times that might be observed under similar experimental conditions. Such a population is referred to as a *conceptual* or *hypothetical population*. There are a number of situations in which we fit questions into the framework of inferential statistics by conceptualizing a population.

Collecting Data

Statistics deals not only with the organization and analysis of data once it has been collected but also with the development of techniques for collecting that data. If data is not properly collected, an investigator might not be able to answer the questions under consideration with a reasonable degree of confidence. One common problem is that the target population—the one about which conclusions are to be drawn—may be different from the population actually sampled. In that case, an investigator must be very cautious about generalizing from the circumstances under which data has been gathered.

For example, advertisers would like various kinds of information about the television-viewing habits of potential customers. The most systematic information of this sort comes from placing monitoring devices in a small number of homes across the USA. It has been conjectured that placement of such devices in and of itself alters viewing behavior, so that characteristics of the sample may be different from those of the target population. As another example, a sample of five engines

with a new design may be experimentally manufactured and tested to investigate efficiency. These five could be viewed as a sample from the conceptual population of all prototypes that could be manufactured under similar conditions, but *not* necessarily as representative of all units manufactured once regular production gets under way. Methods for using sample information to draw conclusions about future production units may be problematic. Similarly, a new drug may be tried on patients who arrive at a clinic (i.e., a voluntary sample), but there may be some question about how typical these patients are. They may not be representative of patients elsewhere or patients at the same clinic next year.

When data collection entails selecting individuals or objects from a list, the simplest method for ensuring a representative selection is to take a **simple random sample**. This is one for which any particular subset of the specified size (e.g., a sample of size 100) has the same chance of being selected. For example, if the list consists of 1,000,000 serial numbers, the numbers 1, 2, …, up to 1,000,000 could be placed on identical slips of paper. After placing these slips in a box and thoroughly mixing, slips could be drawn one by one until the requisite sample size has been obtained. Alternatively (and much to be preferred), a computer's random number generator could be employed to generate 100 distinct numbers between 1 and 1,000,000.

Sometimes alternative sampling methods can be used to make the selection process easier, to obtain extra information, or to increase the degree of precision in conclusions. One such method, *stratified sampling*, entails separating the population units into nonoverlapping groups and taking a sample from each one. For example, a cell phone manufacturer might want information about customer satisfaction for units produced during the previous year. If three different models were manufactured and sold, a separate sample could be selected from each of the three corresponding strata. This would result in information on all three models and ensure that no one model was over-or underrepresented in the entire sample.

Frequently a "convenience" sample is obtained by selecting individuals or objects without systematic randomization. As an example, a collection of bricks may be stacked in such a way that it is extremely difficult for those in the center to be selected. If the bricks on the top and sides of the stack were somehow different from the others, the resulting sample data would not be representative of the population. Often an investigator will assume that such a convenience sample approximates a random sample, in which case a statistician's repertoire of inferential methods can be used; however, this is a judgment call.

Researchers may also collect data by carrying out some sort of designed experiment. This may involve deciding how to allocate several different treatments (such as fertilizers or drugs) to various experimental units (plots of land or patients). Alternatively, an investigator may systematically vary the levels or categories of certain factors (e.g., amount of fertilizer or dose of a drug) and observe the effect on some response (such as corn yield or blood pressure).

Example 1.3 Neonicotinoid insecticides (NNIs) are popular in agricultural use, especially for growing corn, but scientists are increasingly concerned about their effects on bee populations. An article in *Science* (June 30, 2017) described the results of a two-year study in which scientists randomly assigned some bee colonies to be exposed to "field-realistic" levels and durations of NNIs, while other colonies did not have NNI exposure. The researchers found that bees in the colonies exposed to NNIs had a 23% reduced life span, on average, compared to those in nonexposed colonies. One possible explanation for this result is chance variation—i.e., that NNIs really don't affect bee colony health and the observed difference is just "random noise," in the same way that tossing two identical coins 10 times each will usually produce different numbers of heads. However, in this case, inferential methods discussed in this textbook (and in the original article) suggest that chance

variation by itself cannot adequately explain the magnitude of the observed difference, indicating that NNIs may very well be responsible for the reduced average life span.   ∎

Exercises: Section 1.1 (1–13)

1. Give one possible sample of size 4 from each of the following populations:

   a. All daily newspapers published in the USA

   b. All companies listed on the New York Stock Exchange

   c. All students at your college or university

   d. All grade point averages of students at your college or university

2. For each of the following hypothetical populations, give a plausible sample of size 4:

   a. All distances that might result when you throw a football

   b. Page lengths of books published 5 years from now

   c. All possible earthquake strength measurements (Richter scale) that might be recorded in California during the next year

   d. All possible yields (in grams) from a certain chemical reaction carried out in a laboratory

3. Consider the population consisting of all cell phones of a certain brand and model, and focus on whether a cell phone needs service while under warranty.

   a. Pose several probability questions based on selecting a sample of 100 such cell phones.

   b. What inferential statistics question might be answered by determining the number of such cell phones in a sample of size 100 that need warranty service?

4. Give three different examples of concrete populations and three different examples of hypothetical populations. For one each of your concrete and hypothetical populations, give an example of a probability question and an example of an inferential statistics question.

5. The authors of the article "From Dark to Light: Skin Color and Wages among African Americans" (*J. of Human Resources* 2007: 701–738) investigated the association between darkness of skin and hourly wages. For a sample of 948 African Americans, skin color was classified as dark black, medium black, light black, or white.

   a. What variables were recorded for each member of the sample?

   b. Classify each of these variables as quantitative or categorical.

6. *Consumer Reports* compared the actual polyunsaturated fat percentages for different brands of "low-fat" margarine. Twenty-six containers of margarine were purchased; for each one, the brand was noted and the percent of polyunsaturated fat was determined.

   a. What variables were recorded for each margarine container in the sample?

   b. Classify each of these variables as quantitative or categorical.

   c. Give some examples of inferential statistics questions that *Consumer Reports* might try to answer with the data from these 26 margarine containers.

   d. "The average polyunsaturated fat content for the five Parkay margarine containers in the sample was 12.8%." Is the preceding sentence an example of descriptive statistics or inferential statistics?

7. The article "Is There a Market for Functional Wines? Consumer Preferences and Willingness to Pay for Resveratrol-Enriched Red Wine" (*Food Quality and Preference* 2008: 360–371) included the following information for a variety of Spanish wines:

   a. Region of origin

   b. Price of the wine, in euros

   c. Style of wine (young or crianza)

d. Production method (conventional or organic)

e. Type of grapes used (regular or resveratrol-enhanced)

Classify each of these variables as quantitative or categorical.

8. The authors of the article cited in the previous exercise surveyed 300 wine consumers, each of whom tasted two different wines. For each individual in the study, the following information was recorded:

a. Gender

b. Age, in years

c. Monthly income, in euros

d. Educational level (primary, secondary, or university)

e. Willingness to pay (WTP) for the first wine tasted, in euros

f. WTP for the second wine tasted, in euros.

(WTP is a very common measure for consumer products. Researchers ask, "How much would you be willing to pay for this item?") Classify each of the variables (a)–(f) as quantitative or categorical.

9. Many universities and colleges have instituted supplemental instruction (SI) programs, in which a student facilitator meets regularly with a small group of students enrolled in the course to promote discussion of course material and enhance subject mastery. Suppose that students in a large statistics course (what else?) are randomly divided into a control group that will not participate in SI and a treatment group that will participate. At the end of the term, each student's total score in the course is determined.

a. Are the scores from the SI group a sample from an existing population? If so, what is it? If not, what is the relevant conceptual population?

b. What do you think is the advantage of randomly dividing the students into the two groups rather than letting each student choose which group to join?

c. Why didn't the investigators put all students in the treatment group?

10. The California State University (CSU) system consists of 23 campuses, from San Diego State in the south to Humboldt State near the Oregon border. A CSU administrator wishes to make an inference about the average distance between the hometowns of students and their campuses. Describe and discuss several different sampling methods that might be employed.

11. A certain city divides naturally into ten district neighborhoods. A real estate appraiser would like to develop an equation to predict appraised value from characteristics such as age, size, number of bathrooms, distance to the nearest school, and so on. How might she select a sample of single-family homes that could be used as a basis for this analysis?

12. The amount of flow through a solenoid valve in an automobile's pollution control system is an important characteristic. An experiment was carried out to study how flow rate depended on three factors: armature length, spring load, and bobbin depth. Two different levels (low and high) of each factor were chosen, and a single observation on flow was made for each combination of levels.

a. The resulting data set consisted of how many observations?

b. Does this study involve sampling an existing population or a conceptual population?

13. In a famous experiment carried out in 1882, Michelson and Newcomb obtained 66 observations on the time it took for light to travel between two locations in Washington, D.C. A few of the measurements (coded in a certain manner) were 31, 23, 32, 36, 22, 26, 27, and 31.

a. Why are these measurements not identical?

b. Does this study involve sampling an existing population or a conceptual population?

1.2 Graphical Methods in Descriptive Statistics

There are two general types of methods within descriptive statistics: graphical and numerical summaries. In this section we will discuss the first of these types—representing a data set using visual techniques. In Sections 1.3 and 1.4, we will develop some numerical summary measures for data sets. Many visual techniques may already be familiar to you: frequency tables, histograms, pie charts, bar graphs, scatterplots, and the like. Here we focus on a selected few of these techniques that are most useful and relevant to probability and inferential statistics.

Notation

Some general notation will make subsequent discussions easier. The number of observations in a single sample, that is, the **sample size**, will often be denoted by n. So $n = 4$ for the sample of universities {Stanford, Iowa State, Wyoming, Rochester} and also for the sample of pH measurements {6.3, 6.2, 5.9, 6.5}. If two samples are simultaneously under consideration, either m and n or n_1 and n_2 can be used to denote the numbers of observations. Thus if {3.75, 2.60, 3.20, 3.79} and {2.75, 1.20, 2.45} are GPAs for two sets of friends, respectively, then $m = 4$ and $n = 3$.

Given a data set consisting of n observations on some variable x, the individual observations will be denoted by x_1, x_2, x_3, …, x_n. The subscript bears no relation to the magnitude of a particular observation. Thus x_1 will not in general be the smallest observation in the set, nor will x_n typically be the largest. In many applications, x_1 will be the first observation gathered by the experimenter, x_2 the second, and so on. The ith observation in the data set will be denoted by x_i.

Stem-and-Leaf Displays

Consider a numerical data set x_1, x_2, …, x_n for which each x_i consists of at least two digits. A quick way to obtain an informative visual representation of the data set is to construct a *stem-and-leaf display*, or *stem plot*.

Steps for constructing a stem-and-leaf display
1. Select one or more leading digits for the *stem* values. The trailing digits become the *leaves*.
2. List possible stem values in a vertical column.
3. Record the leaf for every observation beside the corresponding stem value.
4. Order the leaves from smallest to largest on each line.
5. Indicate the units for stems and leaves someplace in the display.

If the data set consists of exam scores, each between 0 and 100, the score of 83 would have a stem of 8 and a leaf of 3. For a data set of automobile fuel efficiencies (mpg), all between 8.1 and 47.8, we could use the tens digit as the stem, so 32.6 would then have a leaf of 2.6. Usually, a display based on between 5 and 20 stems is appropriate.

For a simple example, assume a sample of seven test scores: 93, 84, 86, 78, 95, 81, 72. Then the first-pass stem plot would be

```
7|82
8|461
9|35
```

With the leaves ordered this becomes

```
7|28          Stem: tens digit
8|146         Leaf: ones digit
9|35
```

Occasionally stems will be repeated to spread out the stem-and-leaf display. For instance, if the preceding test scores included dozens of values in the 70s, we could repeat the stem 7 twice, using 7L for scores in the low 70s (leaves 0, 1, 2, 3, 4) and 7H for scores in the high 70s (leaves 5, 6, 7, 8, 9).

Example 1.4 Job prospects for students majoring in an engineering discipline continue to be very robust. How much can a new engineering graduate expect to earn? Here are the starting salaries for a sample of 38 civil engineers from one author's home institution (Spring 2016), courtesy of the university's Graduate Status Report:

58,000	62,000	56,160	67,000	66,560	58,240	60,000	61,000	70,000	61,000
65,000	60,000	61,000	80,000	62,500	75,000	60,000	68,000	57,600	65,000
55,000	63,000	60,000	70,000	68,640	72,000	83,000	50,128	56,000	63,000
55,000	52,000	70,000	80,000	60,320	65,000	70,000	65,000		

Figure 1.3 shows a stem-and-leaf display of these 38 starting salaries. Hundreds places and lower have been truncated; for instance, the lowest salary in the sample was $50,128, which is represented by 5|0 in the first row.

```
5L |  02
5H |  5566788                    Stem: $10,000
6L |  000001112233               Leaf: $1000
6H |  55556788
7L |  00002
7H |  5
8L |  003
```

Figure 1.3 Stem-and-leaf display for starting salaries of civil engineering graduates

Typical starting salaries were in the $60,000–$65,000 range, with most graduates starting between $55,000 and $70,000. A lucky (and/or exceptionally talented!) handful of students earned $80,000 or more upon graduation. ∎

Most graphical displays of quantitative data, including the stem-and-leaf display, convey information about the following aspects of the data:

- Identification of a typical or representative value
- Extent of spread about the typical value
- Presence of any gaps in the data
- Extent of symmetry in the distribution of values
- Number and location of peaks
- Presence of any outlying values (i.e. unusually small or large)

Example 1.5 Figure 1.4 presents stem-and-leaf displays for a random sample of lengths of golf courses (yards) that have been designated by *Golf Magazine* as among the most challenging in the USA. Among the sample of 40 courses, the shortest is 6433 yards long, and the longest is 7280 yards. The lengths appear to be distributed in a roughly uniform fashion over the range of values in the sample. Notice that a stem choice here of either a single digit (6 or 7) or three digits (643, …, 728) would yield an uninformative display, the first because of too few stems and the latter because of too many.

a

```
64| 33 35 64 70        Stem: Thousands and hundreds digits
65| 06 26 27 83        Leaf: Tens and ones digits
66| 05 14 94
67| 00 13 45 70 70 90 98
68| 50 70 73 90
69| 00 04 27 36
70| 05 11 22 40 50 51
71| 05 13 31 65 68 69
72| 09 80
```

b

```
Stem-and-leaf of yardage N = 40
Leaf Unit = 10
64  3367
65  0228
66  019
67  0147799
68  5779
69  0023
70  012455
71  013666
72  08
```
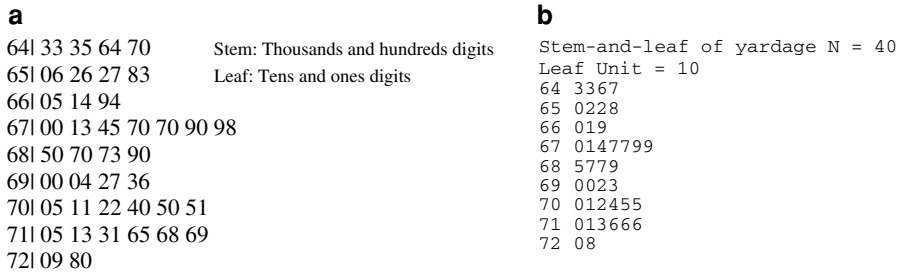
Figure 1.4 Stem-and-leaf displays of golf course yardages: (**a**) two-digit leaves; (**b**) display from Minitab with truncated one-digit leaves   ■

Dotplots

A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically. As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

Example 1.6 For decades, *The Economist* has used its "Big Mac index," defined for any country as the average cost of a McDonald's Big Mac, as a humorous way to compare product costs across nations and also examine the valuation of the US dollar worldwide. Here are values of the Big Mac index, converted to US dollars, for 56 countries reported by *The Economist* on January 1, 2019 (listed in alphabetical order by country name):

2.00	4.35	2.33	3.18	4.55	4.07	5.08	3.89	3.05	3.73
3.77	3.24	3.81	4.60	2.23	4.64	3.23	3.49	2.55	3.03
2.55	2.34	4.58	3.60	2.75	3.46	4.31	2.20	2.54	2.32
4.19	3.18	5.86	2.73	3.31	3.14	2.67	2.80	3.30	2.29
1.65	3.20	4.28	2.24	4.02	3.18	5.84	6.62	2.24	3.72
2.00	1.94	3.81	5.58	4.31	2.80				

Figure 1.5 shows a dotplot of these values. We can see that the average cost of a Big Mac in the USA, $5.58, is higher than in all but three countries (Sweden, Norway, and Switzerland). A typical Big Mac index value is around $3.20, but those values vary substantially across the globe. The distribution extends farther to the right of that typical value than to the left, due to a handful of comparatively large Big Mac prices in the USA and a few other countries.
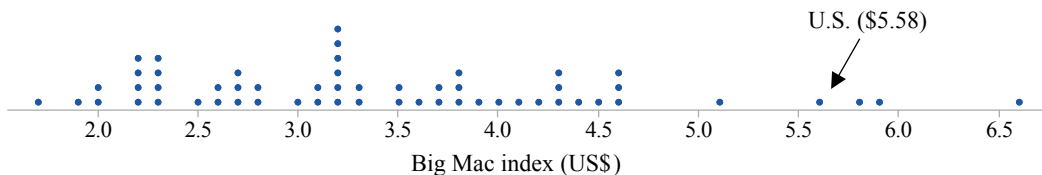
Figure 1.5 A dotplot of the data from Example 1.6

How can the Big Mac index be used to assess the strength of the US dollar? As an example, a Big Mac cost £3.19 in Britain, which converts to $4.07 using the exchange rate at the time the data was collected. Since that same amount of British currency *ought to* buy $5.58 worth of American goods according to the Big Mac index, it appears that the British pound was substantially undervalued at the time. ∎

Histograms

While stem-and-leaf displays and dotplots are useful for smaller data sets, histograms are well-suited to larger samples or the results of a census.

Consider first data resulting from observations on a "counting variable" x, such as the number of traffic citations a person received during the last year, or the number of people arriving for service during a particular period. The **frequency** of any particular x value is simply the number of times that value occurs in the data set. The **relative frequency** of a value is the fraction or proportion of times the value occurs:

$$\text{relative frequency of a value} = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}}$$

Suppose, for example, that our data set consists of 200 observations on x = the number of major defects in a new car of a certain type. If 70 of these x values are 1, then the frequency of the value 1 is (obviously) 70, while the relative frequency of the value 1 is 70/200 = .35. Multiplying a relative frequency by 100 gives a percentage; in the defect example, 35% of the cars in the sample had just one major defect. The relative frequencies, or percentages, are usually of more interest than the frequencies themselves. In theory, the relative frequencies should sum to 1, but in practice the sum may differ slightly from 1 because of rounding. A **frequency distribution** is a tabulation of the frequencies and/or relative frequencies.

Example 1.7 How unusual is a no-hitter or a one-hitter in a major league baseball game, and how frequently does a team get more than 10, 15, or 20 hits? Table 1.1 is a frequency distribution for the number of hits *per team* per game for all games in the 2016 regular season, courtesy of the website www.retrosheet.org.

Table 1.1 Frequency distribution for hits per team in 2016 MLB games

Hits/Team/Game	Number of games	Relative frequency	Hits/Team/Game	Number of games	Relative frequency
0	1	.0002	12	297	.0612
1	19	.0039	13	202	.0416
2	56	.0115	14	176	.0362
3	151	.0311	15	102	.0210
4	279	.0575	16	71	.0146
5	380	.0783	17	56	.0115
6	471	.0970	18	32	.0066
7	554	.1141	19	22	.0045
8	564	.1161	20	3	.0006
9	556	.1145	21	4	.0008
10	469	.0966	22	5	.0010
11	386	.0795		4,856	.9999

The corresponding histogram in Figure 1.6 rises rather smoothly to a single peak and then declines. The histogram extends a bit more to the right (toward large values) than it does on the left—a slight "positive skew."
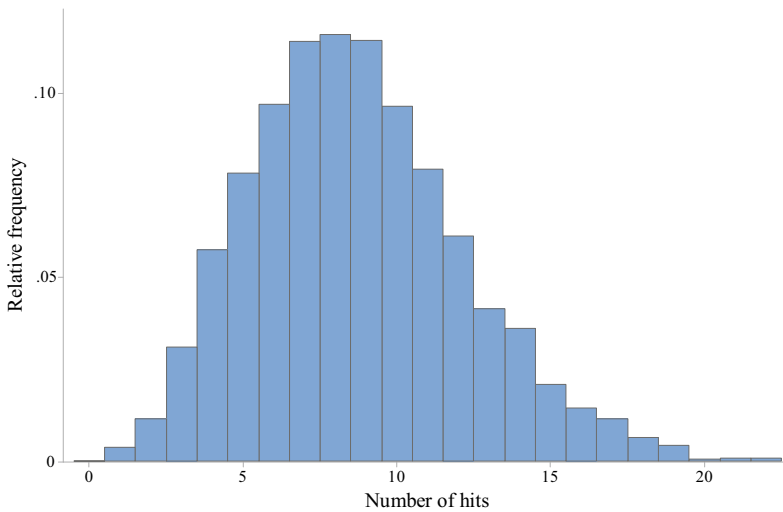
Figure 1.6 Relative frequency histogram of x = number of hits per team per game for the 2016 MLB season

Either from the tabulated information or from the histogram, we can determine the following:

$$\begin{array}{ccccccc} \text{proportion of} & & \text{relative} & & \text{relative} & & \text{relative} \\ \text{instances of at} & = & \text{frequency} & + & \text{frequency} & + & \text{frequency} \\ \text{most two hits} & & \text{for } x = 0 & & \text{for } x = 1 & & \text{for } x = 2 \end{array}$$

$$= \quad .0002 + .0039 + .0115 = .0156$$

Similarly,

$$\text{proportion of instances of between 5 and 10 hits (inclusive)} = .0783 + .0970 + \cdots + .0966 = .6166$$

That is, roughly 62% of the time that season, a team had between 5 and 10 hits (inclusive) in a game.

Incidentally, the only no-hitter of the season (notice the frequency of 1 for $x = 0$) came on April 21, 2016, with Jake Arrieta pitching the complete game for the Chicago Cubs against the Cincinnati Reds. ∎

Constructing a histogram for measurement data (e.g., weights of individuals, reaction times to a particular stimulus) requires subdividing the measurement axis into a suitable number of **class intervals** or **classes**, such that each observation is contained in exactly one class. Suppose, for example, that we have 50 observations on $x =$ fuel efficiency of an automobile (mpg), the smallest of which is 27.8 and the largest of which is 31.4. Then we could use the class boundaries 27.5, 28.0, 28.5, …, and 31.5 as shown here:
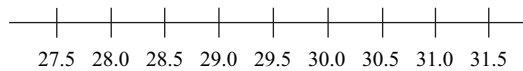
When all class widths are equal, a histogram is constructed as follows: first, mark the class boundaries on a horizontal axis like the one above. Then, above each interval, draw a rectangle whose height is the corresponding relative frequency (or frequency).

One potential difficulty is that occasionally an observation falls on a class boundary and therefore does not lie in exactly one interval, for example, 29.0. We will use the convention that any observation falling on a class boundary will be included in the class *to the right of the observation*. Thus 29.0 would go in the 29.0–29.5 class rather than the 28.5–29.0 class. This is how Minitab constructs a histogram; in contrast, the default histogram in R does it the other way, with 29.0 going into the 28.5–29.0 class.

Example 1.8 Power companies need information about customer usage to obtain accurate forecasts of demands. Investigators from Wisconsin Power and Light determined energy consumption (BTUs) during a particular period for a sample of 90 gas-heated homes. For each home, an adjusted consumption value was calculated to account for weather and house size. This resulted in the accompanying data (part of the stored data set furnace.mtw available in Minitab), which we have ordered from smallest to largest.

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

We let Minitab select the class intervals. The most striking feature of the histogram in Figure 1.7 is its resemblance to a bell-shaped (and therefore symmetric) curve, with the point of symmetry roughly at 10.
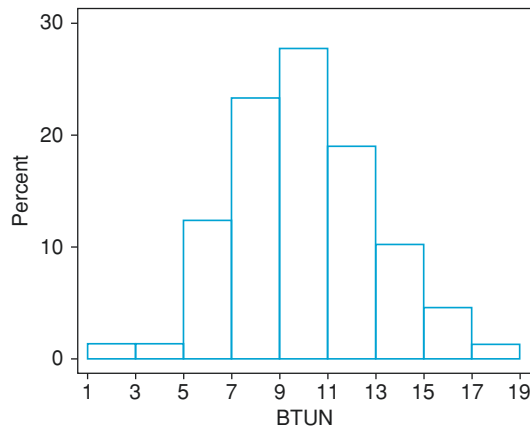
Figure 1.7 Minitab histogram of the energy consumption data from Example 1.8

Class	1–<3	3–<5	5–<7	7–<9	9–<11	11–<13	13–<15	15–<17	17–<19
Frequency	1	1	11	21	25	17	9	4	1
Relative frequency	.011	.011	.122	.233	.278	.189	.100	.044	.011

From the histogram,

$$\begin{array}{l}\text{proportion of} \\ \text{observations} \\ \text{less than } 9\end{array} \approx .01 + .01 + .12 + .23 = .37 \quad \left(\text{exact value} = \tfrac{34}{90} = .378\right)$$

The relative frequency for the 9–11 class is about .27, so we estimate that roughly half of this, or .135, is between 9 and 10. Thus

$$\begin{array}{l}\text{proportion of observations} \\ \text{less than } 10\end{array} \approx .37 + .135 = .505 \,(\text{slightly more than } 50\%)$$

The exact value of this proportion is 47/90 = .522. ∎

There are no hard-and-fast rules concerning either the number of classes or the choice of classes themselves. Between 5 and 20 classes will be satisfactory for most data sets. Generally, the larger the number of observations in a data set, the more classes should be used. A reasonable rule is

$$\text{number of classes} \approx \sqrt{\text{number of observations}}$$

Equal-width classes may not be a sensible choice if a data set "stretches out" to one side or the other. Figure 1.8 shows a dotplot of such a data set. If a large number of short, equal width classes are used,

many classes will have zero frequency. Using a small number of wide equal width classes results in almost all observations falling in just one or two of the classes. A sound choice is to use a few wider intervals near extreme observations and narrower intervals in the region of high concentration. In such situations a *density histogram* must be used.
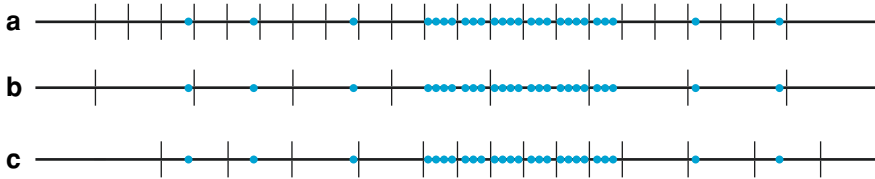
Figure 1.8 Selecting class intervals for "stretched out" dots: (**a**) many short equal width intervals; (**b**) a few wide equal width intervals; (**c**) unequal width intervals

---

DEFINITION   For any class to be used in a histogram, the **density** of the data in that class is defined by

$$\text{density} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

A histogram can then be constructed in which the height of the rectangle over each class is its density. The vertical scale on such a histogram is called a **density scale**.

   When class widths are unequal, not using a density scale will give a picture with distorted areas. For equal class widths, the divisor is the same in each density calculation, and the extra arithmetic simply results in a rescaling of the vertical axis (i.e., the histogram using relative frequency and the one using density will have exactly the same appearance).

   A density histogram does have one interesting property. Multiplying both sides of the formula for density by the class width gives

$$\text{relative frequency} = (\text{class width})(\text{density}) = (\text{rectangle width})(\text{rectangle height})$$
$$= \text{rectangle area}$$

That is, *the area of each rectangle is the relative frequency of the corresponding class.* Furthermore, because the sum of relative frequencies must be 1.0 (except for roundoff), *the total area of all rectangles in a density histogram is 1*. It is always possible to draw a histogram so that the area equals the relative frequency (this is true also for a histogram of counting data)—just use the density scale. This property will play an important role in creating models for certain distributions in Chapter 4.

Example 1.9 The Environmental Protection Agency (EPA) publishes information each year on the estimate gas mileage as well as expected annual fuel cost for hundreds of new vehicles. For 2018, the EPA evaluated 369 different cars and trucks. The fuel cost estimates ranged from \$700 (Toyota Camry Hybrid LE) to \$3800 (Bugatti Chiron). We have divided the expected annual fuel costs of these vehicles, in hundreds of dollars, into five intervals: 7–<10, 10–<15, 15–<20, 20–<25, and 25–38.

Class	7–<10	10–<15	15–<17.5	17.5–<20	20–<25	25–38
Frequency	3	86	103	83	74	20
Relative frequency	.0081	.2331	.2791	.2249	.2005	.0542
Density	.0027	.0466	.1117	.0900	.0401	.0042

The resulting histogram appears in Figure 1.9. The right or upper tail stretches out much farther than does the left or lower tail—a substantial departure from symmetry. Thankfully, high-efficiency/low-cost vehicles predominate, and gas guzzlers are relatively rare.
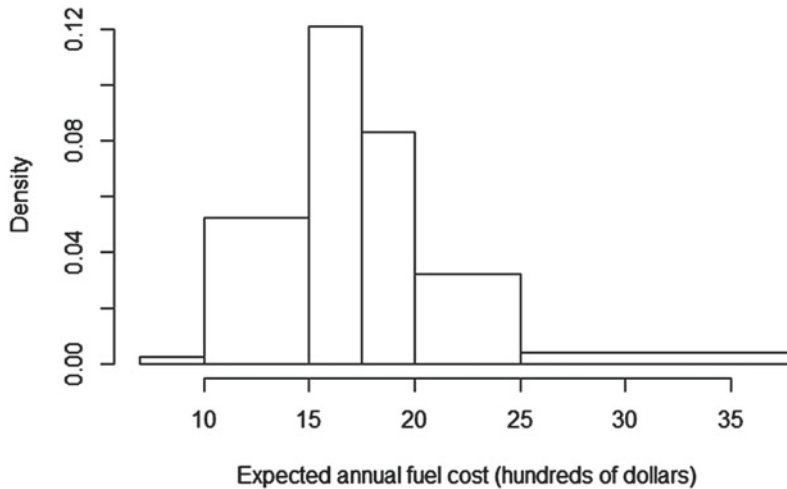
Figure 1.9 R density histogram for the fuel cost data of Example 1.9   ■

Histogram Shapes

Histograms come in a variety of shapes. A **unimodal** histogram is one that rises to a single peak and then declines. A **bimodal** histogram has two different peaks. Bimodality can occur when the data set consists of observations on two quite different kinds of individuals or objects. For example, the histogram of a data set consisting of driving times between San Luis Obispo and Monterey in California would show two peaks, one for those cars that took the inland route (roughly 2.5 h) and another for those cars traveling up the coast (3.5–4 h). A histogram with more than two peaks is said to be **multimodal**.

A histogram is **symmetric** if the left half is a mirror image of the right half. A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail and **negatively skewed** if the stretching is to the left. Figure 1.10 shows "smoothed" histograms, obtained by superimposing a smooth curve on the rectangles, that illustrate various possibilities.
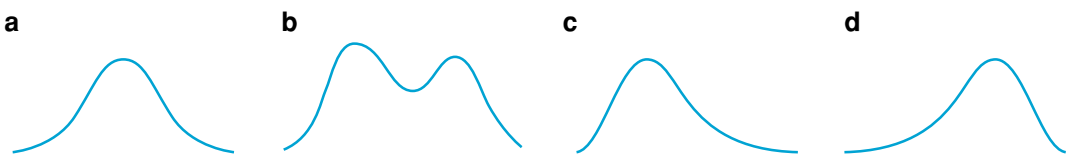
Figure 1.10 Smoothed histograms: (**a**) symmetric unimodal; (**b**) bimodal; (**c**) positively skewed; and (**d**) negatively skewed

Categorical Data

Both a frequency distribution and a **pie chart** or **bar graph** can be constructed when a data set is categorical in nature; generally speaking, statisticians prefer bar graphs over pie charts in most circumstances. Sometimes there will be a natural ordering of categories (freshman, sophomore, junior, senior, graduate student); for such *ordinal data* the categories should be presented in their natural order. In other cases the order will be arbitrary (e.g., Catholic, Jewish, Protestant, and so on); while we have the choice of displaying *nominal data* in any order, it's common to sort the categories in decreasing order of their (relative) frequencies. Either way, the rectangles for the bar graph should have equal width.

Example 1.10 Each member of a sample of 120 individuals owning motorcycles was asked for the name of the manufacturer of his or her bike. The frequency distribution for the resulting data is given in Table 1.2 and the bar chart is shown in Figure 1.11.

Table 1.2 Frequency distribution for motorcycle data

Manufacturer	Frequency	Relative frequency
1. Honda	41	.34
2. Yamaha	27	.23
3. Kawasaki	20	.17
4. Harley-Davidson	18	.15
5. BMW	3	.03
6. Other	11	.09
	120	1.01
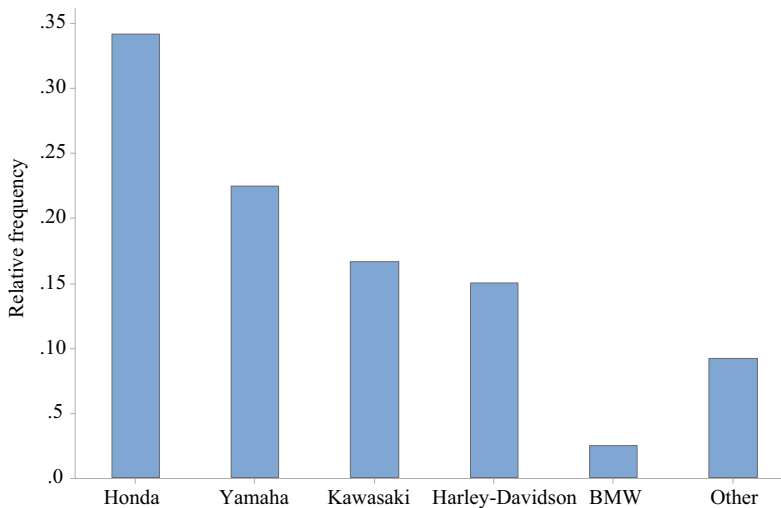
Figure 1.11 Bar chart for motorcycle data   ∎

Multivariate Data

The techniques presented so far have been exclusively for situations in which each observation in a data set is either a single number or a single category. Often, however, our data is *multivariate* in nature. That is, if we obtain a sample of individuals or objects and on each one we make two or more measurements, then each "observation" would consist of several measurements on one individual or object. The sample is bivariate if each observation consists of two measurements or responses, so that the data set can be represented as (x_1, y_1), ..., (x_n, y_n). For example, x might refer to engine size and y to horsepower, or x might refer to brand of cell phone owned and y to academic major. We consider the analysis of multivariate data in several later chapters.

Exercises: Section 1.2 (14–39)

14. Consider the fire resistance time data given in Example 1.2.

    a. Construct a stem-and-leaf display of the data. What appears to be a representative value? Do the observations appear to be highly concentrated about the representative value or rather spread out?

    b. Does the display appear to be reasonably symmetric about a representative value, or would you describe its shape in some other way?

    c. Do there appear to be any outlying fire resistance times?

    d. What proportion of times in this sample exceed 45 min?

15. Construct a stem-and-leaf display for the given batch of exam scores, repeating each stem twice (so, 6L through 9H). What feature of the data is highlighted by this display?

74	89	80	93	64	67	72	70	66	85
89	81	81	71	74	82	85	63	72	81
81	95	84	81	80	70	69	66	60	83
85	98	84	68	90	82	69	72	87	88

16. A sample of 77 individuals working at a particular office was selected and the noise level (dBA) experienced by each one was determined, yielding the following data ("Acceptable Noise Levels for Construction Site Offices," *Build. Serv. Engr. Res. Technol.* 2009: 87–94).

55.3	55.3	55.3	55.9	55.9	55.9	55.9	56.1	56.1	56.1	56.1
56.1	56.1	56.8	56.8	57.0	57.0	57.0	57.8	57.8	57.8	57.9
57.9	57.9	58.8	58.8	58.8	59.8	59.8	59.8	62.2	62.2	63.8
63.8	63.8	63.9	63.9	63.9	64.7	64.7	64.7	65.1	65.1	65.1
65.3	65.3	65.3	65.3	67.4	67.4	67.4	67.4	68.7	68.7	68.7
68.7	69.0	70.4	70.4	71.2	71.2	71.2	73.0	73.0	73.1	73.1
74.6	74.6	74.6	74.6	79.3	79.3	79.3	79.3	83.0	83.0	83.0

    Construct a stem-and-leaf display using repeated stems, and comment on any interesting features of the display.

17. The following data on crack depth (µm) was read from a graph in the article "Effects of Electropolishing on Corrosion and Stress Corrosion Cracking of Alloy 182 in High Temperature Water" (*Corrosion Sci.* 2017: 1–10)

1.5	2.9	3.1	3.3	3.4	3.6	3.7	3.8	3.9	4.1
4.3	4.5	4.6	4.7	4.8	5.2	5.3	5.5	5.6	5.9
6.1	6.9	7.2	7.5	8.0	8.0	8.1	8.2	8.5	9.2
9.5	9.9	10.0	10.5	10.5	10.7	10.9	10.9	11.2	11.3
11.3	11.8	12.0	12.7	14.4	15.7	17.3	18.4	19.9	20.0
21.7	21.8	22.4	26.4	33.7	33.8	34.0	37.8	42.2	46.0
48.6	50.2	51.4	52.4	66.5	76.1	81.1			

    a. Construct a stem-and-leaf display of the data.

    b. What is a typical, or representative, crack depth?

    c. Does the display appear to be highly concentrated or spread out?

    d. Does the distribution of values appear to be reasonably symmetric? If not, how would you describe the departure from symmetry?

    e. Would you describe any observations as being far from the rest of the data (outliers)?

18. The *15th Annual Demographia International Housing Affordability Survey: 2019* reports the "median multiple," the ratio of median home price to median household income, for 188 metropolitan areas in the United States. (A higher median multiple means that it's harder for residents of that area to purchase a home.) The resulting data appears below.

2.6	3.0	4.1	3.1	3.0	3.8	3.9	4.8	3.5	2.8
2.9	4.1	4.3	3.6	3.4	3.1	3.5	5.1	5.3	6.7
4.5	4.9	3.0	2.8	2.6	4.3	2.5	4.5	3.8	3.4
3.6	2.8	2.9	2.8	4.5	4.5	3.2	3.2	3.1	3.7
3.9	2.3	2.7	4.3	5.5	2.8	3.0	2.8	4.3	3.5
2.4	5.6	2.8	3.2	3.0	2.7	5.2	2.9	4.4	2.6
4.9	4.4	3.1	4.7	2.8	3.2	4.1	3.1	3.1	2.5
3.3	2.8	8.6	3.7	2.9	3.0	2.9	3.1	3.9	2.8
3.3	4.0	2.9	3.2	3.4	3.4	3.8	3.2	2.4	3.4
4.9	3.0	3.0	2.8	9.2	3.1	2.8	3.2	3.7	3.4
3.0	4.0	3.4	6.0	5.7	3.8	3.4	3.0	5.0	2.8
4.2	5.3	3.9	3.4	3.1	4.0	5.5	3.4	3.7	2.7
3.9	2.7	4.5	7.1	3.6	2.3	3.4	4.3	2.6	4.4
3.9	5.2	4.3	4.3	3.8	2.7	5.8	3.7	5.6	3.2
2.6	2.2	5.6	5.0	7.5	3.9	4.4	3.9	7.8	8.8
9.4	8.1	7.5	9.6	7.5	4.6	3.2	2.5	5.6	4.1
3.1	2.6	3.2	4.3	3.8	3.0	2.8	5.9	2.3	3.7
4.1	2.5	3.2	4.0	3.1	2.2	5.4	3.5	4.6	3.3
4.0	2.8	5.0	3.3	3.8	4.3	2.8	2.2		

a. Construct a stem-and-leaf display of the data.
b. What is a typical, or representative, median multiple?
c. Describe the shape of the distribution.
d. Values above 5.0 earn the city a "Severely Unaffordable" rating. What proportion of cities in the study have severely unaffordable housing?

19. Do running times of American movies differ somehow from times of French movies? The authors investigated this question by randomly selecting 25 recent movies of each type, resulting in the following running times:

American:	94	90	95	93	128	95	125
	91	104	116	162	102	90	110
	92	113	116	90	97	103	95
	120	109	91	138			
French:	123	116	90	158	122	119	125
	90	96	94	137	102	105	106
	95	125	122	103	96	111	81
	113	128	93	92			

Construct a *comparative* stem-and-leaf display by listing stems in the middle of your paper and then placing the American leaves out to the left and the French leaves out to the right. Then comment on interesting features of the display.

20. The report "Congestion Reduction Strategies" (Texas Transportation Institute, 2005) investigated how much additional time (in hours, per year per traveler) drivers spend in traffic during peak hours for a sample of urban areas. Data on "large" (e.g., Denver) and "very large" (e.g., Houston) urban areas appear below.

Large:	55	55	53	52	51	50	46
	46	43	40	39	38	35	33
	33	30	30	29	26	23	18
	17	14	13	12	10		
Very Large:	93	72	69	67	63	60	58
	57	51	51	49	49	38	

Construct a comparative stem-and-leaf display (see Exercise 19) of this data. Compare and contrast the extra time spent in traffic for drivers in large urban areas and very large urban areas.

21. Temperature transducers of a certain type are shipped in batches of fifty. A sample of 60 batches was selected, and the number of transducers in each batch not conforming to design specifications was determined, resulting in the following data:

2	1	2	4	0	1	3	2	0	5	3	3	1	3	2	4	7	0	2	3
0	4	2	1	3	1	1	3	4	1	2	3	2	2	8	4	5	1	3	1
5	0	2	3	2	1	0	6	4	2	1	6	0	3	3	3	6	1	2	3

a. Determine frequencies and relative frequencies for the observed values of x = number of nonconforming transducers in a batch.
b. What proportion of batches in the sample have at most five nonconforming transducers? What proportion have fewer than five? What proportion have at least five nonconforming units?

c. Draw a histogram of the data using relative frequency on the vertical scale, and comment on its features.

22. *Lotka's law* is used in library science to describe the productivity of authors in a given field. The article "Lotka's Law and Productivity Patterns of Authors in Biomedical Science in Nigeria on HIV/AIDS: a Bibliometric Approach" (*The Electronic Library* 2016: 789–807) provides the following frequency distribution for the number of articles written by various authors on HIV/AIDS over a five-year period in Nigeria:

Number of papers	1	2	3	4	5	6
Frequency	650	90	73	40	35	30
Number of papers	7	8	9	10	11	
Frequency	23	17	15	10	5	

a. Construct a histogram corresponding to this frequency distribution. What is the most interesting feature of the shape of the distribution?
b. What proportion of these authors published at least five papers? More than five papers?
c. Suppose the ten 10s and five 11s had been lumped into a single category displayed as "10+." Would you be able to draw a histogram? Explain.
d. Suppose that instead of the values 10 and 11 being listed separately, they had been combined into a 10–11 category with frequency 15. Would you be able to draw a histogram? Explain.

23. The article "Ecological Determinants of Herd Size in the Thorncraft's Giraffe of Zambia" (*Afric. J. Ecol.* 2010: 962–971) gave the following data (read from a graph) on herd size for a sample of 1570 herds over a 34-year period.

Herd size	1	2	3	4	5	6	7	8
Frequency	589	190	176	157	115	89	57	55
Herd size	9	10	11	12	13	14	15	17
Frequency	33	31	22	10	4	10	11	5
Herd size	18	19	20	22	23	24	26	32
Frequency	2	4	2	2	2	2	1	1

a. What proportion of the sampled herds had just one giraffe?
b. What proportion of the sampled herds had six or more giraffes (characterized in the article as "large herds")?
c. What proportion of the sampled herds had between five and ten giraffes, inclusive?
d. Draw a histogram using relative frequency on the vertical axis. How would you describe the shape of this histogram?

24. The article "Determination of Most Representative Subdivision" (*J. Energy Engr.* 1993: 43–55) gave data on various characteristics of subdivisions that could be used in deciding whether to provide electrical power using overhead lines or underground lines. Here are the values of the variable $x =$ total length of streets within a subdivision:

1280	5320	4390	2100	1240	3060	4770
1050	360	3330	3380	340	1000	960
1320	530	3350	540	3870	1250	2400
960	1120	2120	450	2250	2320	2400
3150	5700	5220	500	1850	2460	5850
2700	2730	1670	100	5770	3150	1890
510	240	396	1419	2109		

a. Construct a stem-and-leaf display using the thousands digit as the stem and the hundreds digit as the leaf, and comment on various features of the display.
b. Construct a histogram using class boundaries 0, 1000, 2000, 3000, 4000, 5000, and 6000. What proportion of subdivisions have total length less than 2000? Between 2000 and 4000? How would you describe the shape of the histogram?

25. The article cited in the previous exercise also gave the following values of the variables $y =$ number of culs-de-sac and $z =$ number of intersections:

y	1	0	1	0	0	2	0	1	1	1	2	1	0	0	1	1	0	1	1
z	1	8	6	1	1	5	3	0	0	4	4	0	0	1	2	1	4	0	4
y	1	1	0	0	0	1	1	2	0	1	2	2	1	1	0	2	1	1	0
z	0	3	0	1	1	0	1	3	2	4	6	6	0	1	1	8	3	3	5
y	1	5	0	3	0	1	1	0	0										
z	0	5	2	3	1	0	0	0	3										

a. Construct a histogram for the y data. What proportion of these subdivisions had no culs-de-sac? At least one cul-de-sac?

b. Construct a histogram for the z data. What proportion of these subdivisions had at most five intersections? Fewer than five intersections?

26. How does the speed of a runner vary over the course of a marathon (a distance of 42.195 km)? Consider determining both the time to run the first 5 km and the time to run between the 35 km and 40 km points, and then subtracting the former time from the latter time. A positive value of this difference corresponds to a runner slowing down toward the end of the race. The accompanying histogram is based on times of runners who participated in several different Japanese marathons ("Factors Affecting Runners' Marathon Performance," *Chance*, Fall 1993: 24–30).

What are some interesting features of this histogram? What is a typical difference value? Roughly what proportion of the runners ran the late distance more quickly than the early distance? (Time differences are in seconds.)
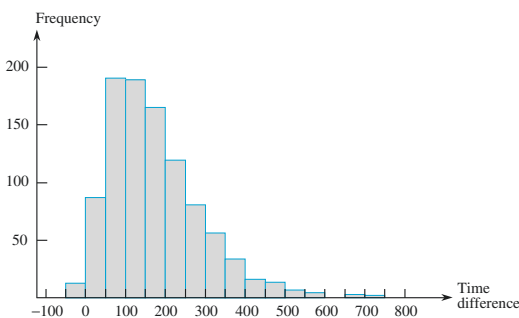
27. America used to be number 1 in the world for percentage of adults with four-year degrees, but it has recently dropped to 19th. Here is data on the percentage of adults age 25 or older in each state who had a four-year degree as of 2015 (listed in alphabetical order, with the District of Columbia included):

23.5	28.0	27.5	21.1	29.5	38.1	37.6	30.0	54.6
27.3	28.8	30.8	25.9	32.3	24.1	26.7	31.0	22.3
22.5	29.0	37.9	40.5	26.9	33.7	20.7	27.1	31.4
29.3	23.0	34.9	36.8	26.3	34.2	28.4	27.7	26.1
24.1	30.8	28.6	31.9	25.8	27.0	24.9	27.6	31.1
36.0	36.3	32.9	19.2	27.8	25.7			

a. Construct a dotplot, and comment on any interesting features. [*Note*: The values 54.6, 40.5, and 19.2 belong to DC, MA, and WV, respectively.]

b. The national percentage of adults age 25 or older with a four-year degree was 29.8% in 2015. Would you obtain that same value by averaging the 51 numbers provided? Why or why not?

28. Tire pressure monitoring systems are increasingly common for vehicles, but such systems rarely include checking the spare tire. The article "A Statistical Study of Tire Pressures on Road-Going Vehicles" (*SAE Intl. J. Passeng. Cars—Mech. Systems* 2016) provided the following data on the amount (psi) by which spare tires in a sample of 95 cars were under-inflated, relative to manufacturer's specifications:

29	25	60	57	7	35	20	23	5	52
58	20	60	40	52	9	7	57	57	55
−6	46	−6	19	32	−11	17	50	1	57
15	5	4	60	50	41	34	6	54	31
4	0	29	19	12	50	52	6	−3	41
8	50	32	12	38	32	46	51	26	20
16	20	30	8	0	42	16	41	35	45
−5	39	25	42	29	3	60	20	1	0
35	30	13	37	13	16	15	25	24	25
11	−12	10	10	5					

a. What does a value of 0 represent here? What does a negative value represent?

b. Construct a relative frequency histogram based on the class boundaries −20,

−10, 0, 10, …, 50, and 60, and comment on features of the distribution.

c. Construct a histogram based on the following class boundaries: −20, 0, 10, 20, 30, 40, and 60.

d. What proportion of spare tires in the sample were within ± 10 psi of their manufacturer-recommended pressure?

29. A transformation of data values by means of some mathematical function, such as $\sqrt{x}$ or $1/x$, can often yield a set of numbers that has "nicer" statistical properties than the original data. In particular, it may be possible to find a function for which the histogram of transformed values is more symmetric (or, even better, more like a bell-shaped curve) than the original data. As an example, the article "The Negative Binomial Distribution as a Model for External Corrosion Defect Counts in Buried Pipelines" (*Corrosion Sci.* 2015: 114–131) reported the number of defects in 50 oil and gas pipeline segments in southern Mexico.

46	518	274	37	46	85	365	40	378	18
29	43	153	23	206	34	25	37	125	84
33	170	63	49	88	54	144	45	27	14
349	148	321	183	148	61	65	127	116	35
57	46	81	156	59	26	88	33	104	44

a. Use class intervals 0–50, 50–<100, … to construct a histogram of the original data.

b. Transform the data by applying $\log_{10}()$ to all 50 values. Use class intervals 1.0–<1.2, 1.2–<1.4, …, 2.6–<2.8 to construct a histogram for the transformed data. What is the effect of the transformation?

30. Unlike most packaged food products, alcohol beverage container labels are not required to show calorie or nutrient content. The article "What Am I Drinking? The Effects of Serving Facts Information on Alcohol Beverage Containers" (*J. of Consumer Affairs* 2008: 81–99) reported on a pilot study in which each individual in a sample was asked to estimate the calorie content of a 12 oz can of light beer known to contain 103 cal. The following information appeared in the article:

Class	Percentage
0–<50	7
50–<75	9
75–<100	23
100–<125	31
125–<150	12
150–<200	3
200–<300	12
300–<500	3

a. Construct a histogram of the data and comment on any interesting features.

b. What proportion of the estimates were at least 100? Less than 200?

31. The report "Majoring in Money 2019" (Sallie Mae) provides the following relative frequency distribution for the credit card balance of a nationally representative sample of $n = 464$ college students:

$0	7%
$1–$100	15%
$101–$500	35%
$501–$1000	14%
$1001–$2500	18%
$2501–$5000	6%
>$5000	5%

a. Approximately how many students in the survey reported a $0 credit card balance?

b. What proportion of students surveyed carry a balance greater than $1000?

c. Based on the information provided, is it possible to construct a histogram of the data? Why or why not?

32. The College Board reports the following Total SAT score distribution for 2018, the first year of the "new" SAT format:

Score range	Frequency
1400–1600	145,023
1200–1390	434,200
1000–1190	741,452
800–990	619,145
600–790	192,267
400–590	4452

a. Create a histogram of this data. Comment on its features.

b. What is a typical, or representative, Total SAT score?

c. What proportion of students in 2018 scored between 800 and 1190?

33. The article "Study on the Life Distribution of Microdrills" (*J. Engr. Manuf.* 2002: 301–305) reported the following observations, listed in increasing order, on drill lifetime (number of holes that a drill machines before it breaks) when holes were drilled in a certain brass alloy.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

a. Construct a frequency distribution and histogram of the data using class boundaries 0, 50, 100, …, and then comment on interesting characteristics.

b. Construct a frequency distribution and histogram of the natural logarithms of the lifetime observations, and comment on interesting characteristics.

c. What proportion of the lifetime observations in this sample are less than 100? What proportion of the observations are at least 200?

34. Consider the following data on type of health complaint (J = joint swelling, F = fatigue, B = back pain, M = muscle weakness, C = coughing, N = nose running/irritation, O = other) made by tree planters. Obtain frequencies and relative frequencies for various categories, and construct a bar chart. (The data is consistent with percentages given in the article "Physiological Effects of Work Stress and Pesticide Exposure in Tree Planting by British Columbia Silviculture Workers," *Ergonomics* 1993: 951–961.)

O	O	N	J	C	F	B	B	F	O	J	O	O	M
O	F	F	O	O	N	O	N	J	F	J	B	O	C
J	O	J	J	F	N	O	B	M	O	J	M	O	B
O	F	J	O	O	B	N	C	O	O	O	M	B	F
J	O	F	N										

35. The report "Motorcycle Helmet Use in 2005—Overall Results" (NHTSA August 2005) included observations made on the helmet use of 1700 motorcyclists across the country. The data is summarized in the accompanying table. (A "noncompliant helmet" failed to meet U.S. Department of Transportation safety guidelines.)

Category	Frequency
No helmet	731
Noncompliant helmet	153
Compliant helmet	816
Total	1700

a. What is the variable in this example? What are the possible values of that variable?

b. Construct the relative frequency distribution of this variable.

c. What proportion of observed motorcyclists wore some kind of helmet?

d. Construct an appropriate graph for this data.

36. The author of the article "Food and Eating on Television: Impacts and Influences" (*Nutr. and Food Sci.* 2000: 24–29) examined hundreds of hours of BBC television footage and categorized food images for both TV programs and commercials. The data presented here is consistent with information in the article; one of the research goals was to compare food images in ads and programs.

	Number of food images	
Food category	TV Programs	Commercials
Milk and dairy products	149	99
Breads, cereals, and potatoes	372	346
Meat, fish, and alternatives	248	198
Fruits and vegetables	694	32
Fatty and sugary foods	322	511
Total	$n = 1785$	$n = 1186$

   a. Why is it inappropriate to compare frequencies (counts) between program images and commercial images?

   b. Obtain the relative frequency distribution for the variable food category among images in TV programs. Create a graph of the distribution.

   c. Repeat part (b) for food images in commercials.

   d. Contrast the two distributions: what are the biggest differences between the types of food images in TV programs and those in commercials?

37. A **Pareto diagram** is a variation of a bar chart for categorical data resulting from a quality control study. Each category represents a different type of product nonconformity or production problem. The categories are ordered so that the one with the largest frequency appears on the far left, then the category with the second-largest frequency, and so on. Suppose the following information on nonconformities in circuit packs is obtained: failed component, 126; incorrect component, 210; insufficient solder, 67; excess solder, 54; missing component, 131. Construct a Pareto diagram.

38. The **cumulative frequency** and **cumulative relative frequency** for a particular class interval are the sum of frequencies and relative frequencies, respectively, for that interval and all intervals lying below it. If, for example, there are four intervals with frequencies 9, 16, 13, and 12, then the cumulative frequencies are 9, 25, 38, and 50, and the cumulative relative frequencies are .18, .50, .76, and 1.00. Compute the cumulative frequencies and cumulative relative frequencies for the data of Exercise 28, using the class intervals in part (c).

39. Fire load (MJ/m^2) is the heat energy that could be released per square meter of floor area by combustion of contents and the structure itself. The article "Fire Loads in Office Buildings" (*J. Struct. Engr*. 1997: 365–368) gave the following cumulative percentages (read from a graph) for fire loads in a sample of 388 rooms:

Value	0	150	300	450	600
Cumulative %	0	19.3	37.6	62.7	77.5
Value	750	900	1050	1200	1350
Cumulative %	87.2	93.8	95.7	98.6	99.1
Value	1500	1650	1800	1950	
Cumulative %	99.5	99.6	99.8	100.0	

   a. Construct a relative frequency histogram and comment on interesting features.

   b. What proportion of fire loads are less than 600? At least 1200?

   c. What proportion of the loads are between 600 and 1200?

1.3 Measures of Center

Visual summaries of data are excellent tools for obtaining preliminary impressions and insights. More formal data analysis often requires the calculation and interpretation of numerical summary measures —numbers that might serve to characterize the data set and convey some of its most important features.

Our primary concern will be with quantitative data. Suppose that our data set is of the form $x_1, x_2,$ …, x_n, where each x_i is a number. What features of such a set of numbers are of most interest and deserve emphasis? One important characteristic of a set of numbers is its "center": a single value that we might consider typical or representative of the entire data set. This section presents methods for

describing the center of a data set; in Section 1.4 we will turn to methods for measuring variability in a set of numbers.

The Mean

For a given set of numbers x_1, x_2, ..., x_n, the most familiar and useful measure of the center is the *mean*, or arithmetic average, of the set. Because we will almost always think of the x_i's as constituting a sample, we will often refer to the arithmetic average as the *sample mean* and denote it by $\bar{x}$.

DEFINITION   The **sample mean** $\bar{x}$ of observations x_1, x_2, ..., x_n is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The numerator of $\bar{x}$ can be written more informally as $\sum x_i$ where the summation is over all sample observations.

For reporting $\bar{x}$, we recommend using decimal accuracy of one digit more than the accuracy of the x_i's. Thus if observations are stopping distances with $x_1 = 125$, $x_2 = 131$, and so on, we might have $\bar{x} = 127.3$ ft.

Example 1.11 Students in a class were assigned to make wingspan measurements at home. The wingspan is the horizontal measurement from fingertip to fingertip with outstretched arms. Here are the measurements (inches) given by 21 of the students.

$x_1 = 60$	$x_2 = 64$	$x_3 = 72$	$x_4 = 63$	$x_5 = 66$	$x_6 = 62$	$x_7 = 75$
$x_8 = 66$	$x_9 = 59$	$x_{10} = 75$	$x_{11} = 69$	$x_{12} = 62$	$x_{13} = 63$	$x_{14} = 61$
$x_{15} = 65$	$x_{16} = 67$	$x_{17} = 65$	$x_{18} = 69$	$x_{19} = 95$	$x_{20} = 60$	$x_{21} = 70$

Figure 1.12 shows a stem-and-leaf display of the data; a wingspan in the 60s appears to be "typical."

```
5H|9
6L|00122334
6H|5566799
7L|02
7H|55
8L|
8H|
9L|
9H|5
```

Figure 1.12 A stem-and-leaf display of the wingspan data

With $\sum x_i = 1408$, the sample mean is

$$\bar{x} = \frac{1408}{21} = 67.0 \text{ in,}$$

a value consistent with information conveyed by the stem-and-leaf display.   ∎

A physical interpretation of $\bar{x}$ demonstrates how it measures the center of a sample. Think of a dotplot in which each dot (i.e., each observation) "weighs" 1 lb. The only point at which a fulcrum

can be placed to balance the system of weights is the point corresponding to the value of $\bar{x}$ (see Figure 1.13). The system balances because, as shown in the next section, $\sum(x_i - \bar{x}) = 0$, so the net total tendency to turn about $\bar{x}$ is 0.
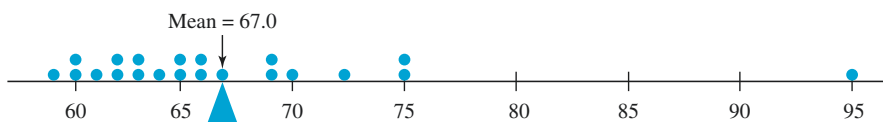
Figure 1.13   The mean as the balance point for a system of weights

Just as $\bar{x}$ represents the average value of the observations in a sample, the average of all values in a population can, in principle, be calculated. This average is called the **population mean** and will be denoted by the Greek letter μ. When there are N values in the population (a finite population), then $\mu =$ (sum of the N population values)$/N$. In Chapters 3 and 4, we will give a more general definition for μ that applies to both finite and (conceptually) infinite populations. In the chapters on statistical inference, we will present methods based on the sample mean for drawing conclusions about a population mean. For example, we might use the sample mean $\bar{x} = 67.0$ computed in Example 1.11 as a *point estimate* (a single number that is our "best" guess) of $\mu =$ the true average wingspan for all students in introductory statistics classes.

The mean suffers from one deficiency that makes it an inappropriate measure of center in some circumstances: its value can be greatly affected by the presence of even a single outlier (i.e., an unusually large or small observation). In Example 1.11, the value $x_{19} = 95$ is obviously an outlier. Without this observation, $\bar{x} = 1313/20 = 65.7$ in; the outlier increases the mean by 1.3 in. The value 95 is clearly an error—this student was only 70 inches tall, and there is no way such a student could have a wingspan of almost 8 ft. As Leonardo da Vinci noticed, wingspan is usually quite close to height. (Note, though, that outliers are often *not* the result of recording errors!)

We will next consider an alternative to the mean, namely the median, that is insensitive to outliers. However, the mean is still by far the most widely used measure of center, largely because there are many populations for which outliers are very scarce. When sampling from such a population (a "normal" or bell-shaped distribution being the most important example), outliers are highly unlikely to enter the sample. The sample mean will then tend to be stable and quite representative of the sample.

The Median
The word *median* is synonymous with "middle," and the sample median is indeed the middle value when the observations are ordered from smallest to largest. When the observations are denoted by x_1, ..., x_n, we will use the symbol $\tilde{x}$ to represent the sample median.

DEFINITION   The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included so that every sample observation appears in the ordered list). Then, if n is odd,

$$\tilde{x} = \left(\tfrac{n+1}{2}\right)\text{th ordered value}$$

whereas if n is even,

$$\tilde{x} = \text{average of the } \left(\tfrac{n}{2}\right)\text{th and } \left(\tfrac{n}{2} + 1\right)\text{th ordered values}$$

Example 1.12 People not familiar with classical music might tend to believe that a composer's instructions for playing a particular piece are so specific that the duration would not depend at all on the performer(s). However, there is typically plenty of room for interpretation, and orchestral conductors and musicians take full advantage of this. We went to the website ArkivMusic.com and selected a sample of 12 recordings of Beethoven's Symphony No. 9 (the "Choral," a stunningly beautiful work), yielding the following durations (min) listed in increasing order:

| 62.3 | 62.8 | 63.6 | 65.2 | 65.7 | 66.4 | 67.4 | 68.4 | 68.8 | 70.8 | 75.7 | 79.0 |

Since $n = 12$ is even, the sample median is the average of the $n/2 = $ 6th and $(n/2 + 1) = $ 7th values from the ordered list:

$$\tilde{x} = \frac{66.4 + 67.4}{2} = 66.90 \, \text{min}$$

Note that half of the durations in the sample are less than 66.90 min, and half are greater than that.

If the largest observation 79.0 had not been included in the sample, the resulting sample median for the $n = 11$ remaining observations would have been the single middle value 66.4 (the $[n + 1]/2 = $ 6th ordered value, i.e., the 6th value in from either end of the ordered list).

The sample mean is $\bar{x} = \sum x_i/n = 816.1/12 = 68.01$ min, a bit more than a full minute larger than the median. The mean is pulled out a bit relative to the median because the sample "stretches out" somewhat more on the upper end than on the lower end. ■

The data in Example 1.12 illustrates an important property of $\tilde{x}$ in contrast to $\bar{x}$. The sample median is very insensitive to a number of extremely small or extremely large data values. If, for example, we increased the two largest x_i's from 75.7 and 79.0 to 95.7 and 99.0, respectively, $\tilde{x}$ would be unaffected. Thus, in the treatment of outlying data values, $\bar{x}$ and $\tilde{x}$ are at opposite ends of a spectrum: $\bar{x}$ is sensitive to even one such value, whereas $\tilde{x}$ is insensitive to a large number of outlying values. Although $\bar{x}$ and $\tilde{x}$ both provide a measure for the center of a data set, they will not in general be equal because they focus on different aspects of the sample.

Analogous to $\tilde{x}$ as the middle value in the sample is a middle value in the population, the **population median**, denoted by $\tilde{\mu}$. As with $\bar{x}$ and μ, we can think of using the sample median $\tilde{x}$ to make an inference about $\tilde{\mu}$. In Example 1.12, we might use $\tilde{x} = 66.90$ min as an estimate of the median duration in the entire population from which the sample was selected. Or, if the median salary for a sample of statisticians was $\tilde{x} = \$96,416$, we might use this as a basis for concluding that the median salary $\tilde{\mu}$ for *all* statisticians exceeds \$90,000.

The population mean μ and median $\tilde{\mu}$ will not generally be identical. If the population distribution is positively or negatively skewed, as shown in Figure 1.14 (p. 29), then $\mu \neq \tilde{\mu}$. When this is the case, in making inferences we must first decide which of the two population characteristics is of greater interest and then proceed accordingly. As an example, according to the report "How America Saves 2019" issued by the Vanguard Funds investment company, the mean retirement fund balance among workers 65 and older is \$192,877, whereas the median balance is just \$58,035. Clearly a small minority of such people has extremely large retirement fund balances, inflating the mean relative to the median; the latter is arguably a better representation of a "typical" retirement fund balance.
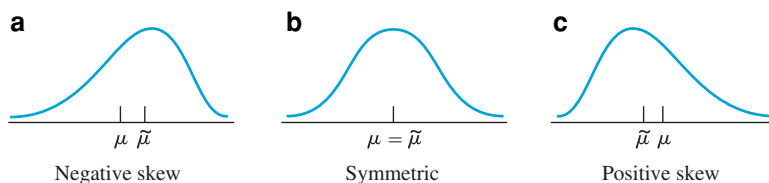
Figure 1.14 Three different shapes for a population distribution

Trimmed Means

The sample mean and sample median are influenced by outlying values to very different degrees—the mean greatly and the median not at all. Since extreme behavior of either type might be undesirable, we briefly consider alternative measures that are neither as sensitive as $\bar{x}$ nor as insensitive as $\tilde{x}$. To motivate these alternatives, note that $\bar{x}$ and $\tilde{x}$ are at opposite extremes of the same "family" of measures: while $\tilde{x}$ is computed by throwing away as many values on each end as one can without eliminating everything (leaving just one or two middle values), to compute $\bar{x}$ one throws away nothing before averaging. Said differently, the mean involves "trimming" 0% from each end of the sample, whereas for the median the maximum possible amount is trimmed from each end. A **trimmed mean** is a compromise between $\bar{x}$ and $\tilde{x}$. A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and the largest 10% of the sample and then averaging what remains.

Example 1.13 Consider the following 20 observations, ordered from smallest to largest, each one representing the lifetime (in hours) of a type of incandescent lamp:

612	623	666	744	883	898	964	970	983	1003
1016	1022	1029	1058	1085	1088	1122	1135	1197	1201

The mean and median of all 20 observations are $\bar{x} = 965.0$ h and $\tilde{x} = 1009.5$ h, respectively. The 10% trimmed mean is obtained by deleting the smallest two observations (612 and 623) and the largest two (1197 and 1201) and then averaging the remaining 16 to obtain $\bar{x}_{tr(10)} = 979.1$ h. The effect of trimming here is to produce a "central value" that is somewhat above the mean ($\bar{x}$ is pulled down by a few small lifetimes) and yet considerably below the median. Similarly, the 20% trimmed mean averages the middle 12 values to obtain $\bar{x}_{tr(20)} = 999.9$, even closer to the median. See Figure 1.15.
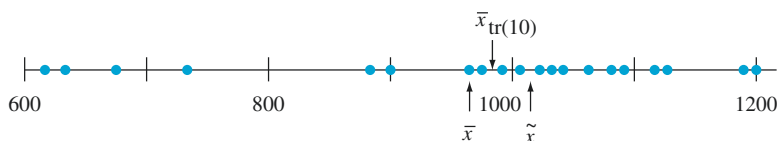
Figure 1.15 Dotplot of lifetimes (in hours) of incandescent lamps ∎

Generally speaking, using a trimmed mean with a moderate trimming proportion (between 5% and 25%) will yield a measure that is neither as sensitive to outliers as the mean nor as insensitive as the median. For this reason, trimmed means have merited increasing attention from statisticians for both descriptive and inferential purposes. More will be said about trimmed means when point estimation is discussed in Chapter 7. As a final point, if the trimming proportion is denoted by α and $n\alpha$ is not an integer, then it is not obvious how the $100\alpha\%$ trimmed mean should be computed. For example, if

$\alpha = .10$ (10%) and $n = 22$, then $n\alpha = (22)(.10) = 2.2$, and we cannot trim 2.2 observations from each end of the ordered sample. In this case, the 10% trimmed mean would be obtained by first trimming two observations from each end and calculating $\bar{x}_{tr}$, then trimming three and calculating $\bar{x}_{tr}$, and finally interpolating between the two values to obtain $\bar{x}_{tr(10)}$.

Exercises: Section 1.3 (40–51)

40. The website realtor.com listed the following sale prices (in $1000s) for a sample of 10 homes sold in 2019 in Los Osos, CA (home town of two of the authors):

    525 830 600 180 129 525 350 490 640 475

    a. Calculate and interpret the sample mean and median.
    b. Suppose the second observation was 930 instead of 830. How would that affect the mean and median?
    c. The two low outliers in the sample were mobile homes. If we excluded those two observations, how would that affect the mean and median?
    d. Calculate a 20% trimmed mean by first trimming the two smallest and two largest observations.
    e. Calculate a 15% trimmed mean.

41. Super Bowl LIII was the lowest scoring (and, to many, the least exciting) Super Bowl of all time. During the game, Los Angeles Rams running back Todd Gurley had just 10 rushing plays, resulting in the following gains in yards:

    5 2 1 2 3 −1 16 2 5 0

    a. Determine the value of the mean.
    b. Determine the value of the median. Why is it so different from the mean?
    c. Calculate a trimmed mean by deleting the smallest and largest observations. What is the corresponding trimming percentage? How does the value of this trimmed mean compare to the mean and median?

42. The minimum injection pressure (psi) for injection molding specimens of high amylose corn was determined for eight different specimens (higher pressure corresponds to greater processing difficulty), resulting in the following observations (from "Thermoplastic Starch Blends with a Polyethylene-Co-Vinyl Alcohol: Processability and Physical Properties," *Polymer Engr. and Sci.* 1994: 17–23):

    15.0   13.0   18.0   14.5   12.0   11.0   8.9   8.0

    a. Determine the values of the sample mean, sample median, and 12.5% trimmed mean, and compare these values.
    b. By how much could the smallest sample observation, currently 8.0, be increased without affecting the value of the sample median?
    c. Suppose we want the values of the sample mean and median when the observations are expressed in kilograms per square inch (ksi) rather than psi. Is it necessary to re-express each observation in ksi, or can the values calculated in part (a) be used directly? [*Hint:* 1 kg = 2.2 lb.]

43. Here is the average weekday circulation (paper plus digital subscriptions) for the top 20 newspapers in the country (247wallst.com, January 24, 2017):

2,237,601	512,118	507,395	424,721	410,587
384,962	356,768	299,538	291,991	285,129
276,445	246,963	245,042	243,376	242,567
232,546	232,372	227,245	215,476	196,286

    a. Which value, the mean or the median, do you anticipate will be higher? Why?
    b. Calculate the mean and median for this data.

44. An article in the *Amer. J. Enol. and Viti.* (2006: 486–490) includes the following

alcohol content measurements (%) for a sample of $n = 35$ port wines.

16.35	17.73	19.62	19.07	19.48	19.45	19.33
18.85	22.75	19.20	19.90	20.00	19.37	21.22
16.20	23.78	20.05	18.68	19.97	19.20	19.50
17.75	23.25	17.85	18.82	17.48	18.00	15.30
19.58	19.08	19.17	19.03	17.15	19.60	22.25

a. Graph the data. Based on the graph, what is a representative value for the alcohol content in port wines?

b. Calculate the mean and the median. Are these values consistent with your answer in (a)? Why or why not?

45. Compute the sample median, 25% trimmed mean, 10% trimmed mean, and sample mean for the microdrill data given in Exercise 33, and compare these measures.

46. A sample of 26 offshore oil workers took part in a simulated escape exercise, resulting in the accompanying data on time (sec) to complete the escape ("Oxygen Consumption and Ventilation During Escape from an Offshore Platform," *Ergonomics* 1997: 281–292):

389	356	359	363	375	424	325	394	402
373	373	370	364	366	364	325	339	393
392	369	374	359	356	403	334	397	

a. Construct a stem-and-leaf display of the data. How does it suggest that the sample mean and median will compare?

b. Calculate the values of the sample mean and median.

c. By how much could the largest time, currently 424, be increased without affecting the value of the sample median? By how much could this value be decreased without affecting the value of the sample median?

d. What are the values of $\bar{x}$ and $\tilde{x}$ when the observations are re-expressed in minutes?

47. Blood pressure values are often reported to the nearest 5 mmHg (100, 105, 110, etc.).

Suppose the actual blood pressure values for nine randomly selected individuals are

118.6 127.4 138.4 130.0 113.7 122.0 108.3 131.5 133.2

a. What is the median of the *reported* blood pressure values?

b. Suppose the blood pressure of the second individual is 127.6 rather than 127.4 (a small change in a single value). How does this affect the median of the reported values? What does this say about the sensitivity of the median to rounding or grouping in the data?

48. Let $x_1, \ldots, x_n$ be a sample, and let a and b be constants with $a \neq 0$. Define a new sample $y_1, \ldots, y_n$ by $y_1 = ax_1 + b$, ..., $y_n = ax_n + b$.

a. How does the sample mean of the y_i's relate to the mean of the x_i's? Verify your conjectures.

b. How does the sample median of the y_i's relate to the median of the x_i's? Substantiate your assertion.

49. An experiment to study the lifetime (in hours) for a certain type of component involved putting ten components into operation and observing them for 100 h. Eight of the components failed during that period, and those lifetimes were recorded. Denote the lifetimes of the two components still functioning after 100 h by 100+. The resulting sample observations were

48 79 100+ 35 92 86 57 100+ 17 29

Which of the measures of center discussed in this section can be calculated, and what are the values of those measures? [*Note:* The data from this study is said to be "right-censored."]

50. A sample of $n = 10$ automobiles was selected, and each was subjected to a 5-mph crash test. Denoting a car with no visible

damage by S (for success) and a car with such damage by F, results were as follows:

S   S   F   S   S   S   F   F   S   S

a. What is the sample proportion of successes?
b. Replace each S with a 1 and each F with a 0. Then calculate $\bar{x}$ for this numerically coded sample. How does $\bar{x}$ compare to the sample proportion of successes?
c. Suppose it is decided to include 15 more cars in the experiment. How many of these would have to be S's to give a sample proportion of .80 for the entire sample of 25 cars?

51. Refer back to Example 1.10, in which 120 motorcycle owners were asked to specify their bikes' manufacturer.
a. Is the variable *manufacturer* quantitative or categorical?
b. Based on the sample data, what would you consider a "typical" or "representative" value for the variable, and why?
c. Suppose the responses were recoded according to the numbering indicated in Table 1.2 (1 = Honda, 2 = Yamaha, etc.), resulting in a data set consisting of 41 1's, 27 2's, and so on. Would it be reasonable to use the mean of these 120 numbers as a representative value? What about the median? Explain.

1.4   Measures of Variability

Reporting a measure of center gives only partial information about a data set or distribution. Different samples or populations may have identical measures of center yet differ from one another in other important ways. Figure 1.16 shows dotplots of three samples with the same mean and median, yet the extent of spread about the center is different for all three samples. The first sample has the largest amount of variability, the second has less variability than the first, and the third has the smallest amount.
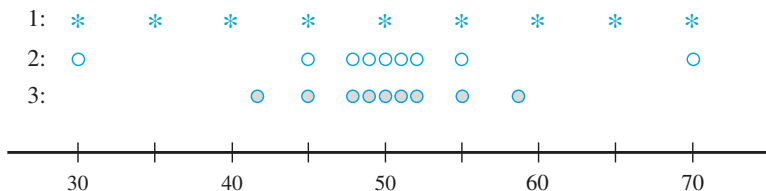
Figure 1.16   Samples with identical measures of center but different amounts of variability

Measures of Variability for Sample Data

The simplest measure of variability in a sample is the **range**, which is the difference between the largest and smallest sample values. Notice that the value of the range for sample 1 in Figure 1.16 is much larger than it is for sample 3, reflecting more variability in the first sample than in the third. A defect of the range, though, is that it depends on only the two most extreme observations and disregards the positions of the remaining $n - 2$ values. Samples 1 and 2 in Figure 1.16 have identical ranges, yet when we take into account the observations between the two extremes, there is much less variability or dispersion in the second sample than in the first.

Our primary measures of variability will involve the n **deviations from the mean**: $x_1 - \bar{x}$, $x_2 - \bar{x}$, $\ldots, x_n - \bar{x}$ obtained by subtracting $\bar{x}$ from each sample observation. A deviation will be positive if the observation is larger than the mean (to the right of the mean on the measurement axis) and negative if the observation is smaller than the mean. If all the deviations are small in magnitude, then all x_i's are close to the mean and there is little variability. On the other hand, if some of the deviations are large in magnitude, then some x_i's lie far from $\bar{x}$, suggesting a greater amount of variability.

A simple way to combine the deviations into a single quantity is to average them (sum them and divide by n). Unfortunately, this does not yield a useful measure, because the positive and negative deviations counteract one another:

$$\text{sum of deviations} = \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Thus the average deviation is always zero. To see why, use standard rules of summation and the fact that $\sum \bar{x} = \bar{x} + \bar{x} + \cdots + \bar{x} = n\bar{x}$:

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = \sum x_i - n\left(\frac{1}{n}\sum x_i\right) = 0$$

Another possibility is to base a measure on the absolute values of the deviations, in particular the mean absolute deviation $\sum |x_i - \bar{x}|/n$. But because the absolute value operation leads to some calculus-related difficulties, statisticians instead work with the squared deviations $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$. Rather than use the average squared deviation $\sum (x_i - \bar{x})^2/n$, for several reasons the sum of squared deviations is divided by $n - 1$ rather than n.

DEFINITION   Let $S_{xx} = \sum (x_i - \bar{x})^2$, the sum of the squared deviations from the mean. Then the **sample standard deviation**, denoted by s, is given by

$$s = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

The quantity s^2 is called the **sample variance**.

The unit for s is the same as the unit for each of the x_i's. If, for example, the observations are fuel efficiencies in miles per gallon, then we might have $s = 2.0$ mpg. A rough interpretation of the sample standard deviation is that it represents the size of a typical deviation from the sample mean within the given sample. Thus if $s = 2.0$ mpg, then some x_i's in the sample are closer than 2.0 to $\bar{x}$, whereas others are farther away; 2.0 is a representative (or "standard") deviation from the mean fuel efficiency. If $s = 3.0$ for a second sample of cars of another type, a typical deviation in this sample is roughly 1.5 times what it is in the first sample, an indication of more variability in the second sample.

Example 1.14 The website www.fueleconomy.gov contains a wealth of information about fuel characteristics of various vehicles. In addition to EPA mileage ratings, there are many vehicles for which users have reported their own values of fuel efficiency (mpg). Consider Table 1.3 with $n = 10$ efficiencies for the 2015 Toyota Camry (for this model, the EPA reports an overall rating of 25 mpg in city driving and 34 mpg in highway driving).

Table 1.3 Data for Example 1.14

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	31.0	2.1	4.41
2	27.8	−1.1	1.21
3	38.3	9.4	88.36
4	27.0	−1.9	3.61
5	23.4	−5.5	30.25
6	30.0	1.1	1.21
7	30.1	1.2	1.44
8	21.5	−7.4	54.76
9	25.4	−3.5	12.25
10	34.5	5.6	31.36
	$\sum x_i = 289.0$ $\bar{x} = 28.9$	$\sum(x_i - \bar{x}) = 0.0$	$\sum(x_i - \bar{x})^2 = 228.86$

The numerator of s^2 is $S_{xx} = 228.86$, from which

$$s = \sqrt{\frac{S_{xx}}{n-1}} = \sqrt{\frac{228.86}{10-1}} = \sqrt{25.43} = 5.04 \text{ mpg}$$

The size of a typical difference between a driver's fuel efficiency and the mean of 28.9 in this sample is roughly 5.04 mpg. ∎

To explain why $n - 1$ rather than n is used to compute s, note first that whereas s measures variability in a sample, there is a measure of population variability called the **population standard deviation**. We will use σ (lowercase Greek letter sigma) to denote the population standard deviation and σ^2 to denote the population variance. When the population is finite and consists of N values,

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 / N$$

which is the average of all squared deviations from the population mean (for the population, the divisor is N and not $N - 1$). More general definitions of σ^2 for (conceptually) infinite populations appear in Chapters 3 and 4.

Just as $\bar{x}$ will be used to make inferences about the population mean μ, we should define the sample standard deviation s so that it can be used to make inferences about σ. Note that σ involves squared deviations about the population mean μ. If we actually knew the value of μ, then we could define the sample standard deviation as the average squared deviation of the sample x_i's about μ. However, the value of μ is almost never known, so the sum of squared deviations about $\bar{x}$ must be used in the definition of s. But the x_i's tend to be closer to their own average $\bar{x}$ than to the population average μ. Using the divisor $n - 1$ rather than n compensates for this tendency. A more formal explanation for this choice appears in Chapter 7.

It is customary to refer to s as being based on $n - 1$ **degrees of freedom** (df). This terminology results from the fact that although s is based on the n quantities $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$, these sum to 0, so specifying the values of any $n - 1$ of the quantities determines the remaining value. For example, if $n = 4$ and $x_1 - \bar{x} = 8$, $x_2 - \bar{x} = -6$, and $x_4 - \bar{x} = -4$, then automatically $x_3 - \bar{x} = 2$, so only three of the four values of $x_i - \bar{x}$ are "freely determined" (3 df).

A Computing Formula for *s*

Typically, software or a calculator is used to compute summary quantities such as $\bar{x}$ and s. Otherwise, computing and squaring the deviations can be tedious, especially if enough decimal accuracy is being used in $\bar{x}$ to guard against the effects of rounding. An alternative formula for the numerator of s^2 circumvents the need for all the subtraction necessary to obtain the deviations.

PROPOSITION   An alternative expression for the numerator of s^2 is

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Proof Because $\bar{x} = \sum x_i / n$, $n\bar{x}^2 = n(\sum x_i)^2 / n^2 = (\sum x_i)^2 / n$. Then,

$$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x} \cdot x_i + \bar{x}^2) = \sum x_i^2 - 2\bar{x} \sum x_i + \sum \bar{x}^2$$

$$= \sum x_i^2 - 2\bar{x} \cdot n\bar{x} + n \cdot \bar{x}^2 = \sum x_i^2 - n(\bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \qquad \blacksquare$$

Example 1.15 Traumatic knee dislocation often requires surgery to repair ruptured ligaments. One measure of recovery is range of motion, measured as the angle formed when, starting with the leg straight, the knee is bent as far as possible. The given data on postsurgical range of motion appeared in the article "Reconstruction of the Anterior and Posterior Cruciate Ligaments After Knee Dislocation" (*Amer. J. Sports Med.* 1999: 189–197):

154	142	137	133	122	126	135	135	108	120	127	134	122

The sum of these 13 sample observations is $\sum x_i = 1695$, and the sum of their squares is

$$\sum x_i^2 = 154^2 + 142^2 + \cdots + 122^2 = 222{,}581$$

Thus the numerator of the sample variance is

$$S_{xx} = \sum x_i^2 - \left(\sum x_i\right)^2 / n = 222{,}581 - (1695)^2 / 13 = 1579.0769$$

from which $s^2 = 1579.0769/12 = 131.59$ and $s = \sqrt{131.59} = 11.47$ degrees. $\qquad \blacksquare$

If our data is rescaled—for instance, changing Celsius temperature measurements to Fahrenheit— the standard deviation of the rescaled data can easily be determined from the standard deviation of the original values.

PROPOSITION Let $x_1, x_2, \ldots, x_n$ be a sample and c be a constant.

1. If $y_1 = x_1 + c$, $y_2 = x_2 + c, \ldots, y_n = x_n + c$, then $s_y = s_x$; and
2. If $y_1 = cx_1, \ldots, y_n = cx_n$, then $s_y^2 = c^2 s_x^2$ and $s_y = |c| s_x$

where s_x is the sample standard deviation of the x_i's and s_y is the sample standard deviation of the y_i's.

Result 1 is intuitive, because adding or subtracting c shifts the location of the data set but leaves distances between data values unchanged. According to Result 2, multiplication of each x_i by c results in s being multiplied by a factor of $|c|$. Verification of these results utilizes the properties $\bar{y} = \bar{x} + c$ and $\bar{y} = c\bar{x}$ (see Exercise 72).

Quartiles and the Interquartile Range

In Section 1.3, we discussed the sensitivity of the sample mean $\bar{x}$ to outliers. Since the standard deviation is based on measurements from the mean, s is also heavily influenced by outliers. (In fact, the effect of outliers on s can be especially severe, since each deviation is squared during computation.) It is therefore desirable to create a measure of variability that is "resistant" to the presence of a few outliers, analogous to the median.

DEFINITION Order the n observations from smallest to largest, and separate the lower half from the upper half; the median is included in both halves if n is odd. The **lower quartile** (or **first quartile**), q_1, is the median of the lower half of the data, and the **upper quartile** (or **third quartile**), q_3, is the median of the upper half.[1]

A measure of spread that is resistant to outliers is the **interquartile range (iqr)**, given by

$$\text{iqr} = q_3 - q_1$$

The term *quartile* comes from the fact that the lower quartile divides the smallest quarter of observations from the remainder of the data set, while the upper quartile separates the top quarter of values from the rest. The interquartile range is unaffected by observations in the smallest 25% or the largest 25% of the data—hence, it is robust against (resistant to) outliers. Roughly speaking, we can interpret the iqr as the range of the "middle 50%" of the observations.

Example 1.16 Consider the ordered fuel efficiency data from Example 1.14:

$$21.5 \quad 23.4 \quad 25.4 \quad 27.0 \quad 27.8 \quad | \quad 30.0 \quad 30.1 \quad 31.0 \quad 34.5 \quad 38.3$$

The vertical line separates the two halves of the data; the median efficiency is $\tilde{x} = (27.8 + 30.0)/2 = 28.9$ mpg, coincidentally exactly the same as the mean. The quartiles are the middle values of the two halves; from the displayed data, we see that

$$q_1 = 25.4 \quad q_3 = 31.0 \quad \Rightarrow \quad \text{iqr} = 31.0 - 25.4 = 5.6 \, \text{mpg}$$

The software package R reports the upper and lower quartiles to be 25.8 and 30.775, respectively, while JMP and Minitab both give 24.9 and 31.875.

---

[1]Different software packages calculate the quartiles (and, thus, the iqr) somewhat differently, for example using different interpolation methods between x values. For smaller data sets, the difference can be noticeable; this is typically less of an issue for larger data sets.

   Imagine that the lowest value had been 10.5 instead of 21.5 (indicating something very wrong with that particular Camry!). Then the sample standard deviation would explode from 5.04 mpg (see Example 1.14) to 7.46 mpg, a nearly 50% increase. Meanwhile, the quartiles and the iqr would not change at all; those quantities would be unaffected by this low outlier. ∎

   The quartiles and interquartile range lead to a popular statistical convention for defining outliers (i.e., unusual observations) first proposed by renowned statistician John Tukey.

DEFINITION	Any observation farther than 1.5iqr from the closest quartile is an **outlier**. An outlier is **extreme** if it is more than 3iqr from the nearest quartile, and it is **mild** otherwise.

That is, outliers are defined to be all x values in the sample that satisfy either

$$x < q_1 - 1.5\text{iqr} \quad \text{or} \quad x > q_3 + 1.5\text{iqr}$$

Boxplots

In Section 1.2, several graphical displays (stem-and-leaf, dotplot, histogram) were introduced as tools for visualizing quantitative data. We now introduce one more graph, the *boxplot*, which relies on the quartiles, iqr, and aforementioned outlier rule. A boxplot shows several of a data set's most prominent features, including center, spread, the extent and nature of any departure from symmetry, and outliers.

Constructing a Boxplot

1. Draw a measurement scale (horizontal or vertical).
2. Draw a rectangle adjacent to this axis beginning at q_1 and ending at q_3 (so rectangle length = iqr).
3. Place a line segment at the location of the median. (The position of the median symbol relative to the two edges conveys information about the skewness of the middle 50% of the data.)
4. Determine which data values, if any, are outliers. Mark each outlier individually. (We may use different symbols for mild and extreme outliers; most statistical software packages do not make a distinction.)
5. Finally, draw "whiskers" out from either end of the rectangle to the smallest and largest observations *that are not outliers*.

Example 1.17 The Clean Water Act and subsequent amendments require that all waters in the USA meet specific pollution reduction goals to ensure that water is "fishable and swimmable." The article "Spurious Correlation in the USEPA Rating Curve Method for Estimating Pollutant Loads" (*J. Environ. Engr.* 2008: 610–618) investigated various techniques for estimating pollutant loads in watersheds; the authors discuss "the imperative need to use sound statistical methods" for this purpose. Among the data considered is the following sample of total nitrogen loads (TN, in kg of nitrogen/day) from a particular Chesapeake Bay location, displayed here in increasing order.

9.69	13.16	17.09	18.12	23.70	24.07	24.29	26.43
30.75	31.54	35.07	36.99	40.32	42.51	45.64	48.22
49.98	50.06	55.02	57.00	58.41	61.31	64.25	65.24
66.14	67.68	81.40	90.80	92.17	92.42	100.82	101.94
103.61	106.28	106.80	108.69	114.61	120.86	124.54	143.27
143.75	149.64	167.79	182.50	192.55	193.53	271.57	292.61
312.45	352.09	371.47	444.68	460.86	563.92	690.11	826.54
1529.35							

Relevant summary quantities are

$$\tilde{x} = 92.17 \qquad q_1 = 45.64 \qquad q_3 = 167.79$$
$$\text{iqr} = 122.15 \quad 1.5\text{iqr} = 183.225 \quad 3\text{iqr} = 366.45$$

Again, software packages may report slightly different values. Subtracting 1.5iqr from the lower quartile gives a negative number, and none of the observations are negative, so there are no outliers on the lower end of the data. However,

$$q_3 + 1.5\text{iqr} = 351.015 \quad \text{and} \quad q_3 + 3\text{iqr} = 534.24$$

Thus the four largest observations—563.92, 690.11, 826.54, and 1529.35—are extreme outliers, and 352.09, 371.47, 444.68, and 460.86 are mild outliers.

The whiskers in the boxplot in Figure 1.17 extend out to the smallest observation 9.69 on the low end and 312.45, the largest observation that is not an outlier, on the upper end. There is some positive skewness in the middle half of the data (the right edge of the box is somewhat further from the median line than is the left edge) and a great deal of positive skewness overall.
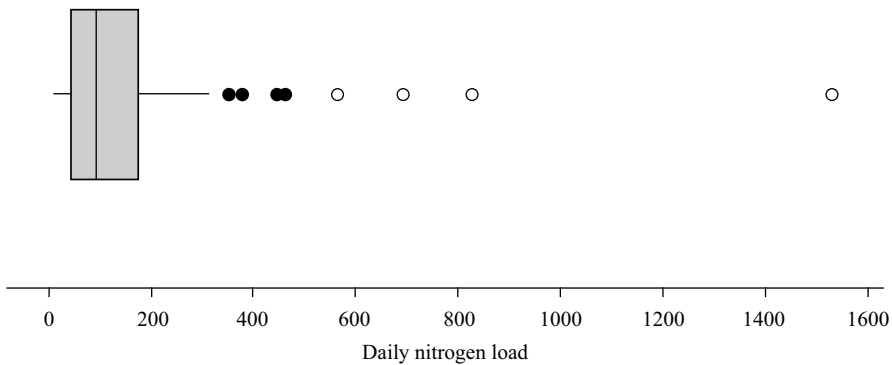
Figure 1.17 A boxplot of the nitrogen load data showing mild and extreme outliers   ∎

Placing individual boxplots side by side can reveal similarities and differences between two or more data sets consisting of observations on the same variable.

Example 1.18 Chronic kidney disease (CKD) affects vital systems throughout the body, including the production of fibrinogen, a protein that helps in the formation of blood clots. (Both too much and too little fibrinogen are dangerous.) The article "Comparison of Platelet Function and Viscoelastic Test Results between Healthy Dogs and Dogs with Naturally Occurring [CKD]" (*Amer. J. Veterinary Res.* 2017: 589-600) compared the fibrinogen levels (mg/dl of blood) in 11 dogs with CKD to 10 dogs with normal kidney function. Figure 1.18 presents a stem-and-leaf display of the data (some values were estimated from a graph in the article).

```
Healthy dogs        Dogs with CKD
          9 | 0 |
      87766 | 1 | 89
         50 | 2 | 579
          1 | 3 | 123
            | 4 | 1
            | 5 | 0
            | 6 |           Stem: Hundreds digit
            | 7 |           Leaf: Tens digit
            | 8 | 2
```

Figure 1.18 Stem-and-leaf display for Example 1.18

Numerical summary quantities are as follows:

	$\bar{x}$	$\tilde{x}$	s	iqr
Healthy	190.7	179.5	57.0	36.0
CKD	353.1	315.0	179.7	107.5

The values of the mean and median suggest that fibrinogen levels are much higher in dogs with CKD than in healthy dogs. Moreover, the variability in fibrinogen levels is much greater in the unhealthy dogs: the interquartile range for dogs with CKD (107.5 mg/dl) is nearly triple the value for healthy dogs. Figure 1.19 shows side-by-side boxplots from the JMP software package. There is obviously a systematic tendency for fibrinogen levels to be higher in the CKD group than the healthy group, and there is much more variability in the former group than in the latter one. Aside from the single outlier in each group, there is a reasonable amount of symmetry in both distributions.
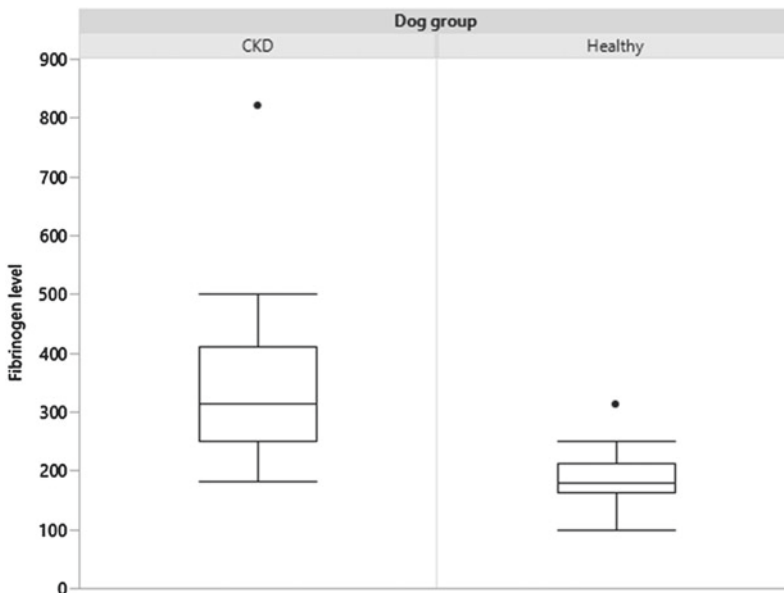
Figure 1.19 Comparative boxplots of the data in Example 1.18, from JMP

The authors of the article conclude that chronic kidney disease in dogs can lead to "hypercoagulability" (i.e., overclotting), which presents very serious health risks.   ■

Exercises: Section 1.4 (52–72)

52. Here is the data on fibrinogen levels (mg/dl) for 10 healthy dogs and 11 dogs with chronic kidney disease discussed in Example 1.18:

Healthy:　99　160　165　170　178　181　190　201
　　　　　250　313
CKD:　　183　190　250　275　290　315　320　330
　　　　　410　500　821

　a. For the data on the 10 healthy dogs, calculate the range, variance, standard deviation, quartiles, and interquartile range.
　b. Repeat part (a) for the 11 dogs with CKD.

53. The article "Oxygen Consumption During Fire Suppression: Error of Heart Rate Estimation" (*Ergonomics* 1991: 1469–1474) reported the following data on oxygen consumption (mL/kg/min) for a sample of ten firefighters performing a fire-suppression simulation:

29.5　49.3　30.6　28.2　28.0　26.3　33.9　29.4　23.5　31.6

　Compute the following:

　a. The sample range
　b. The sample variance s^2 from the definition (by first computing deviations, then squaring them, etc.)
　c. The sample standard deviation
　d. s^2 using the shortcut method

54. The value of Young's modulus (GPa) was determined for cast plates consisting of certain intermetallic substrates, resulting in the following sample observations ("Strength and Modulus of a Molybdenum-Coated Ti-25Al-10Nb-3U-1Mo Intermetallic," *J. Mater. Engr. Perform.* 1997: 46–50):

　116.4　115.9　114.6　115.2　115.8

　a. Calculate $\bar{x}$ and the deviations from the mean.
　b. Use the deviations calculated in part (a) to obtain the sample variance and the sample standard deviation.

　c. Calculate s^2 by using the computational formula for the numerator S_{xx}.
　d. Subtract 100 from each observation to obtain a sample of transformed values. Now calculate the sample variance of these transformed values, and compare it to s^2 for the original data. State the general principle.

55. The accompanying observations on stabilized viscosity (cP) for specimens of a certain grade of asphalt with 18% rubber added are from the article "Viscosity Characteristics of Rubber-Modified Asphalts" (*J. Mater. Civil Engr.* 1996: 153–156):

2781　2900　3013　2856　2888

　a. What are the values of the sample mean and sample median?
　b. Calculate the sample variance using the computational Formula. [*Hint*: First subtract a convenient number from each observation.]

56. Calculate and interpret the values of the sample median, sample mean, and sample standard deviation for the following observations on fracture strength (MPa, read from a graph in "Heat-Resistant Active Brazing of Silicon Nitride: Mechanical Evaluation of Braze Joints," *Welding J.*, Aug. 1997):

87　93　96　98　105　114　128　131　142　168

57. Exercise 46 in Section 1.3 presented a sample of 26 escape times for oil workers in a simulated exercise. Calculate and interpret the sample standard deviation. [*Hint*: $\sum x_i = 9638$ and $\sum x_i^2 = 3{,}587{,}566$.]

58. Acrylamide is a potential carcinogen that forms in certain foods, such as potato chips and French fries. The FDA analyzed McDonald's French fries purchased at seven different locations; the following are the resulting acrylamide levels (in micrograms per kg of food):

$$497 \quad 193 \quad 328 \quad 155 \quad 326 \quad 245 \quad 270$$

Calculate $\sum x_i$ and $\sum x_i^2$ and then s^2 and s.

59. In 1997 a woman sued a computer keyboard manufacturer, charging that her repetitive stress injuries were caused by the keyboard (*Genessy v. Digital Equipment Corp.*). The jury awarded about $3.5 million for pain and suffering, but the court then set aside that award as being unreasonable compensation. In making this determination, the court identified a "normative" group of 27 similar cases and specified a reasonable award as one within two standard deviations of the mean of the awards in the 27 cases. The 27 awards (in $1000s) were

37	60	75	115	135	140	149	150	238
290	340	410	600	750	750	750	1050	1100
1139	1150	1200	1200	1250	1576	1700	1825	2000

from which $\sum x_i = 20{,}179$, $\sum x_i^2 = 24{,}657{,}511$. What is the maximum possible amount that could be awarded under the two standard deviation rule?

60. The US Women's Swimming Team won the 1500 m relay at the 2016 Olympic Games. Here are the completion times, in seconds, for all eight teams that competed in the finals:

$$233.13 \quad 235.00 \quad 235.01 \quad 235.18$$
$$235.49 \quad 235.66 \quad 236.96 \quad 239.50$$

   a. Calculate the sample variance and standard deviation.
   b. If the observations were re-expressed in minutes, what would be the resulting values of the sample variance and sample standard deviation? Answer without actually performing the reexpression.

61. The first four deviations from the mean in a sample of $n = 5$ reaction times were .3, .9, 1.0, and 1.3. What is the fifth deviation from the mean? Give a sample for which these are the five deviations from the mean.

62. Reconsider the data on recent home sales (in $1000s) provided in Exercise 40:

$$525 \quad 830 \quad 600 \quad 180 \quad 129$$
$$525 \quad 350 \quad 490 \quad 640 \quad 475$$

   a. Determine the upper and lower quartiles, and then the iqr.
   b. If the two largest sample values, 830 and 640, had instead been 930 and 740, how would this affect the iqr? Explain.
   c. By how much could the observation 129 be increased without affecting the iqr? Explain.
   d. If an 11th observation, $x_{11} = 845$, is added to the sample, what now is the iqr?

63. Reconsider the court awards data presented in Exercise 59.

   a. What are the values of the quartiles, and what is the value of the iqr?
   b. How large or small does an observation have to be to qualify as an outlier? As an extreme outlier?
   c. Construct a boxplot, and comment on its features.

64. Here is a stem-and-leaf display of the escape time data introduced in Exercise 46.

32	55
33	49
34	
35	6699
36	34469
37	03345
38	9
39	2347
40	23
41	
42	4

   a. Determine the value of the interquartile range.
   b. Are there any outliers in the sample? Any extreme outliers?
   c. Construct a boxplot and comment on its features.
   d. By how much could the largest observation, currently 424, be decreased without affecting the value of the iqr?

65. Many people who believe they may be suffering from the flu visit emergency rooms, where they are subjected to long waits and may expose others or themselves be exposed to various diseases. The article "Drive-Through Medicine: A Novel Proposal for the Rapid Evaluation of Patients During an Influenza Pandemic" (*Annals Emerg. Med.* 2010: 268–273) described an experiment to see whether patients could be evaluated while remaining in their vehicles. The following total processing times (min) for 38 individuals were read from a graph that appeared in the cited article:

9	16	16	17	19	20	20	20
23	23	23	23	24	24	24	24
25	25	26	26	27	27	28	28
29	29	29	30	32	33	33	34
37	43	44	46	48	53		

a. Calculate several different measures of center and compare them.
b. Are there any outliers in this sample? Any extreme outliers?
c. Construct a boxplot and comment on any interesting features.

66. Here is summary information on the alcohol percentage for a sample of 25 beers:

$$q_1 = 4.35 \quad \tilde{x} = 5 \quad q_3 = 5.95$$

The bottom three are 3.20 (Heineken Premium Light), 3.50 (Amstel light), 4.03 (Shiner Light) and the top three are 7.50 (Terrapin All-American Imperial Pilsner), 9.10 (Great Divide Hercules Double IPA), 11.60 (Rogue Imperial Stout).

a. Are there any outliers in the sample? Any extreme outliers?
b. Construct a boxplot, and comment on any interesting features.

67. A company utilizes two different machines to manufacture parts of a certain type. During a single shift, a sample of $n = 20$ parts produced by each machine is obtained, and the value of a particular critical dimension for each part is determined. The comparative boxplot below is constructed from the resulting data. Compare and contrast the two samples.
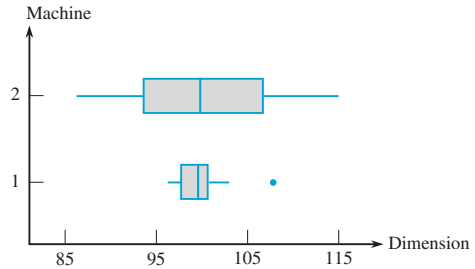
68. Blood cocaine concentration (mg/L) was determined both for a sample of individuals who had died from cocaine-induced excited delirium (ED) and for a sample of those who had died from a cocaine overdose without excited delirium; survival time for people in both groups was at most 6 h. The accompanying data was read from a comparative boxplot in the article "Fatal Excited Delirium Following Cocaine Use" (*J. Forensic Sci.* 1997: 25–31).

ED	0	0	0	0	.1	.1
	.1	.1	.2	.2	.3	.3
	.3	.4	.5	.7	.8	1.0
	1.5	2.7	2.8	3.5	4.0	8.9
	9.2	11.7	21.0			

Non-ED	0	0	0	0	0	.1
	.1	.1	.1	.2	.2	.2
	.3	.3	.3	.4	.5	.5
	.6	.8	.9	1.0	1.2	1.4
	1.5	1.7	2.0	3.2	3.5	4.1
	4.3	4.8	5.0	5.6	5.9	6.0
	6.4	7.9	8.3	8.7	9.1	9.6
	9.9	11.0	11.5	12.2	12.7	14.0
	16.6	17.8				

a. Determine the medians, quartiles, and iqrs for the two samples.
b. Are there any outliers in either sample? Any extreme outliers?
c. Construct a comparative boxplot, and use it as a basis for comparing and contrasting the ED and non-ED samples.

69. At the beginning of the 2007 baseball season each American League team had nine starting position players (this includes the designated hitter but not the pitcher). Here are the salaries for the New York Yankees and the Cleveland Indians in thousands of dollars:

Yankees:	12000	600	491	22709	21600
	13000	13000	15000	23429	
Indians:	3200	3750	396	383	1000
	3750	917	3000	4050	

Construct a comparative boxplot and comment on interesting features. Compare the salaries of the two teams. (The Indians won more games than the Yankees in the regular season and defeated the Yankees in the playoffs.)

70. The article "E-cigarettes as a Source of Toxic and Potentially Carcinogenic Metals" (*Environ. Res.* 2017: 221–225) reports the concentration (μg/L) of cadmium, chromium, lead, manganese, and nickel in 10 cartiges for each of five e-cigarette brands. Here are the lead levels in the 50 cartridges (some values were estimated from a graph in the article):

Brand					
A	500	623	794	1228	1555
	1705	2190	3162	3894	4870
B	3.53	3.67	3.98	10.2	16.4
	20.6	34.0	49.1	126	218
C	7.94	14.3	23.8	44.2	59.3
	79.3	156	204	219	233
D	3.17	3.45	4.21	4.56	4.95
	5.01	5.23	5.34	5.68	5.89
E	4.50	4.89	4.99	5.02	5.06
	5.24	6.43	7.09	8.52	9.82

Because the values are on very different scales, it makes sense to take the logarithms of these values first. Apply a $\log_{10}()$ transformation to these values, construct a comparative boxplot, and comment on what you find.

71. The comparative boxplot below of gasoline vapor coefficients for vehicles in Detroit appeared in the article "Receptor Modeling Approach to [Volatile Organic Compound] Emission Inventory Validation" (*J. Environ. Engr.* 1995: 483–490). Discuss any interesting features.
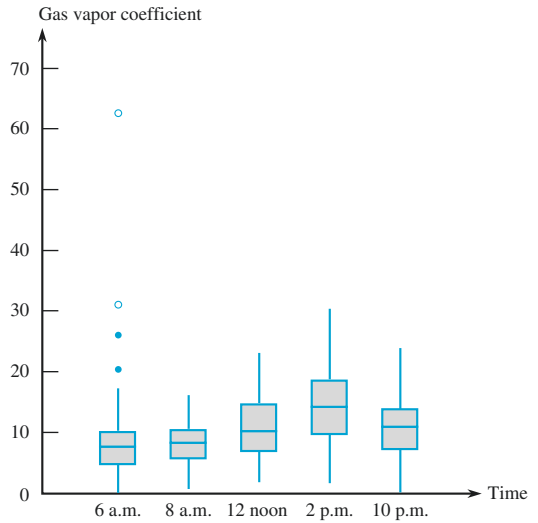
72. Let $x_1, \ldots, x_n$ be a sample, and let a and b be constants. Define a new sample $y_1, \ldots, y_n$ by $y_1 = ax_1 + b$, ..., $y_n = ax_n + b$.

a. How do the sample variance and standard deviation of the y_i's relate to the variance and standard deviation of the x_i's? Verify your conjectures.

b. How does the iqr of the y_i's relate to the iqr of the x_i's? Substantiate your assertion.

Supplementary Exercises: (73–96)

73. The article "Correlation Analysis of Stenotic Aortic Valve Flow Patterns Using Phase Contrast MRI" (*Annals of Biomed. Engr.* 2005: 878–887) included the following data on aortic root diameter (cm) for a sample of patients having various degrees of aortic stenosis (i.e., narrowing of the aortic valve):

Males:	3.7	3.4	3.7	4.0	3.9	3.8	3.4	3.6	3.1
	4.0	3.4	3.8	3.5					
Females:	3.8	2.6	3.2	3.0	4.3	3.5	3.1	3.1	
	3.2	3.0							

    a. Create a comparative stem-and-leaf plot.

    b. Calculate an appropriate measure of center for each set of observations.

    c. Compare and contrast the diameter observations for the two sexes.

74. Consider the following information from a sample of four Wolferman's cranberry citrus English muffins, which are said on the package label to weigh 116 g: $\bar{x} = 104.4$ g, $s = 4.1497$ g, smallest weighs 98.7 g, largest weighs 108.0 g. Determine the values of the two middle sample observations (and don't do it by successive guessing!).

75. Three different C_2F_6 flow rates (SCCM) were considered in an experiment to investigate the effect of flow rate on the uniformity (%) of the etch on a silicon wafer used in the manufacture of integrated circuits, resulting in the following data:

    Flow rate

125	2.6	2.7	3.0	3.2	3.8	4.6
160	3.6	4.2	4.2	4.6	4.9	5.0
200	2.9	3.4	3.5	4.1	4.6	5.1

    Compare and contrast the uniformity observations resulting from these three different flow rates.

76. The amount of radiation received at a greenhouse plays an important role in determining the rate of photosynthesis. The accompanying observations on incoming solar radiation were read from a graph in the article "Radiation Components over Bare and Planted Soils in a Greenhouse" (*Solar Energy* 1990: 1011–1016).

6.3	6.4	7.7	8.4	8.5	8.8	8.9
9.0	9.1	10.0	10.1	10.2	10.6	10.6
10.7	10.7	10.8	10.9	11.1	11.2	11.2
11.4	11.9	11.9	12.2	13.1		

    Use some of the methods discussed in this chapter to describe and summarize this data.

77. The article "Motor Vehicle Emissions Variability" (*J. Air Waste Manag. Assoc.* 1996: 667–675) reported the following hydrocarbon and carbon dioxide measurements using the Federal Testing Procedure for emissions-testing, applied four times each to the same car:

HC (g/mile):	13.8	18.3	32.2	32.5
CO (g/mile):	118	149	232	236

    a. Compute the sample standard deviations for the HC and CO observations. Why should it not be surprising that the CO measurements have a larger standard deviation?

    b. The *sample coefficient of variation* $s/\bar{x}$ (or $100 \cdot s/\bar{x}$) assesses the extent of variability relative to the mean. Values of this coefficient for several different data sets can be compared to determine which data sets exhibit more or less variation. Carry out such a comparison for the given data.

78. The *cost-to-charge ratio* for a hospital is the ratio of the actual cost of care to what the hospital charges for that care. In 2008, the Kentucky Department of Health and Family Services reported the following cost-to-charge ratios, expressed as percents, for 116 Kentucky hospitals:

52.9	49.7	58.1	41.4	66.5	44.1	53.0	49.1
59.8	47.1	44.3	52.3	60.5	59.9	47.1	62.4
47.3	62.1	52.1	47.8	65.1	42.9	38.5	65.9
51.3	52.6	44.9	47.8	60.2	56.4	67.6	31.9
53.9	50.6	72.5	47.8	50.5	25.1	45.0	86.0
53.7	61.2	63.4	51.5	48.6	42.1	49.3	50.0
66.4	64.6	47.4	48.1	45.8	64.7	58.7	56.9
45.9	82.9	46.0	51.0	67.0	49.3	69.5	56.5
55.0	39.2	85.0	46.7	41.6	45.4	71.2	42.7
46.9	39.2	55.3	46.1	43.2	67.7	60.6	68.2
81.6	39.2	54.7	63.5	67.9	50.9	40.4	49.0
54.4	39.2	43.2	43.2	51.7	48.4	50.7	59.4
49.7	60.2	40.2	62.3	41.4	48.6	45.6	46.2
51.4	65.3	31.5	50.6	41.4	82.3	45.2	46.0
58.3	46.3	38.2	59.1				

    (For example, a cost-to-charge ratio of 53.0% means the actual cost of care is 53% of what the hospital charges.) Use various techniques discussed in this chapter to organize, summarize, and describe the data.

79. Fifteen air samples from a certain region were obtained, and for each one the carbon monoxide concentration was determined. The results (in ppm) were

| 9.3 | 10.7 | 8.5 | 9.6 | 12.2 | 15.6 | 9.2 | 10.5 |
| 9.0 | 13.2 | 11.0 | 8.8 | 13.7 | 12.1 | 9.8 | |

Using the interpolation method suggested in Section 1.3, compute the 10% trimmed mean.

80. a. For what value of c is the quantity $\sum (x_i - c)^2$ minimized? [*Hint*: Take the derivative with respect to c, set equal to 0, and solve.]
   b. Using the result of part (a), which of the two quantities $\sum (x_i - \bar{x})^2$ and $\sum (x_i - \mu)^2$ will be smaller than the other (assuming that $\bar{x} \neq \mu$)?

81. The article "A Longitudinal Study of the Development of Elementary School Children's Private Speech" (*Merrill-Palmer Q.* 1990: 443–463) reported on a study of children talking to themselves (private speech). It was thought that private speech would be related to IQ, because IQ is supposed to measure mental maturity, and it was known that private speech decreases as students progress through the primary grades. The study included 33 students whose first-grade IQ scores are given here:

82	96	99	102	103	103	106	107	108	108
108	108	109	110	110	111	113	113	113	113
115	115	118	118	119	121	122	122	127	132
136	140	146							

Use various techniques discussed in this chapter to organize, summarize, and describe the data.

82. The accompanying specific gravity values for various wood types used in construction appeared in the article "Bolted Connection Design Values Based on European Yield Model" (*J. Struct. Engr.* 1993: 2169–2186):

.31	.35	.36	.36	.37	.38	.40	.40	.40
.41	.41	.42	.42	.42	.42	.42	.43	.44
.45	.46	.46	.47	.48	.48	.48	.51	.54
.54	.55	.58	.62	.66	.66	.67	.68	.75

Construct a stem-and-leaf display using repeated stems, and comment on any interesting features of the display.

83. In recent years, some evidence suggests that high indoor radon concentration may be linked to the development of childhood cancers, but many health professionals remain unconvinced. The article "Indoor Radon and Childhood Cancer" (*Lancet* 1991: 1537–1538) presented the accompanying data on radon concentration (Bq/m^3) in two different samples of houses. The first sample consisted of houses in which a child diagnosed with cancer had been residing. Houses in the second sample had no recorded cases of childhood cancer.

Cancer:	3	5	6	7	8	9	9	10	10	10
	11	11	11	11	12	13	13	15	15	15
	16	16	16	17	18	18	18	20	21	21
	22	22	23	23	27	33	34	38	39	45
	57	210								
No cancer:	3	3	5	6	6	7	7	7	8	8
	9	9	9	9	11	11	11	11	11	12
	12	13	14	17	17	21	21	24	24	29
	29	29	29	33	38	39	55	55	85	

   a. Construct a side-by-side stem-and-leaf display, and comment on any interesting features.
   b. Calculate the standard deviation of each sample. Which sample appears to have greater variability, according to these values?
   c. Calculate the iqr for each sample. Now which sample has greater variability, and why is this different than the result of part (b)?

84. Elevated energy consumption during exercise continues after the workout ends. Because calories burned after exercise contribute to weight loss and have other consequences, it is important to understand this process. The paper "Effect of Weight Training Exercise and Treadmill Exercise on Post-Exercise Oxygen Consumption" (*Med. Sci. Sports Exercise* 1998: 518–522) reported the accompanying data from a study in which oxygen consumption (liters) was measured continuously for 30 min for each of 15 subjects both after a weight training exercise and after a treadmill exercise.

a. Construct side-by-side boxplots of the weight and treadmill observations, and comment on what you see.

b. Because the data is in the form of (x, y) pairs, with x and y measurements on the same variable under two different conditions, it is natural to focus on the differences within pairs: $d_1 = x_1 - y_1$, $\ldots, d_n = x_n - y_n$. Construct a boxplot of the sample differences. What does it suggest?

Subject	1	2	3	4	5	6
Weight (x)	14.6	14.4	19.5	24.3	16.3	22.1
Treadmill (y)	11.3	5.3	9.1	15.2	10.1	19.6

Subject	7	8	9	10	11	12
Weight (x)	23.0	18.7	19.0	17.0	19.1	19.6
Treadmill (y)	20.8	10.3	10.3	2.6	16.6	22.4

Subject	13	14	15
Weight (x)	23.2	18.5	15.9
Treadmill (y)	23.6	12.6	4.4

85. Anxiety disorders and symptoms can often be effectively treated with benzodiazepine medications. It is known that animals exposed to stress exhibit a decrease in benzodiazepine receptor binding in the frontal cortex. The paper "Decreased Benzodiazepine Receptor Binding in Prefrontal Cortex in Combat-Related Posttraumatic Stress Disorder" (*Amer. J. Psychiatry* 2000: 1120–1126) described the first study of benzodiazepine receptor binding in individuals suffering from PTSD. The accompanying data on a receptor binding measure (adjusted distribution volume) was read from a graph in the paper.

PTSD: 10 20 25 28 31 35 37 38 38 39 39 42 46
Healthy: 23 39 40 41 43 47 51 58 63 66 67 69 72

Use various methods from this chapter to describe and summarize the data.

86. The article "Can We Really Walk Straight?" (*Amer. J. Phys. Anthropol.* 1992: 19–27) reported on an experiment in which each of 20 healthy men was asked to walk as straight as possible to a target 60 m away at normal speed. Consider the following

observations on cadence (number of strides per second):

.95   .85   .92   .95   .93   .86   1.00   .92   .85   .81
.78   .93   .93   1.05   .93   1.06   1.06   .96   .81   .96

Use the methods developed in this chapter to summarize the data; include an interpretation or discussion wherever appropriate. [*Note*: The author of the article used a rather sophisticated statistical analysis to conclude that people cannot walk in a straight line and suggested several explanations for this.]

87. The **mode** of a numerical data set is the value that occurs most frequently in the set.

a. Determine the mode for the cadence data given in the previous exercise.

b. For a categorical sample, how would you define the modal category?

88. Specimens of three different types of rope wire were selected, and the fatigue limit (MPa) was determined for each specimen, resulting in the accompanying data.

Type 1:	350	350	350	358	370	370	370	371
	371	372	372	384	391	391	392	
Type 2:	350	354	359	363	365	368	369	371
	373	374	376	380	383	388	392	
Type 3:	350	361	362	364	364	365	366	371
	377	377	377	379	380	380	392	

a. Construct a comparative boxplot, and comment on similarities and differences.

b. Construct a comparative dotplot (a dotplot for each sample with a common scale). Comment on similarities and differences.

c. Does the comparative boxplot of part (a) give an informative assessment of similarities and differences? Explain your reasoning.

89. The three measures of center introduced in this chapter are the mean, median, and trimmed mean. Two additional measures of center that are occasionally used are the *midrange*, which is the average of the smallest and largest observations, and the

midquarter, which is the average of the two quartiles. Which of these five measures of center are resistant to the effects of outliers and which are not? Explain your reasoning.

90. The authors of the article "Predictive Model for Pitting Corrosion in Buried Oil and Gas Pipelines" (*Corrosion* 2009: 332–342) provided the data on which their investigation was based.

    a. Consider the following sample of 61 observations on maximum pitting depth (mm) of pipeline specimens buried in clay loam soil.

0.41	0.41	0.41	0.41	0.43	0.43	0.43	0.48	0.48
0.58	0.79	0.79	0.81	0.81	0.81	0.91	0.94	0.94
1.02	1.04	1.04	1.17	1.17	1.17	1.17	1.17	1.17
1.17	1.19	1.19	1.27	1.40	1.40	1.59	1.59	1.60
1.68	1.91	1.96	1.96	1.96	2.10	2.21	2.31	2.46
2.49	2.57	2.74	3.10	3.18	3.30	3.58	3.58	4.15
4.75	5.33	7.65	7.70	8.13	10.41	13.44		

    Construct a stem-and-leaf display in which the two largest values are shown in a last row labeled HI.

    b. Refer back to (a), and create a histogram based on eight classes with 0 as the lower limit of the first class and class widths of .5, .5, .5, .5, 1, 2, 5, and 5, respectively.

    c. The accompanying comparative boxplot shows plots of pitting depth for four different types of soils. Describe its important features.
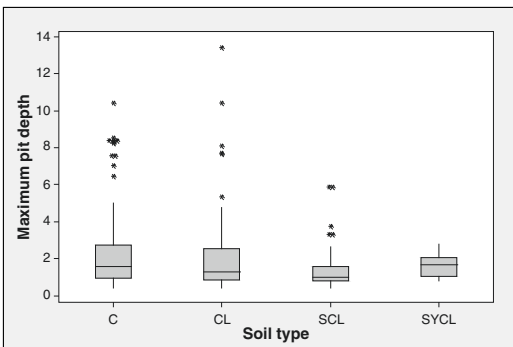
91. Consider a sample $x_1, x_2, \ldots, x_n$ and suppose that the values of $\bar{x}$, s^2, and s have been calculated.

    a. Let $y_i = x_i - \bar{x}$ for $i = 1, \ldots, n$. How do the values of s^2 and s for the y_i's compare to the corresponding values for the x_i's? Explain.

    b. Let $z_i = (x_i - \bar{x})/s$ for $i = 1, \ldots, n$. What are the values of the sample variance and sample standard deviation for the z_i's?

92. Let $\bar{x}_n$ and s_n^2 denote the sample mean and variance for the sample $x_1, \ldots, x_n$ and let $\bar{x}_{n+1}$ and s_{n+1}^2 denote these quantities when an additional observation x_{n+1} is added to the sample.

    a. Show how $\bar{x}_{n+1}$ can be computed from $\bar{x}_n$ and x_{n+1}.

    b. Show that

    $$ns_{n+1}^2 = (n-1)s_n^2 + \frac{n}{n+1}(x_{n+1} - \bar{x}_n)^2$$

    so that s_{n+1}^2 can be computed from x_{n+1}, $\bar{x}_n$, and s_n^2.

    c. Suppose that a sample of 15 strands of drapery yarn has resulted in a sample mean thread elongation of 12.58 mm and a sample standard deviation of .512 mm. A 16th strand results in an elongation value of 11.8. What are the values of the sample mean and sample standard deviation for all 16 elongation observations?

93. Lengths of bus routes for any particular transit system will typically vary from one route to another. The article "Planning of City Bus Routes" (*J. Institut. Engr.* 1995: 211–215) gives the following information on lengths (km) for one particular system:

Length:	6–8	8–10	10–12	12–14	14–16
Frequency:	6	23	30	35	32
Length:	16–18	18–20	20–22	22–24	24–26
Frequency:	48	42	40	28	27
Length:	26–28	28–30	30–35	35–40	40–45
Frequency:	26	14	27	11	2

a. Draw a histogram corresponding to these frequencies.

b. What proportion of these route lengths are less than 20? What proportion of these routes have lengths of at least 30?

c. Roughly what is the value of the 90th percentile of the route length distribution?

d. Roughly what is the median route length?

94. A study carried out to investigate the distribution of total braking time (reaction time plus accelerator-to-brake movement time, in msec) during real driving conditions at 60 km/h gave the following summary information on the distribution of times ("A Field Study on Braking Responses during Driving," *Ergonomics* 1995: 1903–1910):

mean = 535   median = 500   mode = 500
sd = 96   minimum = 220   maximum = 925
5th percentile = 400   10th percentile = 430
90th percentile = 640   95th percentile = 720

What can you conclude about the shape of a histogram of this data? Explain your reasoning.

95. The sample data x_1, x_2, …, x_n sometimes represents a **time series**, where x_t = the observed value of a response variable x at time t. Often the observed series shows a great deal of random variation, which makes it difficult to study longer-term behavior. In such situations, it is desirable to produce a smoothed version of the series. One technique for doing so involves **exponential smoothing**. The value of a smoothing constant α is chosen ($0 < \alpha < 1$). Then with $\bar{x}_t$ defined as the smoothed value at time t, we set $\bar{x}_1 = x_1$, and for $t = 2, 3, …, n$, $\bar{x}_t = \alpha x_t + (1 - \alpha)\bar{x}_{t-1}$.

a. Consider the following time series in which x_t = temperature (°F) of effluent at a sewage treatment plant on day t: 47,

54, 53, 50, 46, 46, 47, 50, 51, 50, 46, 52, 50, 50. Plot each x_t against t on a two-dimensional coordinate system (a time series plot). Does there appear to be any pattern?

b. Calculate the $\bar{x}_t$'s using $\alpha = .1$. Repeat using $\alpha = .5$. Which value of α gives a smoother $\bar{x}_t$ series?

c. Substitute $\bar{x}_{t-1} = \alpha x_{t-1} + (1 - \alpha)\bar{x}_{t-2}$ on the right-hand side of the expression for $\bar{x}_t$, then substitute $\bar{x}_{t-2}$ in terms of x_{t-2} and $\bar{x}_{t-3}$, and so on. On how many of the values x_t, x_{t-1}, …, x_1 does $\bar{x}_t$ depend? What happens to the coefficient on x_{t-k} as k increases?

d. Refer to part (c). If t is large, how sensitive is $\bar{x}_t$ to the initialization $\bar{x}_1 = x_1$? Explain.

96. Consider numerical observations $x_1, …, x_n$. It is frequently of interest to know whether the x_i's are (at least approximately) symmetrically distributed about some value. If n is at least moderately large, the extent of symmetry can be assessed from a stem-and-leaf display or histogram. However, if n is not very large, such pictures are not particularly informative. Consider the following alternative. Let y_1 denote the smallest x_i, y_2 the second-smallest x_i, and so on. Then plot the following pairs as points on a two-dimensional coordinate system: $(y_n - \tilde{x}, \tilde{x} - y_1)$, $(y_{n-1} - \tilde{x}, \tilde{x} - y_2)$, $(y_{n-2} - \tilde{x}, \tilde{x} - y_3)$, …. There are $n/2$ points when n is even and $(n - 1)/2$ when n is odd.

a. What does this plot look like when there is perfect symmetry in the data? What does it look like when observations stretch out more above the median than below it (a long upper tail)?

b. Construct the plot for the nitrogen data presented in Example 1.17, and comment on the extent of symmetry or nature of departure from symmetry.

Probability

Introduction

The term **probability** refers to the study of randomness and uncertainty. In any situation in which one of a number of possible outcomes may occur, the theory of probability provides methods for quantifying the chances, or likelihoods, associated with the various outcomes. The language of probability is constantly used in an informal manner in both written and spoken contexts. Examples include such statements as "It is likely that the Dow Jones Industrial Average will increase by the end of the year," "There is a 50–50 chance that the incumbent will seek reelection," "There will probably be at least one section of that course offered next year," "The odds favor a quick settlement of the strike," and "It is expected that at least 20,000 concert tickets will be sold." In this chapter, we introduce some elementary probability concepts, indicate how probabilities can be interpreted, and show how the rules of probability can be applied to compute the probabilities of many interesting events. The methodology of probability will then permit us to express in precise language such informal statements as those given above.

The study of probability as a branch of mathematics goes back over 300 years, where it had its genesis in connection with questions involving games of chance. Many books are devoted exclusively to probability and explore in great detail numerous interesting aspects and applications of this lovely branch of mathematics. Our objective here is more limited in scope: We will focus on those topics that are central to a basic understanding and also have the most direct bearing on problems of statistical inference.

2.1 Sample Spaces and Events

In probability, an **experiment** refers to any action or activity whose outcome is subject to uncertainty. Although the word *experiment* generally suggests a planned or carefully controlled laboratory testing situation, we use it here in a much wider sense. Thus experiments that may be of interest include tossing a coin once or several times, selecting a card or cards from a deck, weighing a loaf of bread, measuring the commute time from home to work on a particular morning, determining blood types from a group of individuals, or calling people to conduct a survey.

The Sample Space of an Experiment

DEFINITION   The **sample space** of an experiment, denoted by $\mathcal{S}$, is the set of all possible
                     outcomes of that experiment.

Example 2.1 The simplest experiment to which probability applies is one with two possible outcomes. One such experiment consists of examining a single fuse to see whether it is defective. The sample space for this experiment can be abbreviated as $\mathcal{S} = \{N, D\}$, where N represents not defective, D represents defective, and the braces are used to enclose the elements of a set. Another such experiment would involve tossing a thumbtack and noting whether it landed point up or point down, with sample space $\mathcal{S} = \{U, D\}$, and yet another would consist of observing the sex assigned to the next child born at the local hospital, with $\mathcal{S} = \{M, F\}$.   ■

Example 2.2 If we examine three fuses in sequence and note the result of each examination, then an outcome for the entire experiment is any sequence of N's and D's of length 3, so

$\mathcal{S} = \{NNN, NND, NDN, NDD, DNN, DND, DDN, DDD\}$

If we had tossed a thumbtack three times, the sample space would be obtained by replacing N by U in $\mathcal{S}$ above. A similar notational change would yield the sample space for the experiment in which the assigned sexes of three newborn children are observed.   ■

Example 2.3 Two gas stations are located at a certain intersection. Each one has six gas pumps. Consider the experiment in which the number of pumps in use at a particular time of day is observed for each of the stations. An experimental outcome specifies how many pumps are in use at the first station and how many are in use at the second one. One possible outcome is (2, 2), another is (4, 1), and yet another is (1, 4). The 49 outcomes in $\mathcal{S}$ are displayed in the accompanying table.

First station	Second station						
	0	1	2	3	4	5	6
0	(0, 0)	(0, 1)	(0, 2)	(0, 3)	(0, 4)	(0, 5)	(0, 6)
1	(1, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 0)	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 0)	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 0)	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 0)	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 0)	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

The sample space for the experiment in which a six-sided die is thrown twice results from deleting the 0 row and 0 column from the table, giving 36 outcomes.   ■

Example 2.4 If a new cell phone battery has a voltage that is outside certain limits, that battery is characterized as a failure (F); if the battery has a voltage within the prescribed limits, it is a success (S). Suppose an experiment consists of testing each battery as it comes off an assembly line until we first observe a success. Although it may not be very likely, a possible outcome of this experiment is that the first 10 (or 100 or 1000 or …) are F's and the next one is an S. That is, for any positive integer

n, we may have to examine n batteries before seeing the first S. The sample space is $\mathcal{S} = \{S, FS, FFS, FFFS, ...\}$, which contains an infinite number of possible outcomes. The same abbreviated form of the sample space is appropriate for an experiment in which, starting at a specified time, the sex of each newborn infant at a hospital is recorded until the birth of a female is observed. ∎

Events

In our study of probability, we will be interested not only in the individual outcomes but also in various collections of outcomes from $\mathcal{S}$.

DEFINITION   An **event** is any collection (subset) of outcomes contained in the sample space $\mathcal{S}$. An event is said to be **simple** if it consists of exactly one outcome and **compound** if it consists of more than one outcome.

When an experiment is performed, a particular event A is said to occur if the resulting experimental outcome is contained in A. In general, exactly one simple event will occur, but many compound events will occur simultaneously.

Example 2.5   Consider an experiment in which each of three vehicles taking a particular freeway exit turns left (L) or right (R) at the end of the off-ramp. The eight possible outcomes that comprise the sample space are LLL, RLL, LRL, LLR, LRR, RLR, RRL, and RRR. Thus there are eight simple events, among which are $E_1 = \{LLL\}$ and $E_5 = \{LRR\}$. Some compound events include

$A = \{RLL, LRL, LLR\}$ = the event that exactly one of the three vehicles turns right
$B = \{LLL, RLL, LRL, LLR\}$ = the event that at most one of the vehicles turns right
$C = \{LLL, RRR\}$ = the event that all three vehicles turn in the same direction.

Suppose that when the experiment is performed, the outcome is LLL. Then the simple event E_1 has occurred and so also have the events B and C, but not A. ∎

Example 2.6   (Example 2.3 continued) When the number of pumps in use at each of two six-pump gas stations is observed, there are 49 possible outcomes, so there are 49 simple events: $E_1 = \{(0,0)\}$, $E_2 = \{(0,1)\}, \ldots, E_{49} = \{(6,6)\}$. Examples of compound events are

$A = \{(0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$ = the event that the number of pumps in use is the same for both stations
$B = \{(0, 4), (1, 3), (2, 2), (3, 1), (4, 0)\}$ = the event that the total number of pumps in use is four
$C = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ = the event that at most one pump is in use at each station. ∎

Example 2.7   (Example 2.4 continued) The sample space for the cell phone battery experiment contains an infinite number of outcomes, so there are an infinite number of simple events. Compound events include

$A = \{S, FS, FFS\}$ = the event that at most three batteries are examined
$B = \{S, FFS, FFFFS\}$ = the event that exactly one, three, or five batteries are examined
$C = \{FS, FFFS, FFFFFS, \ldots\}$ = the event that an even number of batteries are examined. ∎

Some Relations from Set Theory

An event is nothing but a set, so relationships and results from elementary set theory can be used to study events. The following operations will be used to construct new events from given events.

DEFINITION   1. The **complement** of an event A, denoted by A', is the set of all outcomes in $\mathscr{S}$ that are *not* contained in A.
2. The **intersection** of two events A and B, denoted by $A \cap B$ and read "A *and* B," is the event consisting of all outcomes that are in *both* A *and* B.
3. The **union** of two events A and B, denoted by $A \cup B$ and read "A *or* B," is the event consisting of all outcomes that are *either in* A *or in* B *or in both events* (so that the union includes outcomes for which both A and B occur as well as outcomes for which exactly one occurs)—that is, all outcomes in at least one of the events.

Example 2.8 (Example 2.3 continued) For the experiment in which the number of pumps in use at a *single* six-pump gas station is observed, let A = {0, 1, 2, 3, 4}, B = {3, 4, 5, 6}, and C = {1, 3, 5}. Then

$A \cup B$ = {0, 1, 2, 3, 4, 5, 6} = $\mathscr{S}$ $A \cup C$ = {0, 1, 2, 3, 4, 5}
$A \cap B$ = {3, 4} $A \cap C$ = {1, 3} A' = {5, 6} $(A \cup C)'$ = {6}   ∎

Example 2.9 (Example 2.4 continued) In the cell phone battery experiment, define A, B, and C as in Example 2.7. Then

$$A \cup B = \{S, FS, FFS, FFFFS\}$$
$$A \cap B = \{S, FFS\}$$
$$A' = \{FFFS, FFFFS, FFFFFS, \ldots\}$$

and

$C' = \{S, FFS, FFFFS, \ldots\}$ = the event that an odd number of batteries are examined.   ∎

The complement, intersection, and union operators from set theory correspond to the *not*, *and*, and *or* operators from computer science. Readers with prior programming experience may be aware of an important relationship between these three operators, first discovered by 19-century British mathematician Augustus De Morgan.

DE MORGAN'S LAWS   Let A and B be two events in the sample space of some experiment. Then

1. $(A \cup B)' = A' \cap B'$

2. $(A \cap B)' = A' \cup B'$

De Morgan's laws state that the complement of a union is an intersection, and the complement of an intersection is a union (see Exercise 11).

Sometimes A and B have no outcomes in common, so that the intersection of A and B contains no outcomes.

> **DEFINITION** When A and B have no outcomes in common, they are said to be **disjoint** or **mutually exclusive** events. Mathematicians write this compactly as $A \cap B = \varnothing$, where $\varnothing$ denotes the event consisting of no outcomes whatsoever (the "null" or "empty" event).

Example 2.10 A small city has three automobile dealerships: a GM dealer selling Chevrolets and Buicks; a Ford dealer selling Fords and Lincolns; and a Chrysler dealer selling Rams and Jeeps. If an experiment consists of observing the brand of the next vehicle sold, then the events A = {Chevrolet, Buick} and B = {Ford, Lincoln} are mutually exclusive, because the next vehicle sold cannot be both a GM product and a Ford product. ∎

Venn diagrams are often used to visually represent sample spaces and events. To construct a Venn diagram, draw a rectangle whose interior will represent the sample space $\mathscr{S}$. Then any event A is represented as the interior of a closed curve (often a circle) contained in $\mathscr{S}$. Figure 2.1 shows examples of Venn diagrams.

(a) Venn diagram of events A and B **(b)** Shaded region is $A \cap B$ **(c)** Shaded region is $A \cup B$ **(d)** Shaded region is A' **(e)** Mutually exclusive events
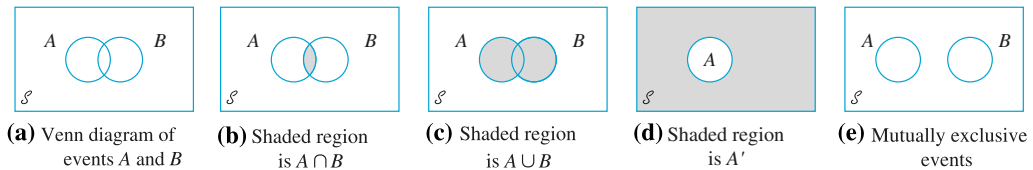
Figure 2.1 Venn diagrams

The operations of union and intersection can be extended to more than two events. For any three events A, B, and C, the event $A \cap B \cap C$ is the set of outcomes contained in *all* three events, whereas $A \cup B \cup C$ is the set of outcomes contained in *at least one* of the three events. A collection of several events is said to be mutually exclusive (or pairwise disjoint) if no two events have any outcomes in common. De Morgan's laws also extend; e.g. $(A \cup B \cup C)' = A' \cap B' \cap C'$.

Exercises Section 2.1 (1–12)

1. Ann and Bev have each applied for several jobs at a local university. Let A be the event that Ann is hired, and let B be the event that Bev is hired. Express in terms of A and B the following events:

   a. Ann is hired but not Bev.
   b. At least one of them is hired.
   c. Exactly one of them is hired.

2. Two voters, Al and Bill, are each choosing between one of three candidates—1, 2, and 3—who are running for city council. An experimental outcome specifies both Al's choice and Bill's choice, e.g., the pair (3, 2).

   a. List all elements of $\mathscr{S}$.
   b. List all outcomes in the event A that Al and Bill make the same choice.
   c. List all outcomes in the event B that neither of them votes for candidate 2.

3. Four universities—1, 2, 3, and 4—are participating in a holiday basketball tournament. In the first round, 1 will play 2 and 3 will play 4. Then the two winners will play for the championship, and the two losers will also play. One possible outcome can be denoted by 1324: 1 beats 2 and 3 beats 4 in first-round games, and then 1 beats 3 and 2 beats 4.

a. List all outcomes in $\mathscr{S}$.

b. Let *A* denote the event that 1 wins the tournament. List outcomes in *A*.

c. Let *B* denote the event that 2 gets into the championship game. List outcomes in *B*.

d. What are the outcomes in $A \cup B$ and in $A \cap B$? What are the outcomes in A'?

4. Suppose that vehicles taking a particular freeway exit can turn right (*R*), turn left (*L*), or go straight (*S*). Consider observing the direction for each of three successive vehicles.

   a. List all outcomes in the event *A* that all three vehicles go in the same direction.

   b. List all outcomes in the event *B* that all three vehicles take different directions.

   c. List all outcomes in the event *C* that exactly two of the three vehicles turn right.

   d. List all outcomes in the event *D* that exactly two vehicles go in the same direction.

   e. List the outcomes in D', $C \cup D$, and $C \cap D$.

5. Three components are connected to form a system as shown in the accompanying diagram. Because the components in the 2–3 subsystem are connected in parallel, that subsystem will function if at least one of the two individual components functions. For the entire system to function, component 1 must function and so must the 2–3 subsystem.
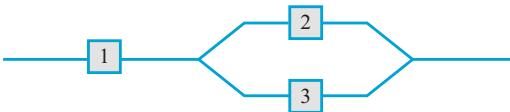
The experiment consists of determining the condition of each component: *S* (success) for a functioning component and *F* (failure) for a nonfunctioning component.

a. What outcomes are contained in the event *A* that exactly two out of the three components function?

b. What outcomes are contained in the event *B* that at least two of the components function?

c. What outcomes are contained in the event *C* that the system functions?

d. List outcomes in C', $A \cup C$, $A \cap C$, $B \cup C$, and $B \cap C$.

6. Each of a sample of four home mortgages is classified as fixed rate (*F*) or variable rate (*V*).

   a. What are the 16 outcomes in $\mathscr{S}$?

   b. Which outcomes are in the event that exactly three of the selected mortgages are fixed rate?

   c. Which outcomes are in the event that all four mortgages are of the same type?

   d. Which outcomes are in the event that at most one of the four is a variable rate mortgage?

   e. What is the union of the events in parts (c) and (d), and what is the intersection of these two events?

   f. What are the union and intersection of the two events in parts (b) and (c)?

7. A family consisting of three persons—*A*, *B*, and *C*—belongs to a medical clinic that always has a doctor at each of stations 1, 2, and 3. During a certain week, each member of the family visits the clinic once and is assigned at random to a station. The experiment consists of recording the station number for each member. One outcome is (1, 2, 1) for *A* to station 1, *B* to station 2, and *C* to station 1.

   a. List the 27 outcomes in the sample space.

   b. List all outcomes in the event that all three members go to the same station.

   c. List all outcomes in the event that all members go to different stations.

   d. List all outcomes in the event that no one goes to station 2.

8. A college library has five copies of a certain text on reserve. Copies 1 and 2 are first printings, whereas 3, 4, and 5 are second printings. A student examines these books in random order, stopping only when a second printing has been selected. One possible outcome is 5, and another is 213.

   a. List the outcomes in $\mathscr{S}$.
   b. Let A denote the event that exactly one book must be examined. What outcomes are in A?
   c. Let B be the event that book 5 is the one selected. What outcomes are in B?
   d. Let C be the event that book 1 is not examined. What outcomes are in C?

9. An academic department has just completed voting by secret ballot for a department head. The ballot box contains four slips with votes for candidate A and three slips with votes for candidate B. Suppose these slips are removed from the box one by one.

   a. List all possible outcomes.
   b. Suppose a running tally is kept as slips are removed. For what outcomes does A remain ahead of B throughout the tally?

10. A construction firm is currently working on three different buildings. Let A_i denote the event that the ith building is completed by the contract date. Use the operations of union, intersection, and complementation to describe each of the following events in terms of A_1, A_2, and A_3, draw a Venn diagram, and shade the region corresponding to each one.

    a. At least one building is completed by the contract date.
    b. All buildings are completed by the contract date.
    c. Only the first building is completed by the contract date.
    d. Exactly one building is completed by the contract date.
    e. Either the first building or both of the other two buildings are completed by the contract date.

11. Use Venn diagrams to verify De Morgan's laws:
    a. $(A \cup B)' = A' \cap B'$
    b. $(A \cap B)' = A' \cup B'$

12. a. In Example 2.10, identify three events that are mutually exclusive.
    b. Suppose there is no outcome common to all three of the events A, B, and C. Are these three events necessarily mutually exclusive? If your answer is yes, explain why; if your answer is no, give a counterexample using the experiment of Example 2.10.

2.2 Axioms, Interpretations, and Properties of Probability

Given an experiment and its sample space $\mathscr{S}$, the objective of probability is to assign to each event A a number $P(A)$, called **the probability of the event A**, which will give a precise measure of the chance that A will occur. To ensure that the probability assignments will be consistent with our intuitive notions of probability, all assignments must satisfy the following axioms (basic properties) of probability.

AXIOM 1   For any event A, $P(A) \geq 0$.

AXIOM 2 $P(\mathscr{S}) = 1$.

AXIOM 3   If $A_1, A_2, A_3, \ldots$ is an infinite collection of disjoint events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

Axiom 1 reflects the intuitive notion that the chance of A occurring should be nonnegative. The sample space is by definition the event that must occur when the experiment is performed ($\mathcal{S}$ contains all possible outcomes), so Axiom 2 says that the maximum possible probability of 1 is assigned to $\mathcal{S}$. The third axiom formalizes the idea that if we wish the probability that at least one of a number of events will occur, and no two of the events can occur simultaneously, then the chance of at least one occurring is the sum of the chances of the individual events.

You might wonder why the third axiom contains no reference to a *finite* collection of disjoint events. It is because the corresponding property for a finite collection can be derived from our three axioms. We want our axiom list to be as short as possible and not contain any property that can be derived from others on the list.

PROPOSITION $P(\varnothing) = 0$, where $\varnothing$ is the null event. This, in turn, implies that the property contained in Axiom 3 is valid for a *finite* collection of events.

Proof First consider the infinite collection $A_1 = \varnothing$, $A_2 = \varnothing$, $A_3 = \varnothing$, …. Since $\varnothing \cap \varnothing = \varnothing$, the events in this collection are disjoint and $\cup A_i = \varnothing$. Axiom 3 then gives

$$P(\varnothing) = \sum P(\varnothing)$$

This can happen only if $P(\varnothing) = 0$.

Now suppose that A_1, A_2, …, A_k are disjoint events, and append to these the infinite collection $A_{k+1} = \varnothing$, $A_{k+2} = \varnothing$, $A_{k+3} = \varnothing$, …. Then the events A_1, A_2, …, A_k, A_{k+1},… are disjoint, since $A \cap \varnothing = \varnothing$ for all events. Again invoking Axiom 3,

$$P\left(\bigcup_{i=1}^{k} A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{k} P(A_i) + \sum_{i=k+1}^{\infty} P(A_i) = \sum_{i=1}^{k} P(A_i) + \sum_{i=k+1}^{\infty} 0 = \sum_{i=1}^{k} P(A_i)$$

as desired.   ∎

Example 2.11 Consider evaluating a refurbished hard drive with a certifier. The certifier either deems the drive acceptable (the outcome A) or unacceptable (the outcome U). The sample space for this event is therefore $\mathcal{S} = \{A, U\}$. The axioms specify $P(\mathcal{S}) = 1$, so the probability assignment will be completed by determining $P(A)$ and $P(U)$. Since A and U are disjoint and their union is $\mathcal{S}$, the foregoing proposition implies that

$$1 = P(\mathcal{S}) = P(A) + P(U)$$

It follows that $P(U) = 1 - P(A)$. One possible assignment of probabilities is $P(A) = .5$, $P(U) = .5$, whereas another possible assignment is $P(A) = .75$, $P(U) = .25$. In fact, letting p represent any fixed number between 0 and 1, $P(A) = p$, $P(U) = 1 - p$ is an assignment consistent with the axioms.   ∎

Example 2.12 (Example 2.4 continued) Consider testing cell phone batteries coming off an assembly line one by one until a battery having a voltage within prescribed limits is found. The simple events are $E_1 = \{S\}$, $E_2 = \{FS\}$, $E_3 = \{FFS\}$, $E_4 = \{FFFS\}$, …. Suppose the probability of any particular battery being satisfactory is .99. Then it can be shown that the probability assignment $P(E_1) = .99$, $P(E_2) = (.01)(.99)$, $P(E_3) = (.01)^2(.99)$, … satisfies the axioms. In particular, because the E_i's are disjoint and $\mathcal{S} = E_1 \cup E_2 \cup E_3 \cup \ldots$, Axioms 2 and 3 require that

$$1 = P(\mathscr{S}) = P(E_1) + P(E_2) + P(E_3) + \cdots = .99[1 + .01 + (.01)^2 + (.01)^3 + \cdots]$$

This can be verified using the formula for the sum of a geometric series:

$$a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1 - r}$$

However, another legitimate (according to the axioms) probability assignment of the same "geometric" type is obtained by replacing .99 by any other number p between 0 and 1 (and .01 by $1 - p$). ∎

Interpreting Probability

Examples 2.11 and 2.12 show that the axioms do not completely determine an assignment of probabilities to events. The axioms serve only to rule out assignments inconsistent with our intuitive notions of probability. In the certifier experiment of Example 2.11, two particular assignments were suggested. The appropriate or correct assignment depends on the nature of the refurbished hard drives and also on one's interpretation of probability. The interpretation that is most frequently used and most easily understood is based on the notion of relative frequencies.

Consider an experiment that can be repeatedly performed in an identical and independent fashion, and let A be an event consisting of a fixed set of outcomes of the experiment. Simple examples of such repeatable experiments include the tack-tossing and die-rolling experiments previously discussed. If the experiment is performed n times, on some of the replications the event A will occur (the outcome will be in the set A), and on others, A will not occur. Let $n(A)$ denote the number of replications on which A does occur. Then the ratio $n(A)/n$ is called the *relative frequency* of occurrence of the event A in the sequence of n replications.

For example, let A be the event that a flight arrives on time at a certain airport. The results of ten such flights (the first ten replications) might be as follows.

Flight	1	2	3	4	5	6	7	8	9	10
On time (did A occur)?	Y	Y	N	Y	N	Y	Y	Y	Y	N
Relative frequency of A	1	1	.667	.75	.6	.667	.714	.75	.778	.7

Figure 2.2 shows the relative frequency, $n(A)/n$, of on-time arrivals as n increases. We see that the relative frequency fluctuates a lot for smaller values of n (this is also visible in the table); but, as n increases, the relative frequency appears to stabilize.
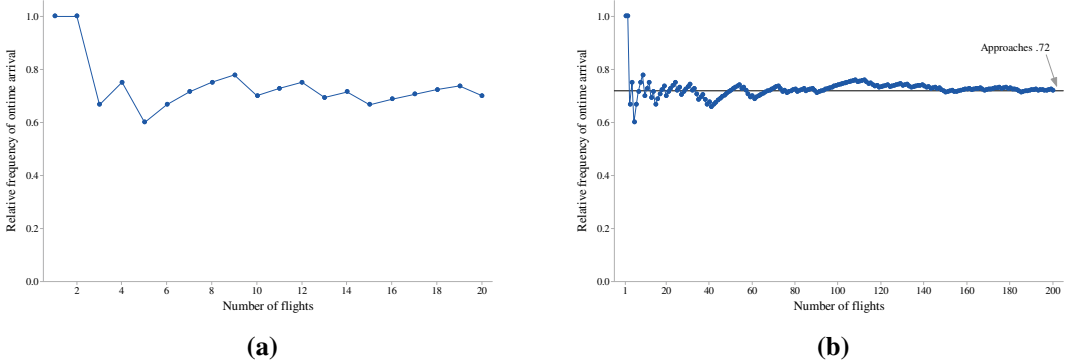

Figure 2.2   (a) Initial fluctuation and (b) eventual stabilization of relative frequency

More generally, both empirical evidence and mathematical theory indicate that any relative frequency of this sort will stabilize as the number of replications n increases. That is, as n gets arbitrarily large, $n(A)/n$ approaches a limiting value we refer to as the **long-run** (or **limiting**) **relative frequency** of the event A. The *objective interpretation of probability* identifies this limiting relative frequency with $P(A)$; e.g., in Figure 2.2b, the limiting relative frequency is .72, and so we say the probability of event A is $P(A) = .72$. A formal justification of this interpretation is provided by the *Law of Large Numbers*, a theorem we'll encounter in Chapter 6.

Suppose that probabilities are assigned to events in accordance with their limiting relative frequencies. Then a statement such as "the probability of a flight arriving on time is .72" means that of a large number of flights, roughly 72% will arrive on time. Similarly, if B is the event that a certain brand of dishwasher will need service while under warranty, then "$P(B) = .1$" is interpreted to mean that in the long run 10% of all such dishwashers will need warranty service. This does *not* mean that exactly 1 out of every 10 will need service, or exactly 20 out of 200 will need service, because 10 and 200 are not the long run. Such misinterpretations of probability as a guarantee on short-term outcomes are at the heart of the infamous *gambler's fallacy*.

This relative frequency interpretation of probability is said to be objective because it rests on a property of the experiment rather than on any particular individual concerned with the experiment. For example, two different observers of a sequence of coin tosses should both use the same probability assignments, since the observers have nothing to do with limiting relative frequency.

In practice, this interpretation is not as objective as it might seem, because the limiting relative frequency of an event will not be known. Thus we will have to assign probabilities based on our beliefs about the limiting relative frequency of events under study. Fortunately, there are many experiments for which there will be a consensus with respect to probability assignments. When we speak of a fair coin, we shall mean $P(H) = P(T) = .5$, and a fair die is one for which limiting relative frequencies of the six outcomes are all equal, suggesting probability assignments $P(1) = \cdots = P(6) = 1/6$.

Because the objective interpretation of probability is based on the notion of limiting frequency, its applicability is limited to experimental situations that are repeatable. Yet the language of probability is often used in connection with situations that are inherently unrepeatable. Examples include: "The chances are good for a peace agreement"; "It is likely that our company will be awarded the contract"; and "Because their best quarterback is injured, I expect them to score no more than 10 points against us." In such situations we would like, as before, to assign numerical probabilities to various outcomes and events (e.g., the probability is .9 that we will get the contract). We must therefore adopt an alternative interpretation of these probabilities. Because different observers may have different prior information and opinions concerning such experimental situations, probability assignments may now differ from individual to individual. Interpretations in such situations are thus referred to as *subjective*. The book by Winkler listed in the references gives a very readable survey of several subjective interpretations. Importantly, even subjective interpretations of probability must satisfy the three axioms (and all properties that follow from the axioms) in order to be valid.

More Probability Properties

COMPLEMENT RULE   For any event A, $P(A) = 1 - P(A')$.

Proof Since by definition of A', $A \cup A' = \mathcal{S}$ while A and A' are disjoint, $1 = P(\mathcal{S}) = P(A \cup A') = P(A) + P(A')$, from which the desired result follows. ∎

This proposition is surprisingly useful because there are many situations in which $P(A')$ is more easily obtained by direct methods than is $P(A)$.

Example 2.13 Consider a system of five identical components connected in series, as illustrated in Figure 2.3.
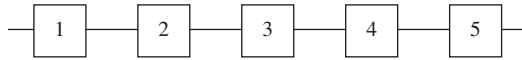
Figure 2.3 A system of five components connected in series

Denote a component that fails by F and one that doesn't fail by S (for success). Let A be the event that the system fails. For A to occur, at least one of the individual components must fail. Outcomes in A include *SSFSS* (1, 2, 4, and 5 all work, but 3 does not), *FFSSS*, and so on. There are, in fact, 31 different outcomes in A! However, A', the event that the system works, consists of the single outcome *SSSSS*. We will see in Section 2.5 that if 90% of all these components do not fail and different components fail independently of one another, then $P(A') = .9^5 = .59$. Thus $P(A) = 1 - .59 = .41$; so among a large number of such systems, roughly 41% will fail. ∎

In general, the Complement Rule is useful when the event of interest can be expressed as "at least …," because the complement "less than …" may be easier to work with. (In some problems, "more than …" is easier to deal with than "at most ….") When you are having difficulty calculating $P(A)$ directly, think of first determining $P(A')$.

PROPOSITION   For any event A, $P(A) \leq 1$.

This follows from the previous proposition: $1 = P(A) + P(A') \geq P(A)$, because $P(A') \geq 0$ by Axiom 1.

When A and B are disjoint, we know that $P(A \cup B) = P(A) + P(B)$. How can this union probability be obtained when the events are not disjoint?

ADDITION RULE   For any events A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Notice that the proposition is valid even if A and B are disjoint, since then $P(A \cap B) = 0$. The key idea is that, in adding $P(A)$ and $P(B)$, the probability of the intersection $A \cap B$ is actually counted twice, so $P(A \cap B)$ must be subtracted out.

Proof Note first that $A \cup B = A \cup (B \cap A')$, as shown in Figure 2.4 (p. 60). Because A and $(B \cap A')$ are disjoint, $P(A \cup B) = P(A) + P(B \cap A')$. But $B = (B \cap A) \cup (B \cap A')$(the union of the part of B in A and the part of B not in A). Furthermore, $(B \cap A)$ and $(B \cap A')$ are disjoint, so $P(B) = P(B \cap A) + P(B \cap A')$. Combining these results gives

$$P(A \cup B) = P(A) + P(B \cap A') = P(A) + [P(B) - P(A \cap B)] = P(A) + P(B) - P(A \cap B)$$
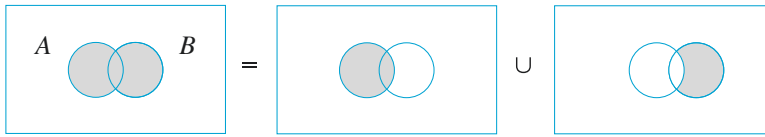
Figure 2.4 Representing $A \cup B$ as a union of disjoint events   ∎

Example 2.14   In a certain residential suburb, 60% of all households get internet service from the local cable company, 80% get television service from that company, and 50% get both services from the company. If a household is randomly selected, what is the probability that it gets at least one of these two services from the company, and what is the probability that it gets exactly one of the services from the company?

With A = {gets internet service from the cable company} and B = {gets television service from the cable company}, the given information implies that $P(A) = .6$, $P(B) = .8$, and $P(A \cap B) = .5$. The Addition Rule then applies to give

$$P(\text{gets at least one of these two services from the company})$$
$$= P(A \cup B) = P(A) + P(B) - P(A \cap B) = .6 + .8 - .5 = .9$$

The event that a household gets *only* television service from the company can be written as $A' \cap B$, i.e. (not internet) and television. Now Figure 2.4 implies that

$$.9 = P(A \cup B) = P(A) + P(A' \cap B) = .6 + P(A' \cap B)$$

from which $P(A' \cap B) = .3$. Similarly, $P(A \cap B') = P(A \cup B) - P(B) = .1$. This is all illustrated in Figure 2.5, from which we see that

$$P(\text{exactly one}) = P(A \cap B') + P(A' \cap B) = .1 + .3 = .4$$
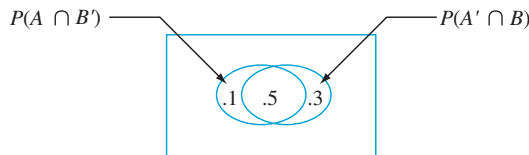
Figure 2.5 Probabilities for Example 2.14   ∎

The probability of a union of more than two events can be computed analogously. For three events A, B, and C, the result is

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

This can be seen by examining a Venn diagram of $A \cup B \cup C$, which is shown in Figure 2.6.
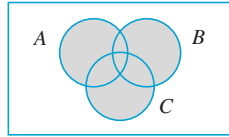
Figure 2.6 $A \cup B \cup C$

When $P(A)$, $P(B)$, and $P(C)$ are added, outcomes in certain intersections are double counted and the corresponding probabilities must be subtracted. But this results in $P(A \cap B \cap C)$ being subtracted once too often, so it must be added back. One formal proof involves applying the Addition Rule to $P((A \cup B) \cup C)$, the probability of the union of the two events $A \cup B$ and C; see Exercise 30. More generally, a result concerning $P(A_1 \cup \cdots \cup A_k)$ can be proved by induction or by other methods. The pattern of additions and subtractions (or, equivalently, the method of deriving such union probability formulas) is often called the **inclusion–exclusion principle**.

Determining Probabilities Systematically

When the number of possible outcomes (simple events) is large, there will be many compound events. A simple way to determine probabilities for these events that avoids violating the axioms and derived properties is to first determine probabilities $P(E_i)$ for all simple events. These should satisfy $P(E_i) \geq 0$ and $\sum_i P(E_i) = 1$. Then the probability of any compound event A is computed by adding together the $P(E_i)$'s for all E_i's in A:

$$P(A) = \sum_{E_i \text{ in } A} P(E_i)$$

Example 2.15 During off-peak hours a commuter train has five cars. Suppose a commuter is twice as likely to select the middle car (#3) as to select either adjacent car (#2 or #4), and is twice as likely to select either adjacent car as to select either end car (#1 or #5). Let $p_i = P(\text{car } i \text{ is selected}) = P(E_i)$. Then we have $p_3 = 2p_2 = 2p_4$ and $p_2 = 2p_1 = 2p_5 = p_4$. This gives

$$1 = \sum P(E_i) = p_1 + 2p_1 + 4p_1 + 2p_1 + p_1 = 10p_1$$

implying $p_1 = p_5 = .1$, $p_2 = p_4 = .2$, and $p_3 = .4$. The probability that one of the three middle cars is selected (a compound event) is then $p_2 + p_3 + p_4 = .8$. ∎

Equally Likely Outcomes

In many experiments consisting of N outcomes, it is reasonable to assign equal probabilities to all N simple events. These include such obvious examples as tossing a fair coin or fair die once (or any fixed number of times), or selecting one or several cards from a well-shuffled deck of 52. With $p = P(E_i)$ for every i,

$$1 = \sum_{i=1}^{N} P(E_i) = \sum_{i=1}^{N} p = p \cdot N \quad \text{so } p = \frac{1}{N}$$

That is, if there are N possible outcomes, then the probability assigned to each is $1/N$.

Now consider an event A, with $N(A)$ denoting the number of outcomes contained in A. Then

$$P(A) = \sum_{E_i \,\text{in}\, A} P(E_i) = \sum_{E_i \,\text{in}\, A} \frac{1}{N} = \frac{N(A)}{N}$$

Once we have counted the number N of outcomes in the sample space, to compute the probability of any event we must count the number of outcomes contained in that event and take the ratio of the two numbers. Thus when outcomes are equally likely, computing probabilities reduces to counting.

Example 2.16 When two dice are rolled separately, there are $N = 36$ outcomes (delete the first row and column from the table in Example 2.3). If both the dice are fair, all 36 outcomes are equally likely, so $P(E_i) = 1/36$ for each simple event. The event $A = \{$sum of two numbers is 8$\}$ consists of the five outcomes (2, 6), (3, 5), (4, 4), (5, 3), and (6, 2), so

$$P(A) = \frac{N(A)}{N} = \frac{5}{36} \qquad\qquad \blacksquare$$

The next section of this book develops some useful counting methods.

Exercises: Section 2.2 (13–30)

13. A mutual fund company offers its customers several different funds: a money market fund, three different bond funds (short, intermediate, and long term), two stock funds (moderate and high risk), and a balanced fund. Among customers who own shares in just one fund, the percentages of customers in the different funds are as follows:

Money market	20%	High-risk stock	18%
Short bond	15%	Moderate-risk stock	25%
Intermediate bond	10%	Balanced	7%
Long bond	5%		

    A customer who owns shares in just one fund is randomly selected.

    a. What is the probability that the selected individual owns shares in the balanced fund?
    b. What is the probability that the individual owns shares in a bond fund?

    c. What is the probability that the selected individual does not own shares in a stock fund?

14. Consider randomly selecting a student at a certain university, and let A denote the event that the selected individual has a Visa credit card and B be the analogous event for a MasterCard. Suppose that $P(A) = .5, P(B) = .4$, and $P(A \cap B) = .25$.

    a. Compute the probability that the selected individual has at least one of the two types of cards (i.e., the probability of the event $A \cup B$).
    b. What is the probability that the selected individual has neither type of card?
    c. Describe, in terms of A and B, the event that the selected student has a Visa card but not a MasterCard, and then calculate the probability of this event.

15. A consulting firm presently has bids out on three projects. Let $A_i = \{$awarded project $i\}$, for $i = 1, 2, 3$, and suppose that $P(A_1) = .22$, $P(A_2) = .25, P(A_3) = .28, P(A_1 \cap A_2) = .11$, $P(A_1 \cap A_3) = .05$, $P(A_2 \cap A_3) = .07$, and

$P(A_1 \cap A_2 \cap A_3) = .01$. Express in words each of the following events, and compute the probability of each event:

a. $A_1 \cup A_2$
b. $A_1' \cap A_2'$ [Hint: $(A_1 \cup A_2)' = A_1' \cap A_2'$]
c. $A_1 \cup A_2 \cup A_3$
d. $A_1' \cap A_2' \cap A_3'$
e. $A_1' \cap A_2' \cap A_3$
f. $(A_1' \cap A_2') \cup A_3$

16. A particular state will elect both a governor and a senator. Let A be the event that a randomly selected voter has a favorable view of a certain party's senatorial candidate, and let B be the corresponding event for that party's gubernatorial candidate. Suppose that $P(A') = .44$, $P(B') = .57$, and $P(A \cup B) = .68$.

a. What is the probability that a randomly selected voter has a favorable view of both candidates?
b. What is the probability that a randomly selected voter has an unfavorable view of at least one of these candidates?
c. What is the probability that a randomly selected voter has a favorable view of exactly one of these candidates?

17. Consider the type of clothes dryer (gas or electric) purchased by each of five different customers at a certain store.

a. If the probability that at most one of these customers purchases an electric dryer is .428, what is the probability that at least two purchase an electric dryer?
b. If $P(\text{all five purchase gas}) = .116$ and $P(\text{all five purchase electric}) = .005$, what is the probability that at least one of each type is purchased?

18. An individual is presented with three different glasses of cola, labeled C, D, and P. He is asked to taste all three and then list them in order of preference. Suppose the same cola has actually been put into all three glasses.

a. What are the simple events in this ranking experiment, and what probability would you assign to each one?
b. What is the probability that C is ranked first?
c. What is the probability that C is ranked first and D is ranked last?

19. Let A denote the event that the next request for assistance from a statistical software consultant relates to the SPSS package, and let B be the event that the next request is for help with SAS. Suppose that $P(A) = .30$ and $P(B) = .50$.

a. Why is it not the case that $P(A) + P(B) = 1$?
b. Calculate $P(A')$.
c. Calculate $P(A \cup B)$.
d. Calculate $P(A' \cap B')$.

20. A box contains four 40-W bulbs, five 60-W bulbs, and six 75-W bulbs. If bulbs are selected one by one in random order, what is the probability that at least two bulbs must be selected to obtain one that is rated 75 W?

21. Human visual inspection of solder joints on printed circuit boards can be very subjective. Part of the problem stems from the numerous types of solder defects (e.g., pad nonwetting, knee visibility, voids) and even the degree to which a joint possesses one or more of these defects. Consequently, even highly trained inspectors can disagree on the disposition of a particular joint. In one batch of 10,000 joints, inspector A found 724 that were judged defective, inspector B found 751 such joints, and 1159 of the joints were judged defective by at least one of the inspectors. Suppose that one of the 10,000 joints is randomly selected.

a. What is the probability that the selected joint was judged to be defective by neither of the two inspectors?
b. What is the probability that the selected joint was judged to be defective by inspector B but not by inspector A?

22. A factory operates three different shifts. Over the last year, 200 accidents have occurred at the factory. Some of these can be attributed at least in part to unsafe working conditions, whereas the others are unrelated to working conditions. The accompanying table gives the percentage of accidents falling in each type of accident–shift category.

Shift	Unsafe conditions	Unrelated to conditions
Day	10%	35%
Swing	8%	20%
Night	5%	22%

Suppose one of the 200 accident reports is randomly selected from a file of reports, and the shift and type of accident are determined.

a. What are the simple events?
b. What is the probability that the selected accident was attributed to unsafe conditions?
c. What is the probability that the selected accident did not occur on the day shift?

23. An insurance company offers four different deductible levels—none, low, medium, and high—for its homeowner's policyholders and three different levels—low, medium, and high—for its automobile policyholders. The accompanying table gives proportions for the various categories of policyholders who have both types of insurance. For example, the proportion of individuals with both low homeowner's deductible and low auto deductible is .06 (6% of all such individuals).

	Homeowner's			
Auto	N	L	M	H
L	.04	.06	.05	.03
M	.07	.10	.20	.10
H	.02	.03	.15	.15

Suppose an individual having both types of policies is randomly selected.

a. What is the probability that the individual has a medium auto deductible and a high homeowner's deductible?
b. What is the probability that the individual has a low auto deductible? A low homeowner's deductible?
c. What is the probability that the individual is in the same category for both auto and homeowner's deductibles?
d. Based on your answer in part (c), what is the probability that the two categories are different?
e. What is the probability that the individual has at least one low deductible level?
f. Using the answer in part (e), what is the probability that neither deductible level is low?

24. The route used by a driver in commuting to work contains two intersections with traffic signals. The probability that he must stop at the first signal is .4, the analogous probability for the second signal is .5, and the probability that he must stop at one or more of the two signals is .6. What is the probability that he must stop

a. At both signals?
b. At the first signal but not at the second one?
c. At exactly one signal?

25. The computers of six faculty members in a certain department are to be replaced. Two of the faculty members have selected laptop machines, and the other four have chosen desktop machines. Suppose that only two of the setups can be done on a particular day, and the two computers to be set up are randomly selected from the six (implying 15 equally likely outcomes; if the computers are numbered 1, 2, …, 6,

then one outcome consists of computers 1 and 2, another consists of computers 1 and 3, and so on).

a. What is the probability that both selected setups are for laptop computers?
b. What is the probability that both selected setups are desktop machines?
c. What is the probability that at least one selected setup is for a desktop computer?
d. What is the probability that at least one computer of each type is chosen for setup?

26. Use the axioms to show that if one event A is contained in another event B (i.e., A is a subset of B), then $P(A) \leq P(B)$. [*Hint*: For such A and B, A and $B \cap A'$ are disjoint and $B = A \cup (B \cap A')$, as can be seen from a Venn diagram.] For general A and B, what does this imply about the relationship among $P(A \cap B)$, $P(A)$, and $P(A \cup B)$?

27. The three major options on a car model are an automatic transmission (A), a sunroof (B), and an upgraded stereo (C). If 70% of all purchasers request A, 80% request B, 75% request C, 85% request A or B, 90% request A or C, 95% request B or C, and 98% request A or B or C, compute the probabilities of the following events. [*Hint*: "A or B" is the event that at least one of the two options is requested; try drawing a Venn diagram and labeling all regions.]

a. The next purchaser will request at least one of the three options.
b. The next purchaser will select none of the three options.

c. The next purchaser will request only an automatic transmission and neither of the other two options.
d. The next purchaser will select exactly one of these three options.

28. A certain system can experience three different types of defects. Let A_i ($i = 1, 2, 3$) denote the event that the system has a defect of type i. Suppose that

$$P(A_1) = .12 \quad P(A_2) = .07 \quad P(A_3) = .05$$
$$P(A_1 \cup A_2) = .13 \quad P(A_1 \cup A_3) = .14$$
$$P(A_2 \cup A_3) = .10 \quad P(A_1 \cap A_2 \cap A_3) = .01$$

a. What is the probability that the system does not have a type 1 defect?
b. What is the probability that the system has both type 1 and type 2 defects?
c. What is the probability that the system has both type 1 and type 2 defects but not a type 3 defect?
d. What is the probability that the system has at most two of these defects?

29. In Exercise 7, suppose that any incoming individual is equally likely to be assigned to any of the three stations irrespective of where other individuals have been assigned. What is the probability that

a. All three family members are assigned to the same station?
b. At most two family members are assigned to the same station?
c. Every family member is assigned to a different station?

30. Apply the Addition Rule to the union of the two events ($A \cup B$) and C in order to verify the formula for $P(A \cup B \cup C)$.

2.3   Counting Methods

When the various outcomes of an experiment are equally likely (the same probability is assigned to each simple event), the task of computing probabilities reduces to counting. In particular, if N is the number of outcomes in a sample space and $N(A)$ is the number of outcomes contained in an event A, then

$$P(A) = \frac{N(A)}{N} \tag{2.1}$$

If a list of the outcomes is available or easy to construct and N is small, then the numerator and denominator of Equation (2.1) can be obtained without the benefit of any general counting principles.

There are, however, many experiments for which the effort involved in constructing such a list is prohibitive because N is quite large. By exploiting some general counting rules, it is possible to compute probabilities of the form (2.1) without a listing of outcomes. These rules are also useful in many problems involving outcomes that are not equally likely. Several of the rules developed here will be used in studying probability distributions in the next chapter.

The Fundamental Counting Principle

Our first counting rule applies to any situation in which an event consists of ordered pairs of objects and we wish to count the number of such pairs. By an ordered pair, we mean that, if O_1 and O_2 are objects, then the pair (O_1, O_2) is different from the pair (O_2, O_1). For example, if an individual selects one airline for a trip from Los Angeles to Chicago and a second one for continuing on to New York, one possibility is (American, United), another is (United, American), and still another is (United, United).

PROPOSITION	If the first element or object of an ordered pair can be selected in n_1 ways, and for each of these n_1 ways the second element of the pair can be selected in n_2 ways, then the number of pairs is $n_1 n_2$.

Example 2.17   A homeowner doing some remodeling requires the services of both a plumbing contractor and an electrical contractor. If there are 12 plumbing contractors and 9 electrical contractors available in the area, in how many ways can the contractors be chosen? If we denote the plumbers by $P_1, \ldots, P_{12}$ and the electricians by $Q_1, \ldots, Q_9$, then we wish the number of pairs of the form (P_i, Q_j). With $n_1 = 12$ and $n_2 = 9$, the proposition yields $N = (12)(9) = 108$ possible ways of choosing the two types of contractors.   ∎

In Example 2.17, the choice of the second element of the pair did not depend on which first element was chosen or occurred. As long as there is the same number of choices of the second element for each first element, the foregoing proposition is valid even when the set of possible second elements depends on the first element.

Example 2.18   A family has just moved to a new city and requires the services of both an obstetrician and a pediatrician. There are two easily accessible medical clinics, each having two obstetricians and three pediatricians. The family will obtain maximum health insurance benefits by joining a clinic and selecting both doctors from that clinic. In how many ways can this be done? Denote the obstetricians by O_1, O_2, O_3, and O_4 and the pediatricians by $P_1, \ldots, P_6$. Then we wish the number of pairs (O_i, P_j) for which O_i and P_j are associated with the same clinic. Because there are four obstetricians, $n_1 = 4$, and for each there are three choices of pediatrician, so $n_2 = 3$. Applying the proposition rule gives $N = n_1 n_2 = 12$ possible choices.   ∎

If a six-sided die is tossed five times in succession, then each possible outcome is an ordered collection of five numbers such as (1, 3, 1, 2, 4) or (6, 5, 2, 2, 2). We will call an ordered collection of k objects a **k-tuple** (so a pair is a 2-tuple and a triple is a 3-tuple). Each outcome of the die-tossing experiment is then a 5-tuple. The following theorem, called the Fundamental Counting Principle, generalizes the previous proposition to k-tuples.

FUNDAMENTAL COUNTING PRINCIPLE	Suppose a set consists of ordered collections of k elements (k-tuples) and that there are n_1 possible choices for the first element; for each choice of the first element, there are n_2 possible choices of the second element; …; for each possible choice of the first $k - 1$ elements, there are n_k choices of the kth element. Then there are $n_1 n_2 \cdot \ldots \cdot n_k$ possible k-tuples.

Example 2.19   (Example 2.17 continued) Suppose the home remodeling job involves first purchasing several kitchen appliances. They will all be purchased from the same dealer, and there are five dealers in the area. With the dealers denoted by $D_1, \ldots, D_5$, there are $N = n_1 n_2 n_3 = (5)(12)(9) = 540$ 3-tuples of the form (D_i, P_j, Q_k), so there are 540 ways to choose first an appliance dealer, then a plumbing contractor, and finally an electrical contractor.   ∎

Example 2.20   (Example 2.18 continued) If each clinic has both three specialists in internal medicine and two general surgeons, there are $n_1 n_2 n_3 n_4 = (4)(3)(3)(2) = 72$ ways to select one doctor of each type such that all doctors practice at the same clinic.   ∎

Tree Diagrams

In many counting and probability problems, a **tree diagram** can be used to represent pictorially all the possibilities. The tree diagram associated with Example 2.18 appears in Figure 2.7. Starting from a point on the left side of the diagram, for each possible first element of a pair a straight-line segment emanates rightward. Each of these lines is referred to as a *first-generation branch*. Now for any given first-generation branch we construct another line segment emanating from the tip of the branch for each possible choice of a second element of the pair. Each such line segment is a *second-generation branch*. Because there are four obstetricians, there are four first-generation branches, and three pediatricians for each obstetrician yield three second-generation branches emanating from each first-generation branch.
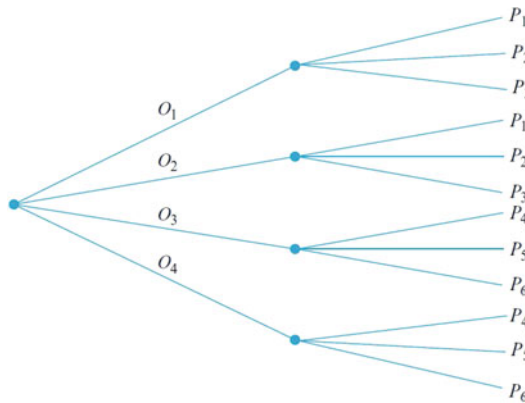
Figure 2.7   Tree diagram for Example 2.18

Permutations

So far the successive elements of a k-tuple were selected from entirely different sets (e.g., appliance dealers, then plumbers, and finally electricians). In several tosses of a die, the set from which successive elements are chosen is always $\{1, 2, 3, 4, 5, 6\}$, but the choices are made "with replacement" so that the same element can appear more than once. If the die is rolled once, there are obviously 6 possible outcomes; for two rolls, there are $6^2 = 36$ possibilities, since we distinguish $(3, 5)$ from $(5, 3)$. In general, if k selections are made with replacement from a set of n distinct objects (such as the six sides of a die), then the total number of possible outcomes is n^k.

We now consider a fixed set consisting of n distinct elements and suppose that a k-tuple is formed by selecting successively from this set *without replacement* so that an element can appear in at most one of the k positions.

DEFINITION   Any ordered sequence of k objects taken without replacement from a set of n distinct objects is called a **permutation** of size k of the objects. The number of permutations of size k that can be constructed from the n objects is denoted by $_nP_k$.

The number of permutations of size k is obtained immediately from the Fundamental Counting Principle. The first element can be chosen in n ways; for each of these n ways the second element can be chosen in $n - 1$ ways; and so on. Finally, for each way of choosing the first $k - 1$ elements, the kth element can be chosen in $n - (k - 1) = n - k + 1$ ways, so

$$_nP_k = n(n - 1)(n - 2) \cdot \cdots \cdot (n - k + 2)(n - k + 1)$$

Example 2.21 Ten teaching assistants are available for grading papers in a particular course. The first exam consists of four questions, and the professor wishes to select a different assistant to grade each question (only one assistant per question). In how many ways can assistants be chosen to grade the exam? Here n = the number of assistants = 10 and k = the number of questions = 4. The number of different grading assignments is then $_{10}P_4 = (10)(9)(8)(7) = 5040$. ∎

Example 2.22 *The Birthday Problem.* Disregarding the possibility of a February 29 birthday, suppose a randomly selected individual is equally likely to have been born on any one of the other 365 days. If ten people are randomly selected, what is the probability that all have different birthdays?

Imagine we draw ten days, *with replacement*, from the calendar to represent the birthdays of the ten randomly selected people. One possible outcome of this selection would be (March 31, December 30, …, September 27, February 12). There are 365^{10} such outcomes. The number of outcomes among them with no repeated birthdays is

$$(365)(364) \cdot \cdots \cdot (356) = {_{365}P_{10}}$$

(any of the 365 calendar days may be selected first; if March 31 is chosen, any of the other 364 days is acceptable for the second selection; and so on). Hence, the probability all ten randomly selected people have different birthdays equals $_{365}P_{10}/365^{10} = .883$. Equivalently, there's only a .117 chance that at least two people out of these ten will share a birthday. It's worth noting that the first probability can be rewritten as

$$\frac{_{365}P_{10}}{365^{10}} = \frac{365}{365} \cdot \frac{364}{365} \cdot \cdots \cdot \frac{356}{365}$$

We may think of each fraction as representing the chance the next birthday selected will be different from all previous ones. (This is an example of *conditional probability*, the topic of the next section.)

Now replace 10 with k (i.e., k randomly selected birthdays); what is the smallest k for which there is at least a 50–50 chance that two or more people will have the same birthday? Most people incorrectly guess that we need a very large group of people for this to be true; the most common guess is that 183 people are required (half the days on the calendar). But the required value of k is actually much smaller: the probability that k randomly selected people all have different birthdays equals $_{365}P_k/365^k$, which not surprisingly decreases as k increases. Figure 2.8 displays this probability for increasing values of k. As it turns out, the smallest k for which this probability falls below .5 is just $k = 23$. That is, there is less than a 50–50 chance (.4927, to be precise) of 23 randomly selected people all having different birthdays, and thus a probability .5073 that at least two people in a random sample of 23 will share a birthday.
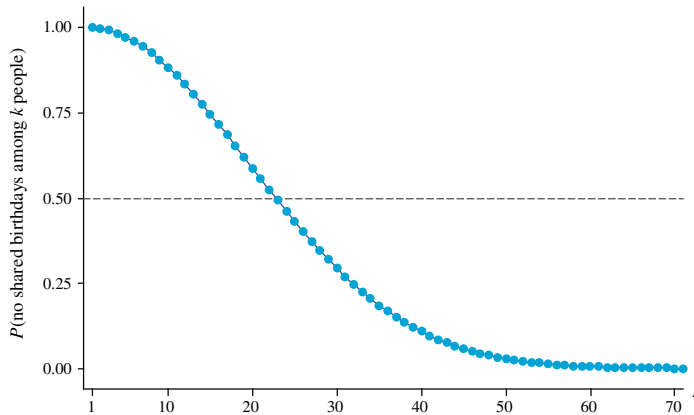
Figure 2.8 P(no birthday match) in Example 2.22   ∎

The expression for $_nP_k$ can be rewritten with the aid of factorial notation. Recall that 7! (read "7 factorial") is compact notation for the descending product of integers $(7)(6)(5)(4)(3)(2)(1)$. More generally, for any positive integer m, $m! = m(m-1)(m-2) \cdot \cdots \cdot (2)(1)$. This gives $1! = 1$, and we also define $0! = 1$.

Using factorial notation, $(10)(9)(8)(7) = (10)(9)(8)(7)(6!)/6! = 10!/6!$. More generally,

$$_nP_k = n(n-1) \cdot \cdots \cdot (n-k+1)$$
$$= \frac{n(n-1) \cdot \cdots \cdot (n-k+1)(n-k)(n-k-1) \cdot \cdots \cdot (2)(1)}{(n-k)(n-k-1) \cdot \cdots \cdot (2)(1)}$$

which becomes

$$_nP_k = \frac{n!}{(n-k)!}$$

For example, $_9P_3 = 9!/(9-3)! = 9!/6! = 9 \cdot 8 \cdot 7 \cdot 6!/6! = 9 \cdot 8 \cdot 7$. Note also that because $0! = 1$, $_nP_n = n!/(n-n)! = n!/0! = n!/1 = n!$, as it should.

Combinations

Often the objective is to count the number of *unordered* subsets of size k that can be formed from a set consisting of n distinct objects. For example, in bridge it is only the 13 cards in a hand and not the order in which they are dealt that is important; in the formation of a committee, the order in which committee members are listed is frequently unimportant.

DEFINITION   Given a set of n distinct objects, any unordered subset of size k of the objects is called a **combination**. The number of combinations of size k that can be formed from n distinct objects will be denoted by $\binom{n}{k}$ or $_nC_k$.

The number of combinations of size k from a particular set is smaller than the number of permutations because, when order is disregarded, some of the permutations correspond to the same combination. Consider, for example, the set $\{A, B, C, D, E\}$ consisting of five elements. There are $_5P_3 = 5!/(5-3)! = 60$ permutations of size 3. There are six permutations of size 3 consisting of the elements A, B, and C because these three can be ordered $3 \cdot 2 \cdot 1 = 3! = 6$ ways: (A, B, C), (A, C, B), (B, A, C), (B, C, A), (C, A, B), and (C, B, A). These six *permutations* are equivalent to the single *combination* $\{A, B, C\}$. Similarly, for any other combination of size 3, there are 3! permutations, each obtained by ordering the three objects. Thus,

$$60 = {_5P_3} = \binom{5}{3} \cdot 3! \quad \text{so} \quad \binom{5}{3} = \frac{60}{3!} = 10$$

These ten combinations are

$$\{A, B, C\} \quad \{A, B, D\} \quad \{A, B, E\} \quad \{A, C, D\} \quad \{A, C, E\}$$
$$\{A, D, E\} \quad \{B, C, D\} \quad \{B, C, E\} \quad \{B, D, E\} \quad \{C, D, E\}$$

When there are n distinct objects, any permutation of size k is obtained by ordering the k unordered objects of a combination in one of $k!$ ways, so the number of permutations is the product of $k!$ and the number of combinations. This gives

$$_nC_k \quad \text{or} \quad \binom{n}{k} = \frac{_nP_k}{k!} = \frac{n!}{k!(n-k)!}$$

Notice that $\binom{n}{n} = 1$ and $\binom{n}{0} = 1$ because there is only one way to choose a set of (all) n elements or of no elements, and $\binom{n}{1} = n$ since there are n subsets of size 1.

Example 2.23 A bridge hand consists of any 13 cards selected from a 52-card deck without regard to order. There are $\binom{52}{13} = 52!/(13! \cdot 39!)$ different bridge hands, which works out to approximately 635 billion. Since there are 13 cards in each suit, the number of hands consisting entirely of clubs and/or spades (no red cards) is $\binom{26}{13} = 26!/(13! \cdot 13!) = 10,400,600$. One of these $\binom{26}{13}$ hands consists entirely of spades, and one consists entirely of clubs, so there are $\left[\binom{26}{13} - 2\right]$ hands that

consist entirely of clubs and spades with both suits represented in the hand. Suppose a bridge hand is dealt from a well-shuffled deck (i.e., 13 cards are randomly selected from among the 52 possibilities) and let

A = {the hand consists entirely of spades and clubs with both suits represented}
B = {the hand consists of exactly two suits}

   The $N = \binom{52}{13}$ possible outcomes are equally likely, so

$$P(A) = \frac{N(A)}{N} = \frac{\binom{26}{13} - 2}{\binom{52}{13}} = .0000164$$

Since there are $\binom{4}{2} = 6$ combinations consisting of two suits, of which spades and clubs is one such combination,

$$P(B) = \frac{N(B)}{N} = \frac{6\left[\binom{26}{13} - 2\right]}{\binom{52}{13}} = .0000983$$

That is, a hand consisting entirely of cards from exactly two of the four suits will occur roughly once in every 10,000 hands. If you play bridge only once a month, it is likely that you will never be dealt such a hand.   ∎

Example 2.24   A university has received a shipment of 25 new laptops for staff and faculty, of which 10 have AMD processors and 15 have Intel chips. If 6 of these 25 laptops are selected at random to be checked by a technician, what is the probability that exactly 3 of those selected have Intel processors (so that the other 3 are AMD)?

   Let D_3 = {exactly 3 of the 6 selected have Intel processors}. Assuming that any particular set of 6 laptops is as likely to be chosen as is any other set of 6, we have equally likely outcomes, so $P(D_3) = N(D_3)/N$, where N is the number of ways of choosing 6 laptops from the 25 and $N(D_3)$ is the number of ways of choosing 3 with AMD processors and 3 with Intel chips. Thus $N = \binom{25}{6}$. To obtain $N(D_3)$, think of first choosing 3 of the 15 Intel laptops and then 3 of the AMD laptops. There are $\binom{15}{3}$ ways of choosing the 3 with Intel processors, and there are $\binom{10}{3}$ ways of choosing the 3 with AMD processors; by the Fundamental Counting Principle, $N(D_3)$ is the product of these two numbers. So,

$$P(D_3) = \frac{N(D_3)}{N} = \frac{\binom{15}{3}\binom{10}{3}}{\binom{25}{6}} = \frac{\frac{15!}{3!12!} \cdot \frac{10!}{3!7!}}{\frac{25!}{6!19!}} = .3083$$

Next, let D_4 = {exactly 4 of the 6 laptops selected have Intel processors} and define D_5 and D_6 in an analogous manner. Notice that the events D_3, D_4, D_5, and D_6 are disjoint. Thus, the probability that *at least* 3 laptops with Intel processors are selected is

$$P(D_3 \cup D_4 \cup D_5 \cup D_6) = P(D_3) + P(D_4) + P(D_5) + P(D_6)$$

$$= \frac{\binom{15}{3}\binom{10}{3}}{\binom{25}{6}} + \frac{\binom{15}{4}\binom{10}{2}}{\binom{25}{6}} + \frac{\binom{15}{5}\binom{10}{1}}{\binom{25}{6}} + \frac{\binom{15}{6}\binom{10}{0}}{\binom{25}{6}} = .8530 \qquad \blacksquare$$

Exercises: Section 2.3 (31–48)

31. The College of Science Student Council has one representative from each of the five science departments (biology, chemistry, statistics, mathematics, physics). In how many ways can
    a. Both a council president and a vice president be selected?
    b. A president, a vice president, and a secretary be selected?
    c. Two members be selected for the Dean's Council?

32. A friend is giving a dinner party. Her current wine supply includes 8 bottles of zinfandel, 10 of merlot, and 12 of cabernet (she drinks only red wine), all from different wineries.
    a. If she wants to serve 3 bottles of zinfandel and serving order is important, how many ways are there to do this?
    b. If 6 bottles of wine are to be randomly selected from the 30 for serving, how many ways are there to do this?
    c. If 6 bottles are randomly selected, how many ways are there to obtain two bottles of each variety?
    d. If 6 bottles are randomly selected, what is the probability that this results in two bottles of each variety being chosen?
    e. If 6 bottles are randomly selected, what is the probability that all of them are the same variety?

33. a. Beethoven wrote 9 symphonies and Mozart wrote 27 piano concertos. If a university radio station announcer wishes to play first a Beethoven symphony and then a Mozart concerto, in how many ways can this be done?
    b. The station manager decides that on each successive night (7 days per week), a Beethoven symphony will be played, followed by a Mozart piano concerto, followed by a Schubert string quartet (of which there are 15). For roughly how many years could this policy be continued before exactly the same program would have to be repeated?

34. A chain of home electronics stores is offering a special price on a complete set of components (receiver, CD/MP3 player, speakers). A purchaser is offered a choice of manufacturer for each component:

Receiver	Kenwood, Onkyo, Pioneer, Sony, Yamaha
CD/MP3 player	Onkyo, Pioneer, Sony, Panasonic
Speakers	Boston, Infinity, Polk

A switchboard display in the store allows a customer to connect any selection of components (consisting of one of each type). Use the product rules to answer the following questions:
    a. In how many ways can one component of each type be selected?
    b. In how many ways can components be selected if both the receiver and the CD/MP3 player are to be Sony?
    c. In how many ways can components be selected if none is to be Sony?

d. In how many ways can a selection be made if at least one Sony component is to be included?

e. If someone flips switches on the selection in a completely random fashion, what is the probability that the system selected contains at least one Sony component? Exactly one Sony component?

35. A particular iPod playlist contains 100 songs, of which 10 are by the Beatles. Suppose the shuffle feature is used to play the songs in random order (the randomness of the shuffling process is investigated in "Does Your iPod *Really* Play Favorites?" (*The Amer. Statistician* 2009: 263–268)). What is the probability that the first Beatles song heard is the fifth song played?

36. A local bar stocks 12 American beers, 8 Mexican beers, and 9 German beers. You ask the bartender to pick out a five-beer "sampler" for you. Assume the bartender makes the five selections at random and without replacement.

a. What is the probability you get at least four American beers?

b. What is the probability you get five beers from the same country?

37. The statistics department at the authors' university participates in an annual volleyball tournament. Suppose that all 16 department members are willing to play.

a. How many different six-person volleyball rosters could be generated? (That is, how many years could the department participate in the tournament without repeating the same six-person team?)

b. The statistics department faculty consists of 5 women and 11 men. How many rosters comprised of exactly 2 women and 4 men be generated?

c. The tournament's rules actually require that each team includes *at least* two women. Under this rule, how many valid teams could be generated?

d. Suppose this year the department decides to randomly select its six players. What is the probability the randomly selected team has exactly two women? At least two women?

38. A production facility employs 20 workers on the day shift, 15 workers on the swing shift, and 10 workers on the graveyard shift. A quality control consultant is to select 6 of these workers for in-depth interviews. Suppose the selection is made in such a way that any particular group of 6 workers has the same chance of being selected as does any other group (drawing 6 slips without replacement from among 45).

a. How many selections result in all 6 workers coming from the day shift? What is the probability that all 6 selected workers will be from the day shift?

b. What is the probability that all 6 selected workers will be from the same shift?

c. What is the probability that at least two different shifts will be represented among the selected workers?

d. What is the probability that at least one of the shifts will be unrepresented in the sample of workers?

39. An academic department with five faculty members narrowed its choice for department head to either candidate *A* or candidate *B*. Each member then voted on a slip of paper for one of the candidates. Suppose there are actually three votes for *A* and two for *B*. If the slips are selected for tallying in random order, what is the probability that *A* remains ahead of *B* throughout the vote count (for example, this event occurs if the selected ordering is *AABAB*, but not for *ABBAA*)?

40. An experimenter is studying the effects of temperature, pressure, and type of catalyst on yield from a chemical reaction. Three different temperatures, four different pressures, and five different catalysts are under consideration.

a. If any particular experimental run involves the use of a single temperature, pressure, and catalyst, how many experimental runs are possible?

b. How many experimental runs involve use of the lowest temperature and two lowest pressures?

41. Refer to the previous exercise and suppose that five different experimental runs are to be made on the first day of experimentation. If the five are randomly selected from among all the possibilities, so that any group of five has the same probability of selection, what is the probability that a different catalyst is used on each run?

42. A box in a certain supply room contains four 40-W lightbulbs, five 60-W bulbs, and six 75-W bulbs. Suppose that three bulbs are randomly selected.

a. What is the probability that exactly two of the selected bulbs are rated 75 W?

b. What is the probability that all three of the selected bulbs have the same rating?

c. What is the probability that one bulb of each type is selected?

d. Suppose now that bulbs are to be selected one by one until a 75-W bulb is found. What is the probability that it is necessary to examine at least six bulbs?

43. Fifteen telephones have just been received at an authorized service center. Five of these telephones are cellular, five are cordless, and the other five are corded phones. Suppose that these components are randomly allocated the numbers 1, 2, ..., 15 to establish the order in which they will be serviced.

a. What is the probability that all the cordless phones are among the first ten to be serviced?

b. What is the probability that after servicing ten of these phones, phones of only two of the three types remain to be serviced?

c. What is the probability that two phones of each type are among the first six serviced?

44. Three molecules of type A, three of type B, three of type C, and three of type D are to be linked together to form a chain molecule. One such chain molecule is $ABCDABCDABCD$, and another is $BCDDAAABDBCC$.

a. How many such chain molecules are there? [*Hint:* If the three A's were distinguishable from one another—A_1, A_2, A_3—and the B's, C's, and D's were also, how many molecules would there be? How is this number reduced when the subscripts are removed from the A's?]

b. Suppose a chain molecule of the type described is randomly selected. What is the probability that all three molecules of each type end up next to each other (such as in $BBBAAADDDCCC$)?

45. Three married couples have purchased theater tickets and are seated in a row consisting of just six seats. If they take their seats in a completely random fashion (random order), what is the probability that Jim and Paula (husband and wife) sit in the two seats on the far left? What is the probability that Jim and Paula end up sitting next to one another? What is the probability that at least one of the wives ends up sitting next to her husband?

46. A popular Dilbert cartoon strip (popular among statisticians, anyway) shows an allegedly "random" number generator producing the sequence 999999 with the accompanying comment, "That's the problem with randomness: you can never be sure." Most people would agree that 999999 seems less "random" than, say, 703928, but in what sense is that true? Imagine we randomly generate a six-digit number; i.e., we make six draws with replacement from the digits 0 through 9.

a. What is the probability of generating 999999?
b. What is the probability of generating 703928?
c. What is the probability of generating a sequence of six identical digits?
d. What is the probability of generating a sequence with *no* identical digits? (Comparing the answers to (c) and (d) gives some sense of why some sequences feel intuitively more random than others.)
e. Here's a real challenge: what is the probability of generating a sequence with exactly one repeated digit?

47. Show that $\binom{n}{k} = \binom{n}{n-k}$. Give an interpretation involving subsets.

48. Consider a group of 10 children.

a. How many ways can the children be split into groups of sizes 2, 3, and 5? [*Hint*: First select 2 children from the original 10, then 3 from the remaining 8. Apply the Fundamental Counting Principle.]
b. Verify that your answer to (a) is equivalent to $\frac{10!}{2!3!5!}$.
c. Generalize the previous result by showing that the number of ways to partition n objects into groups of sizes $k_1, \ldots, k_r$ (with $k_1 + \cdots + k_r = n$) is equal to $\frac{n!}{k_1!\cdots k_r!}$.

2.4   Conditional Probability

The probabilities assigned to various events depend on what is known about the experimental situation when the assignment is made. Subsequent to the initial assignment, partial information relevant to the outcome of the experiment may become available. Such information may cause us to revise some of our probability assignments. For a particular event A, we have used $P(A)$ to represent the probability assigned to A; we now think of $P(A)$ as the original or "unconditional" probability of the event A.

In this section, we examine how the information "an event B has occurred" affects the probability assigned to A. For example, A might refer to an individual having a particular disease in the presence of certain symptoms. If a blood test is performed on the individual and the result is negative (B = negative blood test), then the probability of having the disease will change—it should decrease, but not usually to zero, since blood tests are not infallible.

Example 2.25 Complex components are assembled in a plant that uses two different assembly lines, A and A'. Line A uses older equipment than A', so it is somewhat slower and less reliable. Suppose on a given day line A has assembled 8 components, of which 2 have been identified as defective (B) and 6 as nondefective (B'), whereas A' has produced 1 defective and 9 nondefective components. This information is summarized in the accompanying table.

Line	Condition	
	B	B'
A	2	6
A'	1	9

Unaware of this information, the sales manager randomly selects 1 of these 18 components for a demonstration. Prior to the demonstration,

$$P(\text{line } A \text{ component selected}) = P(A) = \frac{N(A)}{N} = \frac{8}{18} = .444$$

However, if the chosen component turns out to be defective, then the event B has occurred, so the component must have been 1 of the 3 in the B column of the table. Since these 3 components are equally likely among themselves, the probability the component was selected from line A, *given that event B has occurred*, is

$$P(A, \text{ given } B) = \frac{2}{3} = \frac{2/18}{3/18} = \frac{P(A \cap B)}{P(B)} \tag{2.2}$$

∎

In Equation (2.2), the conditional probability is expressed as a ratio of unconditional probabilities. The numerator is the probability of the intersection of the two events, whereas the denominator is the probability of the conditioning event B. A Venn diagram illuminates this relationship (Figure 2.9).
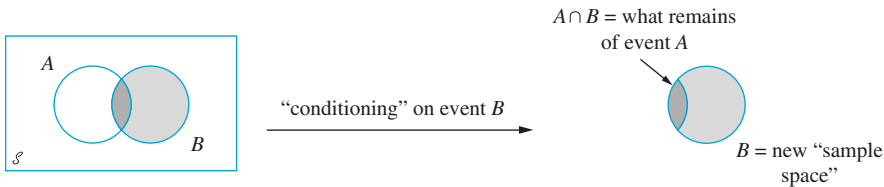
Figure 2.9 Motivating the definition of conditional probability

Given that B has occurred, the relevant sample space is no longer $\mathscr{S}$ but consists of just outcomes in B, and A has occurred if and only if one of the outcomes in the intersection $A \cap B$ occurred. So the conditional probability of A given B should, logically, be the ratio of the likelihoods of these two events.

The Definition of Conditional Probability

Example 2.25 demonstrates that when outcomes are equally likely, computation of conditional probabilities can be based on intuition. When experiments are more complicated, though intuition may fail us, we want to have a general definition of conditional probability that will yield intuitive answers in simple problems. Figure 2.9 and Equation (2.2) suggest the appropriate definition.

DEFINITION   For any two events A and B with $P(B) > 0$, the **conditional probability of A given that B has occurred**, denoted $P(A|B)$, is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.3}$$

Example 2.26 Suppose that of all individuals buying a new iPhone, 60% include a heavy-duty phone case in their purchase, 40% include a portable battery, and 30% include both a heavy-duty case and a portable battery. Consider randomly selecting an iPhone buyer and let A = {heavy-duty case purchased} and B = {portable battery purchased}. Then $P(A) = .60$, $P(B) = .40$, and $P(\text{both purchased}) = P(A \cap B) = .30$. Given that the selected individual purchased a portable battery, the probability that a heavy-duty case was also purchased is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.30}{.40} = .75$$

That is, of all those purchasing a portable battery, 75% purchased a heavy-duty phone case. Similarly,

$$P(\text{battery}|\text{case}) = P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{.30}{.60} = .50$$

Notice that $P(A|B) \neq P(A)$ and $P(B|A) \neq P(B)$. Notice also that $P(A|B) \neq P(B|A)$: these represent two different probabilities computed using different pieces of "given" information. ∎

Example 2.27 A culture website includes three sections entitled "Art" (A), "Books" (B), and "Cinema" (C). Reading habits of a randomly selected reader with respect to these sections are

Read regularly	A	B	C	$A \cap B$	$A \cap C$	$B \cap C$	$A \cap B \cap C$
Probability	.14	.23	.37	.08	.09	.13	.05

Figure 2.10 encapsulates this information.
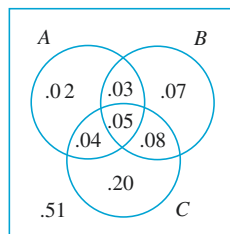
Figure 2.10   Venn diagram for Example 2.27

We thus have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.08}{.23} = .348$$

$$P(A|B \cup C) = \frac{P(A \cap (B \cup C))}{P(B \cup C)} = \frac{.04 + .05 + .03}{.47} = \frac{.12}{.47} = .255$$

$$P(A|\text{reads at least one}) = P(A|A \cup B \cup C) = \frac{P(A \cap (A \cup B \cup C))}{P(A \cup B \cup C)}$$

$$= \frac{P(A)}{P(A \cup B \cup C)} = \frac{.14}{.49} = .286$$

and

$$P(A \cup B|C) = \frac{P((A \cup B) \cap C)}{P(C)} = \frac{.04 + .05 + .08}{.37} = .459 \qquad \blacksquare$$

The Multiplication Rule for $P(A \cap B)$

The definition of conditional probability yields the following result, obtained by multiplying both sides of Equation (2.3) by $P(B)$.

MULTIPLICATION RULE $\qquad\qquad P(A \cap B) = P(A|B) \cdot P(B)$

This rule is important because it is often the case that $P(A \cap B)$ is desired, whereas both $P(B)$ and $P(A|B)$ can be specified from the problem description. By reversing the roles of A and B, the Multiplication Rule can also be written as $P(A \cap B) = P(B|A) \cdot P(A)$.

Example 2.28 Four individuals have responded to a request by a blood bank for blood donations. None of them has donated before, so their blood types are unknown. Suppose only type O+ is desired and only one of the four actually has this type. If the potential donors are selected in random order for typing, what is the probability that at least three individuals must be typed to obtain the desired type?

Define $B = \{$first type not O+$\}$ and $A = \{$second type not O+$\}$. Since three of the four potential donors are not O+, $P(B) = 3/4$. Given that the first person typed is not O+, two of the three individuals left are not O+, and so $P(A|B) = 2/3$. The Multiplication Rule now gives

$$\begin{aligned}
P(\text{at least three individuals are typed}) &= P(\text{first two typed are not O+}) \\
&= P(A \cap B) \\
&= P(A|B) \cdot P(B) \\
&= \frac{2}{3} \cdot \frac{3}{4} = \frac{6}{12} \\
&= .5 \qquad\qquad \blacksquare
\end{aligned}$$

The Multiplication Rule is most useful when the experiment consists of several stages in succession. The conditioning event B then describes the outcome of the first stage and A the outcome of the second, so that $P(A|B)$—conditioning on what occurs first—will often be known. The rule is easily extended to experiments involving more than two stages. For example,

$$\begin{aligned}
P(A_1 \cap A_2 \cap A_3) &= P(A_3|A_1 \cap A_2) \cdot P(A_1 \cap A_2) \\
&= P(A_3|A_1 \cap A_2) \cdot P(A_2|A_1) \cdot P(A_1)
\end{aligned} \qquad (2.4)$$

where A_1 occurs first, followed by A_2, and finally A_3.

Example 2.29 Using Equation (2.4) for the blood typing experiment of Example 2.28,

$P(\text{third type is O+})$

$\quad = P(\text{third is} \mid \text{first isn't} \cap \text{second isn't}) \cdot P(\text{second isn't} \mid \text{first isn't}) \cdot P(\text{first isn't})$

$\quad = \dfrac{1}{2} \cdot \dfrac{2}{3} \cdot \dfrac{3}{4} = \dfrac{1}{4} = .25$ ∎

When the experiment of interest consists of a sequence of several stages, it is convenient to represent these with a tree diagram. Once we have an appropriate tree diagram, probabilities and conditional probabilities can be entered on the various branches; this will make repeated use of the Multiplication Rule quite straightforward.

Example 2.30 An online retailer sells three different brands of Bluetooth earbuds. Of its earbud sales, 50% are brand 1 (the least expensive), 30% are brand 2, and 20% are brand 3. Each manufacturer offers a 1-year warranty. It is known that 25% of brand 1's earbuds will be returned within the 1-year warranty period, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively.

1. What is the probability that a randomly selected purchaser has bought brand 1 earbuds that will be returned while under warranty?
2. What is the probability that a randomly selected purchaser has earbuds that will be returned while under warranty?
3. If a customer returns earbuds under warranty, what is the probability that they are brand 1 earbuds? Brand 2? Brand 3?

The first stage of the problem involves a customer selecting one of the three brands of earbud. Let $A_i = \{\text{brand } i \text{ is purchased}\}$, for $i = 1$, 2, and 3. Then $P(A_1) = .50$, $P(A_2) = .30$, and $P(A_3) = .20$. Once a brand of earbud is selected, the second stage involves observing whether the selected earbuds get returned during the warranty period. With $B = \{\text{returned}\}$ and $B' = \{\text{not returned}\}$, the given information implies that $P(B|A_1) = .25$, $P(B|A_2) = .20$, and $P(B|A_3) = .10$.

The tree diagram representing this experimental situation appears in Figure 2.11 (p. 80). The initial branches correspond to different brands of earbuds; there are two second-generation branches emanating from the tip of each initial branch, one for "returned" and the other for "not returned." The probability $P(A_i)$ appears on the ith initial branch, whereas the conditional probabilities $P(B|A_i)$ and $P(B'|A_i)$ appear on the second-generation branches. To the right of each second-generation branch corresponding to the occurrence of B, we display the product of probabilities on the branches leading out to that point. This is simply the Multiplication Rule in action. The answer to question 1 is thus $P(A_1 \cap B) = P(B|A_1) \cdot P(A_1) = .125$. The answer to question 2 is

$$P(B) = P[(\text{brand } 1 \text{ and returned}) \text{ or } (\text{brand 2 and returned}) \text{ or } (\text{brand 3 and returned})]$$
$$= P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$
$$= .125 + .060 + .020 = .205$$

Finally,

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.125}{.205} = .61$$
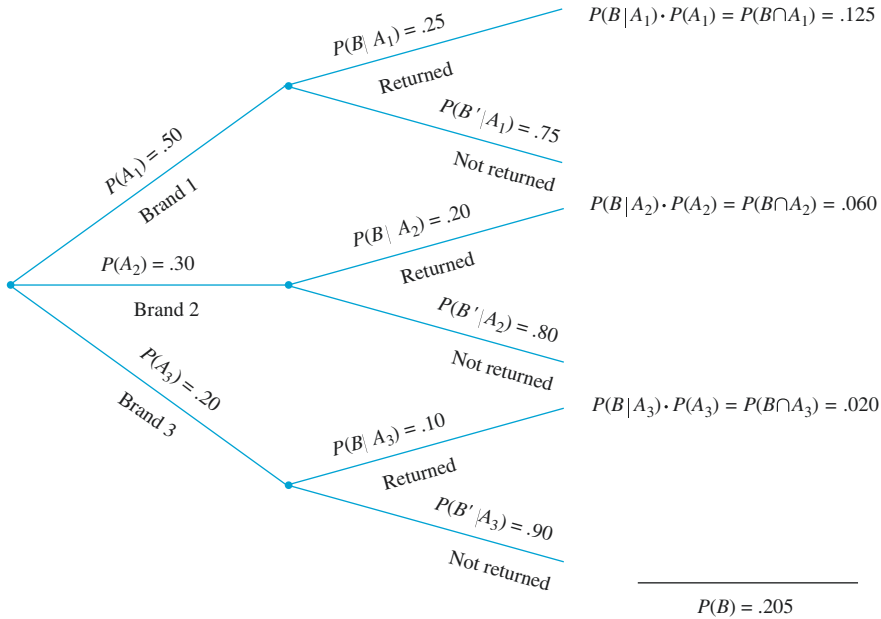
Figure 2.11 Tree diagram for Example 2.30

$$P(A_2|B) = \frac{P(A_2 \cap B)}{P(B)} = \frac{.060}{.205} = .29$$

and

$$P(A_3|B) = 1 - P(A_1|B) - P(A_2|B) = .10$$

Notice that the initial or **prior probability** of brand 1 is .50, whereas once it is known that the selected earbuds were returned, the **posterior probability** of brand 1 increases to .61. This is because brand 1 earbuds are more likely to be returned under warranty than are the other brands. In contrast, the posterior probability of brand 3 is $P(A_3|B) = .10$, which is much less than the prior probability $P(A_3) = .20$. ∎

The Law of Total Probability and Bayes' Theorem

The computation of a posterior probability $P(A_j|B)$ from given prior probabilities $P(A_i)$ and conditional probabilities $P(B|A_i)$ occupies a central position in elementary probability. The general rule for such computations, which is really just a simple application of the Multiplication Rule, goes back to the Reverend Thomas Bayes, who lived in the eighteenth century. To state it we first need another result. Recall that events $A_1, \ldots, A_k$ are mutually exclusive if no two have any common outcomes. The events are *exhaustive* if $A_1 \cup \cdots \cup A_k = \mathcal{S}$, so that one A_i must occur.

LAW OF TOTAL PROBABILITY	Let $A_1, \ldots, A_k$ be mutually exclusive and exhaustive events. Then for any other event B,

$$P(B) = P(B|A_1) \cdot P(A_1) + \cdots + P(B|A_k) \cdot P(A_k)$$
$$= \sum_{i=1}^{k} P(B|A_i)P(A_i) \tag{2.5}$$

Proof Because the A_i's are mutually exclusive and exhaustive, if B occurs it must be in conjunction with exactly one of the A_i's. That is, $B = (A_1 \text{ and } B)$ or ... or $(A_k \text{ and } B) = (A_1 \cap B) \cup \cdots \cup (A_k \cap B)$, where the events $(A_i \cap B)$ are mutually exclusive. This "partitioning of B" is illustrated in Figure 2.12. Thus

$$P(B) = \sum_{i=1}^{k} P(A_i \cap B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

as desired.
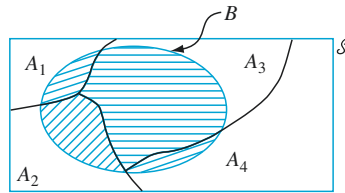
Figure 2.12 Partition of B by mutually exclusive and exhaustive A_i's ∎

An example of the use of Equation (2.5) appeared in answering question 2 of Example 2.30, where $A_1 = \{\text{brand } 1\}$, $A_2 = \{\text{brand } 2\}$, $A_3 = \{\text{brand } 3\}$, and $B = \{\text{returned}\}$.

Example 2.31 A certain university has three colleges: Letters & Science (45% of the student body), Business (32%), and Engineering (23%). Of the students in the College of Letters & Science, 11% traveled out of state during the most recent spring break, compared to 14% in Business and just 3% in Engineering. If we select a student completely at random from this student body, what's the probability he/she traveled out of state for spring break?

Define $A_1 = \{\text{the student belongs to Letters \& Science}\}$; define A_2 and A_3 similarly for Business and Engineering, respectively. Let $B = \{\text{the student traveled out of state for spring break}\}$. The percentages provided above imply that

$$P(A_1) = .45 \quad P(A_2) = .32 \quad P(A_3) = .23$$
$$P(B|A_1) = .11 \quad P(B|A_2) = .14 \quad P(B|A_3) = .03$$

Notice that A_1, A_2, A_3 form a partition of the sample space (the student body). Apply the Law of Total Probability:

$$P(B) = (.11)(.45) + (.14)(.32) + (.03)(.23) = .1012 \qquad \blacksquare$$

BAYES' THEOREM   Let $A_1, \ldots, A_k$ be a collection of mutually exclusive and exhaustive events with $P(A_i) > 0$ for $i = 1, \ldots, k$. Then for any other event B for which $P(B) > 0$,

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)} \quad j = 1, \ldots, k \qquad (2.6)$$

The transition from the second to the third expression in (2.6) rests on using the Multiplication Rule in the numerator and the Law of Total Probability in the denominator.

The proliferation of events and subscripts in (2.6) can be a bit intimidating to probability newcomers. When $k = 2$, so that the partition of $\mathscr{S}$ consists of just $A_1 = A$ and $A_2 = A'$, Bayes' Theorem becomes

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A')P(B|A')}$$

As long as there are relatively few events in the partition, a tree diagram (as in Example 2.30) can be used as a basis for calculating posterior probabilities without ever referring explicitly to Bayes' theorem.

Example 2.32 *Incidence of a rare disease.* Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test has been developed. The test is such that when an individual actually has the disease, a positive result will occur 99% of the time, whereas an individual without the disease will show a positive test result only 2% of the time. If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

[*Note*: The *sensitivity* of this test is 99%, whereas the *specificity*—how specific positive results are to this disease—is 98%. As an indication of the accuracy of medical tests, an article in the October 29, 2010, *New York Times* reported that the sensitivity and specificity for a new DNA test for colon cancer were 86% and 93%, respectively. The PSA test for prostate cancer has sensitivity 85% and specificity about 30%, while the mammogram for breast cancer has sensitivity 75% and specificity 92%. And then there are Covid19 tests. All tests are less than perfect.]

To use Bayes' theorem, let $A_1 = \{$individual has the disease$\}$, $A_2 = \{$individual does not have the disease$\}$, and $B = \{$positive test result$\}$. Then $P(A_1) = .001$, $P(A_2) = .999$, $P(B|A_1) = .99$, and $P(B|A_2) = .02$. The tree diagram for this problem is in Figure 2.13 (p. 83).

Next to each branch corresponding to a positive test result, the Multiplication Rule yields the recorded probabilities. Therefore, $P(B) = .00099 + .01998 = .02097$, from which we have

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.00099}{.02097} = .047$$

This result seems counterintuitive; because the diagnostic test appears so accurate, we expect someone with a positive test result to be highly likely to have the disease, whereas the computed conditional probability is only .047. However, because the disease is rare and the test isn't perfectly reliable, most positive test results arise from errors rather than from diseased individuals. The probability of having the disease has increased by a multiplicative factor of 47 (from prior .001 to posterior .047); but to get a further increase in the posterior probability, a diagnostic test with much
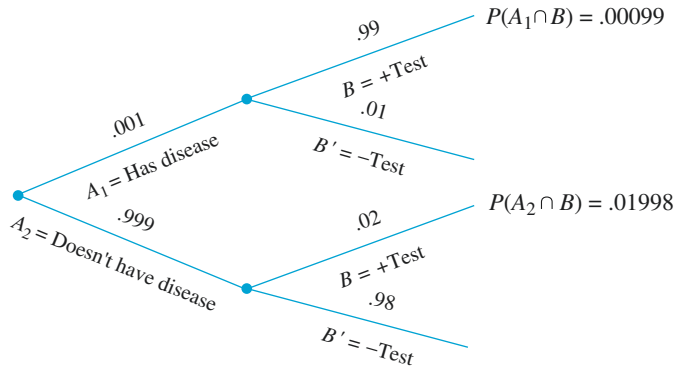
Figure 2.13   Tree diagram for the rare disease problem

smaller error rates is needed. If the disease were not so rare (e.g., 25% incidence in the population), then the error rates for the present test would provide good diagnoses.

   This example shows why it makes sense to be tested for a rare disease only if you are in a high-risk group. For example, most of us are at low risk for HIV infection, so testing would not be indicated, but those who are in a high-risk group should be tested for HIV. For some diseases the degree of risk is strongly influenced by age. Young women are at low risk for breast cancer and should not be tested, but older women do have increased risk and need to be tested. There is some argument about where to draw the line. If we can find the incidence rate for our group and the sensitivity and specificity for the test, then we can do our own calculation to see if a positive test result would be informative.   ■

   An important contemporary application of Bayes' theorem is in the identification of spam e-mail messages. A nice expository article on this appears in *Statistics: A Guide to the Unknown* (see the bibliography).

Exercises: Section 2.4 (49–73)

49. The population of a particular country consists of three ethnic groups. Each individual belongs to one of the four major blood groups. The accompanying *joint probability table* gives the proportions of individuals in the various ethnic group–blood group combinations.

Ethnic group	Blood group			
	O	A	B	AB
1	.082	.106	.008	.004
2	.135	.141	.018	.006
3	.215	.200	.065	.020

Suppose that an individual is randomly selected from the population, and define events by A = {type A selected}, B = {type B selected}, and C = {ethnic group 3 selected}.

a. Calculate $P(A)$, $P(C)$, and $P(A \cap C)$.
b. Calculate both $P(A|C)$ and $P(C|A)$ and explain in context what each of these probabilities represents.
c. If the selected individual does not have type B blood, what is the probability that he or she is from ethnic group 1?

50. Suppose an individual is randomly selected from the population of all adult males living in the United States. Let A be the event that the selected individual is over 6 ft in height, and let B be the event that the selected individual is a professional basketball player. Which do you think is larger, $P(A|B)$ or $P(B|A)$? Why?

51. Return to the credit card scenario of Exercise 14, where A = {Visa}, B = {MasterCard}, $P(A)$ = .5, $P(B)$ = .4, and $P(A \cap B)$ = .25.

Calculate and interpret each of the following probabilities (a Venn diagram might help).

a. $P(B|A)$
b. $P(B'|A)$
c. $P(A|B)$
d. $P(A'|B)$
e. Given that the selected individual has at least one card, what is the probability that he or she has a Visa card?

52. Reconsider the system defect situation described in Exercise 28.

a. Given that the system has a type 1 defect, what is the probability that it has a type 2 defect?
b. Given that the system has a type 1 defect, what is the probability that it has all three types of defects?
c. Given that the system has at least one type of defect, what is the probability that it has exactly one type of defect?
d. Given that the system has both of the first two types of defects, what is the probability that it does not have the third type of defect?

53. If two bulbs are randomly selected from the box of lightbulbs described in Exercise 42 and at least one of them is found to be rated 75 W, what is the probability that both of them are 75-W bulbs? Given that at least one of the two selected is not rated 75 W, what is the probability that both selected bulbs have the same rating?

54. A department store sells sport shirts in three sizes (small, medium, and large), three patterns (plaid, print, and stripe), and two sleeve lengths (long and short). The accompanying tables give the proportions of shirts sold in the various category combinations.

Short-sleeved

| | Pattern | | |
Size	Pl	Pr	St
S	.04	.02	.05
M	.08	.07	.12
L	.03	.07	.08

Long-sleeved

| | Pattern | | |
Size	Pl	Pr	St
S	.03	.02	.03
M	.10	.05	.07
L	.04	.02	.08

a. What is the probability that the next shirt sold is a medium, long-sleeved, print shirt?
b. What is the probability that the next shirt sold is a medium print shirt?
c. What is the probability that the next shirt sold is a short-sleeved shirt? A long-sleeved shirt?
d. What is the probability that the size of the next shirt sold is medium? That the pattern of the next shirt sold is a print?
e. Given that the shirt just sold was a short-sleeved plaid, what is the probability that its size was medium?
f. Given that the shirt just sold was a medium plaid, what is the probability that it was short-sleeved? Long-sleeved?

55. One box contains six red balls and four green balls, and a second box contains seven red balls and three green balls. A ball is randomly chosen from the first box and placed in the second box. Then a ball is randomly selected from the second box and placed in the first box.

a. What is the probability that a red ball is selected from the first box and a red ball is selected from the second box?
b. At the conclusion of the selection process, what is the probability that the numbers of red and green balls in the first box are identical to the numbers at the beginning?

56. A system consists of two identical pumps, #1 and #2. If one pump fails, the system will still operate. However, because of the added strain, the extra remaining pump is now more likely to fail than was originally the case. That is, $r = P(\#2 \text{ fails}|\#1 \text{ fails}) > P(\#2 \text{ fails}) = q$. If at least one pump fails by

the end of the pump design life in 7% of all systems and both pumps fail during that period in only 1%, what is the probability that pump #1 will fail during the pump design life?

57. A certain shop repairs both audio and video components. Let A denote the event that the next component brought in for repair is an audio component, and let B be the event that the next component is an MP3 player (so the event B is contained in A). Suppose that $P(A) = .6$ and $P(B) = .05$. What is $P(B|A)$?

58. In Exercise 15, A_i = {awarded project i}, for $i = 1, 2, 3$. Use the probabilities given there to compute the following probabilities, and explain in words the meaning of each one.

   a. $P(A_2|A_1)$
   b. $P(A_2 \cap A_3|A_1)$
   c. $P(A_2 \cup A_3|A_1)$
   d. $P(A_1 \cap A_2 \cap A_3|A_1 \cup A_2 \cup A_3)$

59. Refer back to the culture website scenario in Example 2.27.

   a. Given that someone regularly reads at least one of the three sections listed (Arts, Books, Cinema), what is the probability she reads all three?
   b. Given that someone regularly reads all three sections, what is the probability she reads at least one? [Think carefully!]

60. Three plants manufacture hard drives and ship them to a warehouse for distribution. Plant I produces 54% of the warehouse's inventory with a 4% defect rate. Plant II produces 35% of the warehouse's inventory with an 8% defect rate. Plant III produces the remainder of the warehouse's inventory with a 12% defect rate.

   a. Draw a tree diagram to represent this information.
   b. A warehouse inspector selects one hard drive at random. What is the probability that it is a defective hard drive and from Plant II?

   c. What is the probability that a randomly selected hard drive is defective?
   d. Suppose a hard drive is defective. What is the probability that it came from Plant II?

61. For any events A and B with $P(B) > 0$, show that $P(A \mid B) + P(A' \mid B) = 1$.

62. If $P(B|A) > P(B)$ show that $P(B'|A) < P(B')$. [Hint: Add $P(B'|A)$ to both sides of the given inequality, and then use the result of the previous exercise.]

63. Show that for any three events A, B, and C with $P(C) > 0$, $P(A \cup B|C) = P(A \mid C) + P(B \mid C) - P(A \cap B \mid C)$.

64. At a gas station, 40% of the customers use regular gas (A_1), 35% use midgrade gas (A_2), and 25% use premium gas (A_3). Of those customers using regular gas, only 30% fill their tanks (event B). Of those customers using midgrade gas, 60% fill their tanks, whereas of those using premium, 50% fill their tanks.

   a. What is the probability that the next customer will request midgrade gas and fill the tank ($A_2 \cap B$)?
   b. What is the probability that the next customer fills the tank?
   c. If the next customer fills the tank, what is the probability that regular gas is requested? Midgrade gas? Premium gas?

65. Seventy percent of the light aircraft that disappear while in flight in a certain country are subsequently discovered. Of the aircraft that are discovered, 60% have an emergency locator, whereas 90% of the aircraft not discovered do not have such a locator. Suppose a light aircraft has disappeared.

   a. If it has an emergency locator, what is the probability that it will not be discovered?
   b. If it does not have an emergency locator, what is the probability that it will be discovered?

66. Components of a certain type are shipped to a supplier in batches of ten. Suppose that 50% of all such batches contain no defective components, 30% contain one defective component, and 20% contain two defective components. Two components from a batch are randomly selected and tested. What are the probabilities associated with 0, 1, and 2 defective components being in the batch under each of the following conditions?

    a. Neither tested component is defective.
    b. One of the two tested components is defective.
    [*Hint*: Draw a tree diagram with three first-generation branches for the three different types of batches.]

67. Verify the multiplication rule for conditional probabilities: $P(A \cap B|C) = P(A|B \cap C) \cdot P(B|C)$.

68. For customers purchasing a full set of tires at a particular tire store, consider the events

    A = {tires purchased were made in the United States}
    B = {purchaser has tires balanced immediately}
    C = {purchaser requests front-end alignment}

    along with A', B', and C'. Assume the following unconditional and conditional probabilities:

    $P(A) = .75 \quad P(B|A) = .9 \quad P(B|A') = .8$
    $P(C|A \cap B) = .8 \quad P(C|A \cap B') = .6$
    $P(C|A' \cap B) = .7 \quad P(C|A' \cap B') = .3$

    a. Construct a tree diagram consisting of first-, second-, and third-generation branches, and place an event label and appropriate probability next to each branch.
    b. Compute $P(A \cap B \cap C)$.
    c. Compute $P(B \cap C)$.
    d. Compute $P(C)$.
    e. Compute $P(A|B \cap C)$, the probability of a purchase of U.S. tires given that both balancing and an alignment were requested.

69. A professional organization (for statisticians, of course) sells term life insurance and major medical insurance. Of those who have just life insurance, 70% will renew next year, and 80% of those with only a major medical policy will renew next year. However, 90% of policyholders who have both types of policy will renew at least one of them next year. Of the policyholders 75% have term life insurance, 45% have major medical, and 20% have both.

    a. Calculate the percentage of policyholders that will renew at least one policy next year.
    b. If a randomly selected policyholder does in fact renew next year, what is the probability that he or she has both types of policies?

70. At a large university, in the never-ending quest for a satisfactory textbook, the Statistics Department has tried a different text during each of the last three quarters. During the fall quarter, 500 students used the text by Professor Mean; during the winter quarter, 300 students used the text by Professor Median; and during the spring quarter, 200 students used the text by Professor Mode. A survey at the end of each quarter showed that 200 students were satisfied with Mean's book, 150 were satisfied with Median's book, and 160 were satisfied with Mode's book. If a student who took statistics during one of these quarters is selected at random and admits to having been satisfied with the text, is the student most likely to have used the book by Mean, Median, or Mode? Who is the least likely author? [*Hint*: Draw a tree diagram or use Bayes' theorem.]

71. A friend who lives in Los Angeles makes frequent consulting trips to Washington, D.C.; 50% of the time she travels on airline #1, 30% of the time on airline #2, and the remaining 20% of the time on airline #3. For airline #1, flights are late into D.C. 30% of the time and late into L.A. 10% of the time. For airline #2, these percentages are

25 and 20%, whereas for airline #3 the percentages are 40 and 25%. If we learn that on a particular trip she arrived late at exactly one of the two destinations, what are the posterior probabilities of having flown on airlines #1, #2, and #3? Assume that the chance of a late arrival in L.A. is unaffected by what happens on the flight to D.C. [*Hint*: From the tip of each first-generation branch on a tree diagram, draw three second-generation branches labeled, respectively, 0 late, 1 late, and 2 late.]

72. Suppose a single gene controls the color of hamsters: black (*B*) is dominant and brown (*b*) is recessive. Hence, a hamster will be black unless its genotype is *bb*. Two hamsters, each with genotype *Bb*, mate and produce a single offspring. The laws of genetic recombination state that each parent is equally likely to donate either of its two alleles (*B* or *b*), so the offspring is equally likely to be any of *BB*, *Bb*, *bB*, or *bb* (the middle two are genetically equivalent).

   a. What is the probability their offspring has black fur?
   b. Given that their offspring has black fur, what is the probability its genotype is *Bb*?

73. Refer back to the scenario of the previous exercise. In the figure below, the genotypes of both members of Generation I are known, as is the genotype of the male member of Generation II. We know that hamster II2 must be black-colored thanks to her father, but suppose that we don't know her genotype exactly (as indicated by *B* in the figure).
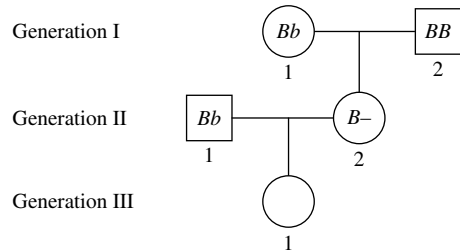
a. What are the possible genotypes of hamster II2, and what are the corresponding probabilities?
b. If we observe that hamster III1 has a black coat (and hence at least one *B* gene), what is the probability her genotype is *Bb*?
c. If we later discover (through DNA testing on poor little hamster III1) that her genotype is *BB*, what is the posterior probability that her mom is also *BB*?

2.5   Independence

The definition of conditional probability enables us to revise the probability $P(A)$ originally assigned to A when we are subsequently informed that another event B has occurred; the new probability of A is $P(A|B)$. In our examples, it was frequently the case that $P(A|B)$ differed from the unconditional probability $P(A)$, indicating that the information "B has occurred" resulted in a change in the chance of A occurring. There are other situations, though, in which the chance that A will occur or has occurred is not affected by knowledge that B has occurred, so that $P(A|B) = P(A)$. It is then natural to think of A and B as independent events, meaning that the occurrence or nonoccurrence of one event has no bearing on the chance that the other will occur.

DEFINITION   Two events A and B are **independent** if $P(A|B) = P(A)$ and are **dependent** otherwise.

The definition of independence might seem "unsymmetrical" because we do not demand that $P(B|A) = P(B)$ also. However, using the definition of conditional probability and the Multiplication Rule,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} \qquad (2.7)$$

The right-hand side of Equation (2.7) is $P(B)$ if and only if $P(A|B) = P(A)$ (independence), so the equality in the definition implies the other equality (and vice versa). It is also straightforward to show that if A and B are independent, then so are the following pairs of events: (1) A' and B, (2) A and B', and (3) A' and B' (see Exercise 77).

Example 2.33   Consider an ordinary deck of 52 cards comprised of the four "suits" spades, hearts, diamonds, and clubs, with each suit consisting of the 13 ranks ace, king, queen, jack, ten, …, and two. Suppose someone randomly selects a card from the deck and reveals to you that it is a face card (that is, a king, queen, or jack). What now is the probability that the card is a spade? If we let $A = \{\text{spade}\}$ and $B = \{\text{face card}\}$, then $P(A) = 13/52$, $P(B) = 12/52$ (there are three face cards in each of the four suits), and $P(A \cap B) = P(\text{spade and face card}) = 3/52$. Thus

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3/52}{12/52} = \frac{3}{12} = \frac{1}{4} = \frac{13}{52} = P(A)$$

Therefore, the likelihood of getting a spade is not affected by knowledge that a face card had been selected. Intuitively this is because the fraction of spades among face cards (3 out of 12) is the same as the fraction of spades in the entire deck (13 out of 52). It is also easily verified that $P(B|A) = P(B)$, so knowledge that a spade has been selected does not affect the likelihood of the card being a jack, queen, or king. ∎

Example 2.34   Let A and B be any two mutually exclusive events with $P(A) > 0$. For example, for a randomly chosen automobile, let $A = \{\text{car is blue}\}$ and $B = \{\text{car is red}\}$. Since the events are mutually exclusive, if B occurs, then A cannot possibly have occurred, so $P(A|B) = 0 \neq P(A)$. The message here is that *if two events are mutually exclusive, they cannot be independent*. When A and B are mutually exclusive, the information that A occurred says something about B (it cannot have occurred), so independence is precluded. ∎

$P(A \cap B)$ When Events Are Independent

Frequently the nature of an experiment suggests that two events A and B should be assumed independent. This is the case, for example, if a manufacturer receives a circuit board from each of two different suppliers, each board is tested on arrival, and $A = \{\text{first is defective}\}$ and $B = \{\text{second is defective}\}$. If $P(A) = .1$, it should also be the case that $P(A|B) = .1$; knowing the condition of the second board shouldn't provide information about the condition of the first. Our next result shows how to compute $P(A \cap B)$ when the events are independent.

PROPOSITION A and B are independent if and only if

$$P(A \cap B) = P(A) \cdot P(B) \qquad (2.8)$$

Proof By the Multiplication Rule, $P(A \cap B) = P(A|B) \cdot P(B)$, and this equals $P(A) \cdot P(B)$ if and only if $P(A|B) = P(A)$. ∎

Because of the equivalence of independence with Equation (2.8), the latter can be used as a definition of independence.[1]

Example 2.35 It is known that 3% of a certain machine tool manufacturing company's band saws break down within the first six months of ownership, compared to only 1% of its industrial lathes. If a machine shop purchases both a band saw and a lathe made by this company, what is the probability that both machines will break down within six months?

Let A denote the event that the band saw breaks down in the first six months, and define B analogously for the industrial lathe. Then $P(A) = .03$ and $P(B) = .01$. Assuming that the two machines function independently of each other, the desired probability is

$$P(A \cap B) = P(A) \cdot P(B) = (.03)(.01) = .0003$$

The probability that neither machine breaks down in that time period is

$$P(A' \cap B') = P(A') \cdot P(B') = (.97)(.99) = .9603$$

Note that, although the independence assumption is reasonable here, it can be questioned. In particular, if heavy use causes a breakdown in one machine, it could also cause trouble for the other one. ∎

Example 2.36 Each day, Monday through Friday, a batch of components sent by a first supplier arrives at a certain inspection facility. Two days a week, a batch also arrives from a second supplier. Eighty percent of all supplier 1's batches pass inspection, and 90% of supplier 2's do likewise. What is the probability that, on a randomly selected day, two batches pass inspection? We will answer this assuming that on days when two batches are tested, whether the first batch passes is independent of whether the second batch does so. Figure 2.14 displays the relevant information.
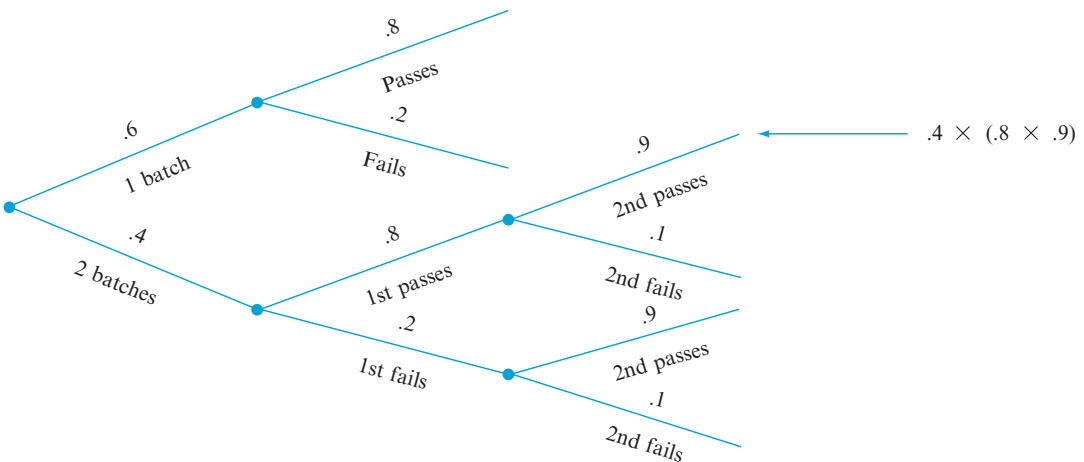
Figure 2.14 Tree diagram for Example 2.36

---

[1]However, the multiplication property is satisfied if $P(B) = 0$, yet $P(A|B)$ is not defined in this case. To make the multiplication property completely equivalent to the definition of independence, we should append to that definition that A and B are also independent if either $P(A) = 0$ or $P(B) = 0$.

$$P(\text{two pass}) = P(\text{two received} \cap \text{both pass})$$
$$= P(\text{both pass}|\text{two received}) \cdot P(\text{two received})$$
$$= [(.8)(.9)](.4) = .288 \qquad \blacksquare$$

Independence of More Than Two Events

The notion of independence of two events can be extended to collections of more than two events. Although it is possible to extend the definition for two independent events by working in terms of conditional and unconditional probabilities, it is more direct and less cumbersome to proceed along the lines of the last proposition.

DEFINITION   Events $A_1, \ldots, A_n$ are **mutually independent** if for every k ($k = 2, 3, \ldots, n$) and every subset of distinct indices $i_1, i_2, \ldots, i_k$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \cdots \cdot P(A_{i_k})$$

To paraphrase the definition, the events are mutually independent if the probability of the intersection of any subset of the n events is equal to the product of the individual probabilities. In using this multiplication property for more than two independent events, it is legitimate to replace one or more of the A_i's by their complements (e.g., if A_1, A_2, and A_3 are independent events, then so are A_1', A_2', and A_3'). As was the case with two events, we frequently specify at the outset of a problem the independence of certain events. The definition can then be used to calculate the probability of an intersection.

Example 2.37 The article "Reliability Evaluation of Solar Photovoltaic Arrays" (*Solar Energy* 2002: 129–141) presents various configurations of solar photovoltaic arrays consisting of crystalline silicon solar cells. Consider first the system illustrated in Figure 2.15a. There are two subsystems connected in parallel, each one containing three cells. In order for the system to function, at least one of the two parallel subsystems must work. Within each subsystem, the three cells are connected in series, so a subsystem will work only if all cells in the subsystem work. Consider a particular lifetime value t_0, and suppose we want to determine the probability that the system lifetime exceeds t_0. Let A_i denote the event that the lifetime of cell i exceeds t_0 ($i = 1, 2, \ldots, 6$). We assume that the A_i's are independent events (whether any particular cell lasts more than t_0 hours has no bearing on whether any other cell does) and that $P(A_i) = .9$ for every i since the cells are all made the same way.
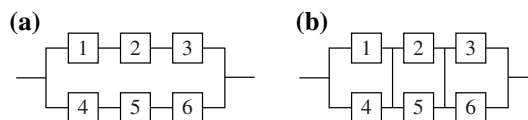
Figure 2.15   System configurations for Example 2.37: (**a**) series–parallel; (**b**) total-cross-tied

Then

$$
\begin{aligned}
P(\text{system lifetime exceeds } t_0) &= P[(A_1 \cap A_2 \cap A_3) \cup (A_4 \cap A_5 \cap A_6)] \\
&= P(A_1 \cap A_2 \cap A_3) + P(A_4 \cap A_5 \cap A_6) \\
&\quad - P[(A_1 \cap A_2 \cap A_3) \cap (A_4 \cap A_5 \cap A_6)] \\
&= (.9)(.9)(.9) + (.9)(.9)(.9) - (.9)(.9)(.9)(.9)(.9)(.9) \\
&= .927
\end{aligned}
$$

Alternatively,

$$
\begin{aligned}
P(\text{system lifetime exceeds } t_0) &= 1 - P(\text{both subsystem lives are} \le t_0) \\
&= 1 - [P(\text{subsystem life is} \le t_0)]^2 \\
&= 1 - [1 - P(\text{subsystem life is} > t_0)]^2 \\
&= 1 - [1 - .9^3]^2 = .927
\end{aligned}
$$

Next consider the total-cross-tied system shown in Figure 2.15b, obtained from the series–parallel array by connecting ties across each column of junctions. Now the system fails as soon as an entire column fails, and system lifetime exceeds t_0 only if the life of every column does so. For this configuration,

$$
\begin{aligned}
P(\text{system lifetime exceeds } t_0) &= [P(\text{column lifetime exceeds } t_0)]^3 \\
&= [1 - P(\text{column lifetime is} \le t_0)]^3 \\
&= [1 - P(\text{both cells in a column have lifetime} \le t_0)]^3 \\
&= [1 - (1 - .9)^2]^3 = .970 \qquad \blacksquare
\end{aligned}
$$

Exercises: Section 2.5 (74–92)

74. Reconsider the credit card scenario of Exercise 51 (Section 2.4), and show that A and B are dependent first by using the definition of independence and then by verifying that the multiplication property does not hold.

75. An oil exploration company currently has two active projects, one in Asia and the other in Europe. Let A be the event that the Asian project is successful and B be the event that the European project is successful. Suppose that A and B are independent events with $P(A) = .4$ and $P(B) = .7$.

   a. If the Asian project is not successful, what is the probability that the European project is also not successful? Explain your reasoning.

   b. What is the probability that at least one of the two projects will be successful?

   c. Given that at least one of the two projects is successful, what is the probability that only the Asian project is successful?

76. In Exercise 15, is any A_i independent of any other A_i? Answer using the multiplication property for independent events.

77. If A and B are independent events, show that A' and B are also independent. [*Hint*: First establish a relationship among $P(A' \cap B)$, $P(B)$, and $P(A \cap B)$.]

78. Suppose that the proportions of blood phenotypes in a particular population are as follows:

A	B	AB	O
.42	.10	.04	.44

Assuming that the phenotypes of two randomly selected individuals are independent of each other, what is the probability that both phenotypes are O? What is the probability that the phenotypes of two randomly selected individuals match?

79. The probability that a grader will make a marking error on any particular question of a multiple-choice exam is .1. If there are ten questions and questions are marked independently, what is the probability that no errors are made? That at least one error is made? If there are n questions and the probability of a marking error is p rather than .1, give expressions for these two probabilities.

80. An aircraft seam requires 25 rivets. The seam will have to be reworked if any of these rivets is defective. Suppose rivets are defective independently of one another, each with the same probability.

   a. If 20% of all seams need reworking, what is the probability that a rivet is defective?
   b. How small should the probability of a defective rivet be to ensure that only 10% of all seams need reworking?

81. A boiler has five identical relief valves. The probability that any particular valve will open on demand is .95. Assuming independent operation of the valves, calculate P(at least one valve opens) and P(at least one valve fails to open).

82. Two pumps connected in parallel fail independently of each other on any given day. The probability that only the older pump will fail is .10, and the probability that only the newer pump will fail is .05. What is the probability that the pumping system will fail on any given day (which happens if both pumps fail)?

83. Consider the system of components connected as in the accompanying picture.
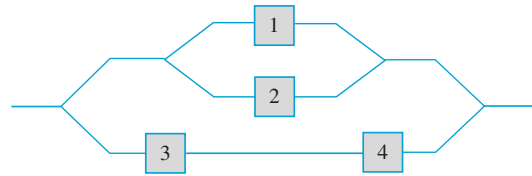
Components 1 and 2 are connected in parallel, so that subsystem works if and only if either 1 or 2 works; since 3 and 4 are connected in series, that subsystem works if and only if both 3 and 4 work. If components work independently of one another and P(component works) = .9, calculate P(system works).

84. Refer back to the series–parallel system configuration introduced in Example 2.37, and suppose that there are only two cells rather than three in each parallel subsystem [in Figure 2.15a, eliminate cells 3 and 6, and re-number cells 4 and 5 as 3 and 4]. Using $P(A_i)$ = .9, the probability that system lifetime exceeds t_0 is easily seen to be .9639. To what value would .9 have to be changed in order to increase the system lifetime reliability from .9639 to .99? [*Hint*: Let $P(A_i) = p$, express system reliability in terms of p, and then let $x = p^2$.]

85. Consider independently rolling two fair dice, one red and the other green. Let A be the event that the red die shows 3 dots, B be the event that the green die shows 4 dots, and C be the event that the total number of dots showing on the two dice is 7.

   a. Are these events pairwise independent (i.e., are A and B independent events, are A and C independent, and are B and C independent)?
   b. Are the three events mutually independent?

86. Components arriving at a distributor are checked for defects by two different inspectors (each component is checked by both inspectors). The first inspector detects 90% of all defectives that are present, and the second inspector does likewise. At least

one inspector fails to detect a defect on 20% of all defective components. What is the probability that the following occur?

a. A defective component will be detected only by the first inspector? By exactly one of the two inspectors?

b. All three defective components in a batch escape detection by both inspectors (assuming inspections of different components are independent of one another)?

87. A quality control inspector is inspecting newly produced items for faults. The inspector searches an item for faults in a series of independent "fixations," each of a fixed duration. Given that a flaw is actually present, let p denote the probability that the flaw is detected during any one fixation (this model is discussed in "Human Performance in Sampling Inspection," *Hum. Factors* 1979: 99–105).

a. Assuming that an item has a flaw, what is the probability that it is detected by the end of the second fixation (once a flaw has been detected, the sequence of fixations terminates)?

b. Give an expression for the probability that a flaw will be detected by the end of the nth fixation.

c. If when a flaw has not been detected in three fixations, the item is passed, what is the probability that a flawed item will pass inspection?

d. Suppose 10% of all items contain a flaw [P(randomly chosen item is flawed) = .1]. With the assumption of part (c), what is the probability that a randomly chosen item will pass inspection (it will automatically pass if it is not flawed, but could also pass if it is flawed)?

e. Given that an item has passed inspection (no flaws in three fixations), what is the probability that it is actually flawed? Calculate for $p = .5$.

88. a. A lumber company has just taken delivery on a lot of 10,000 2×4 boards. Suppose that 20% of these boards (2000) are actually too green to be used in first-quality construction. Two boards are selected at random, one after the other. Let A = {the first board is green} and B = {the second board is green}. Compute $P(A)$, $P(B)$, and $P(A \cap B)$ (a tree diagram might help). Are A and B independent?

b. With A and B independent and $P(A) = P(B) = .2$, what is $P(A \cap B)$? How much difference is there between this answer and $P(A \cap B)$ in part (a)? For purposes of calculating $P(A \cap B)$, can we assume that A and B of part (a) are independent to obtain essentially the correct probability?

c. Suppose the lot consists of ten boards, of which two are green. Does the assumption of independence now yield approximately the correct answer for $P(A \cap B)$? What is the critical difference between the situation here and that of part (a)? When do you think that an independence assumption would be valid in obtaining an approximately correct answer to $P(A \cap B)$?

89. Refer to the assumptions stated in Exercise 83, and answer the question posed there for the system in the accompanying picture. How would the probability change if this were a subsystem connected in parallel to the subsystem pictured in Figure 2.15a?
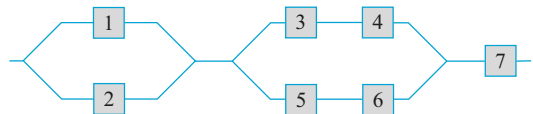
90. Professor Stander Deviation can take one of two routes on his way home from work. On the first route, there are four railroad

crossings. The probability that he will be stopped by a train at any particular one of the crossings is .1, and trains operate independently at the four crossings. The other route is longer but there are only two crossings, independent of each other, with the same stoppage probability for each as on the first route. On a particular day, Professor Deviation has a meeting scheduled at home for a certain time. Whichever route he takes, he calculates that he will be late if he is stopped by trains at least half the crossings encountered.

a. Which route should he take to minimize the probability of being late to the meeting?
b. If he tosses a fair coin to decide on a route and he is late, what is the probability that he took the four-crossing route?

91. Suppose identical tags are placed on both the left ear and the right ear of a fox. The fox is then let loose for a period of time. Consider the two events C_1 = {left ear tag is lost} and C_2 = {right ear tag is lost}. Let $p = P(C_1) = P(C_2)$, and assume C_1 and C_2 are independent events. Derive an expression (involving p) for the probability that exactly one tag is lost given that at most

one is lost ("Ear Tag Loss in Red Foxes," *J. Wildlife Manag.* 1976: 164–167).

92. It's a commonly held misconception that if you play the lottery n times, and the probability of winning each time is $1/N$, then your chance of winning at least once is n/N. That's true if you buy n tickets in one week, but not if you buy a single ticket in each of n independent weeks. Let's explore further.

a. Suppose you play a game n independent times, with $P(\text{win}) = 1/N$ each time. Find an expression for the probability you win at least once. [*Hint:* Consider the complement.]
b. How does your answer to (a) compare to n/N for the easy task of rolling a 4 on a fair die (so $N = 6$) in $n = 3$ tries? In $n = 6$ tries? In $n = 10$ tries?
c. Now consider a weekly lottery where you must guess the 6 winning numbers from 1 to 49, so $N = \binom{49}{6}$. If you play this lottery every week for a year ($n = 52$), how does your answer to (a) compare to n/N?
d. Show that when n is much smaller than N, the fraction n/N is not a bad approximation to (a). [*Hint:* Use the binomial theorem from high school algebra.]

2.6   Simulation of Random Events

As probability models in engineering and the sciences have grown in complexity, many problems have arisen that are too difficult to attack "analytically," i.e., using just mathematical tools such as those in the previous sections. Instead, computer simulation provides us an effective way to estimate probabilities of very complicated events (and, in later chapters, of other properties of random phenomena). In this section, we introduce the principles of probability simulation, demonstrate a few examples with R code, and discuss the precision of simulated probabilities.

Suppose an investigator wishes to determine $P(A)$, but either the experiment on which A is defined or the A event itself is so complicated as to preclude the use of probability rules and properties. The general method for *estimating* this probability via computer simulation is as follows:

- Write a program that simulates (mimics) the underlying random experiment.
- Run the program many times, with each run independent of all others.
- During each run, record whether or not the event A of interest occurs.

If the simulation is run a total of n independent times, then the estimate of $P(A)$, denoted by $\hat{P}(A)$, is

$$\hat{P}(A) = \frac{\text{number of times } A \text{ occurs}}{\text{number of runs}} = \frac{n(A)}{n}$$

For example, if we run a simulation program 10,000 times and the event of interest A occurs in 6174 of those runs, then our estimate of $P(A)$ is $\hat{P}(A) = 6174/10,000 = .6174$. Notice that our definition is consistent with the long-run relative frequency interpretation of probability discussed in Section 2.2.

The Backbone of Simulation: Random Number Generators

All modern software packages are equipped with a function called a **random number generator (RNG)**. A typical call to this function (such as ran or rand) will return a single, supposedly "random" number, though such functions typically permit the user to request a vector or even a matrix of "random" numbers. It is more proper to call these results pseudorandom numbers, since there is actually a deterministic (i.e., nonrandom) algorithm by which the software generates these values. We will not discuss the details of such algorithms here; see the book by Law listed in the references. What will matter to us are the following two characteristics:

1. Each number created by an RNG is as likely to be any particular number in the interval [0, 1) as it is to be any other number in this interval (up to computer precision, anyway).[2]
2. Successive values created by RNGs are independent, in the sense that we cannot predict the next value to be generated from the current value (unless we somehow know the exact parameters of the underlying algorithm).

A typical simulation program manipulates numbers on the interval [0, 1) in a way that mimics the experiment of interest; several examples are provided below. Arguably the most important building block for such programs is the ability to simulate a basic event that occurs with a *known* probability, p. Since RNGs produce values equally likely to be anywhere in the interval [0, 1), it follows that in the long run a proportion p of them will lie in the interval [0, p). So, suppose we need to simulate an event B with $P(B) = p$. In each run of our simulation program, we can call for a single "random" number, which we'll call u, and apply the following rules:

- If $0 \le u < p$, then event B has occurred on this run of the program.
- If $p \le u < 1$, then event B has *not* occurred on this run of the program.

Example 2.38 Let's begin with an example in which the exact probability can be obtained analytically, so that we may verify that our simulation method works. Suppose we have two independent devices which function with probabilities .6 and .7, respectively. What is the probability both devices function? That at least one device functions?

---

[2]In the language of Chapter 4, the numbers produced by an RNG follow essentially a *uniform* distribution on the interval [0, 1).

Let B_1 and B_2 denote the events that the first and second devices function, respectively; we know that $P(B_1) = .6$, $P(B_2) = .7$, and B_1 and B_2 are independent. Our first goal is to estimate the probability of $A = B_1 \cap B_2$, the event that both devices function. The following "pseudo-code" will allow us to obtain $\hat{P}(A)$.

0. Set a counter for the number of times A occurs to zero.
    Repeat n times:

1. Generate two random numbers, u_1 and u_2. (These will help us determine whether B_1 and B_2 occur, respectively.)
2. If $u_1 < .6$ AND $u_2 < .7$, then A has occurred. Add 1 to the count of occurrences of A.

Once the n runs are complete, then $\hat{P}(A) = $ (count of the occurrences of A)$/n$.

Figure 2.16 shows actual implementation code in R. We ran the program with $n = 10{,}000$ (as in the code) twice; the event A occurred 4218 times in the first run and 4157 the second time, providing estimated probabilities of $\hat{P}(A) = .4218$ and $.4157$, respectively. Compare these to the exact probability of A: by independence, $P(A) = P(B_1)P(B_2) = (.6)(.7) = .42$. Our simulation estimates were both "in the ballpark" of the right answer. We'll discuss the precision of these estimates shortly.

```
A=0
for(i in 1:10000){
        u1=runif(1); u2=runif(1)
        if(u1<.6 && u2<.7){
                A=A+1
        }
}
```

Figure 2.16   R code for Example 2.38

By replacing the "and" operator && in the above code with "or" operator ‖, we can estimate the probability at least one device functions, $P(B_1 \cup B_2)$. In one simulation (again with $n = 10{,}000$), the event $B_1 \cup B_2$ occurred 8802 times, giving the estimate $\hat{P}(B_1 \cup B_2) = .8802$. This is quite close to the exact probability:

$$P(B_1 \cup B_2) = 1 - P(B_1' \cap B_2') = 1 - (1 - .6)(1 - .7) = .88 \qquad \blacksquare$$

Example 2.39   Consider the following game: You'll flip a coin 25 times, winning \$1 each time it lands heads (H) and losing \$1 each time it lands tails (T). Unfortunately for you, the coin is biased in such a way that $P(H) = .4$ and $P(T) = .6$. What's the probability you come out ahead; i.e., you have more money at the end of the game than you had at the beginning? We'll use simulation to find out.

Now each run of the simulation requires 25 "random" objects: the results of the 25 coin tosses. What's more, we need to keep track of how much money you have won or lost at the end of the 25 tosses. Let $A = \{$you come out ahead$\}$, and use the following pseudo-code:

0. Set a counter for the number of times A occurs to zero.
    Repeat n times:

1. Set your initial dollar amount to zero.
2. Generate 25 random numbers $u_1, \ldots, u_{25}$.

3. For each $u_i < .4$, heads was tossed, so add 1 to your dollar amount. For each $u_i \geq .4$, the flip was tails and 1 is deducted.
4. If the final dollar amount is positive (i.e., \$1 or greater), add 1 to the count of occurrences for A.

Once the n runs are complete, then $\hat{P}(A) = $ (count of the occurrences of A)$/n$.

R code for Example 2.39 appears in Figure 2.17. Our code gave a final count of 1567 occurrences of A, out of 10,000 runs. Thus, the estimated probability that you come out ahead in this game is $\hat{P}(A) = 1567/10,000 = .1567$.

```
A=0
for (i in 1:10000){
      dollar=0
      for (j in 1:25){
            u=runif(1)
            if (u<.4){
                  dollar=dollar+1
            }
            else{dollar=dollar-1}
      }
      if (dollar>0){
            A=A+1
      }
}
```

Figure 2.17   R code for Example 2.39   ∎

Readers familiar with basic programming will recognize that many "for" loops like those in the preceding examples can be sped up by *vectorization*, i.e., by using a function call that generates all the required random numbers simultaneously, rather than one at a time. Similarly, the if/else statements used in the preceding programs to determine whether a random number lies in an interval can be rewritten in terms of true/false bits, which automatically generate a 1 if a statement is true and a 0 otherwise. For example, the R code

```
if (u < .5){
  A = A+1
}
```

can be replaced by the single line of code

```
A = A+(u < .5);
```

If the statement in parentheses is true, R assigns a value 1 to `(u<.5)`, and so 1 is added to the count `A`.

The previous two examples have both assumed independence of certain events: the functionality of neighboring devices, or the outcomes of successive coin flips. With the aid of some built-in functions within R, we can also simulate counting experiments similar to those in Section 2.3, even though draws without replacement from a finite population are not independent. To illustrate, let's use simulation to estimate some of the combinatorial probabilities from Section 2.3.

Example 2.40   Consider again the situation presented in Example 2.24: A university warehouse has received a shipment of 25 laptops, of which 10 have AMD processors and 15 have Intel chips; a particular technician will check 6 of these 25 laptops, selected at random. Of interest is the probability of the event D_3 = {exactly 3 of the 6 selected have Intel chips}. Although the initial probability of selecting a laptop with an Intel processor is 15/25, successive selections are not independent (the conditional probability that the next laptop also has an Intel chip is *not* 15/25). So, the method of the preceding examples does not apply.

Instead, we use the sampling tool built into our software, as follows:

0. Set a counter for the number of times D_3 occurs to zero.
    Repeat *n* times:

1. Sample 6 numbers, *without replacement*, from the integers 1 through 25. (1–15 correspond to the labels for the 15 laptops with Intel processors, and 16–25 identify the 10 with AMD processors.)
2. Count how many of these 6 numbers fall between 1 and 15, inclusive.
3. If exactly 3 of these 6 numbers fall between 1 and 15, add 1 to the count of occurrences for D_3.
    Once the *n* runs are complete, then $\hat{P}(D_3) = $ (count of the occurrences of D_3)$/n$.

R code for this example appears in Figure 2.18. Vital to the execution of this simulation is the fact that R (like many statistical software packages) has a built-in mechanism for randomly sampling without replacement from a finite set of objects (here, the integers 1–25). For more information on this function, type `help(sample)` in R.

In the code, the line `sum(chips<=15)` performs two actions. First, `chips<=15` converts each

```
D=0
for (i in 1:10000){
    chips=sample(25,6)
    intel=sum(chips<=15)
    if (intel==3){
        D=D+1
    }
}
```

Figure 2.18   R code for Example 2.40

of the 6 numbers in the vector `chips` into a 1 if the entry is between 1 and 15 (and into a 0 otherwise). Second, `sum()` adds up the 1's and 0's, which is equivalent to identifying how many 1's appear (i.e., how many of the 6 numbers fell between 1 and 15).

Our code resulted in event D_3 occurring 3054 times, so $\hat{P}(D_3) = 3054/10{,}000 = .3054$, which is quite close to the "exact" answer of .3083 found in Example 2.24. The other probability of interest, the chance of randomly selecting *at least* 3 laptops with Intel processors, can be estimated by modifying one line of code: change `intel==3` to `intel>=3`. One simulation provided a count of 8522 occurrences in 10,000 trials, for an estimated probability of .8522 (close to the combinatorial solution of .8530). ■

Precision of Simulation

In Example 2.38, we gave two different estimates $\hat{P}(A)$ for a probability $P(A)$. Which is more "correct"? Without knowing $P(A)$ itself, there's no way to tell. However, thanks to the theory we will develop in subsequent chapters, we can quantify the precision of simulated probabilities. Of course, we must have written code that faithfully simulates the random experiment of interest. Further, we

assume that the results of each single run of our program are independent of the results of all other runs. (This generally follows from the aforementioned independence of computer-generated random numbers.)

If this is the case, then a measure of the disparity between the true probability $P(A)$ and the estimated probability $\hat{P}(A)$ based on n runs of the simulation is given by

$$\sqrt{\frac{\hat{P}(A)[1 - \hat{P}(A)]}{n}} \qquad (2.9)$$

This measure of precision is called the **(estimated) standard error** of the estimate $\hat{P}(A)$; see Section 3.5 for a derivation. The standard error is analogous to the standard deviation from Chapter 1. Expression (2.9) tells us that the amount by which $\hat{P}(A)$ typically differs from $P(A)$ depends upon two values: $\hat{P}(A)$ itself and the number of runs n. You can make sense of the former this way: if $P(A)$ is very small, then $\hat{P}(A)$ will presumably be small as well, in which case they cannot deviate by very much since both are bounded below by zero. (Standard error quantifies the *absolute* difference between them, not the relative difference.) A similar comment applies if $P(A)$ is very large, i.e., near 1.

As for the relationship to n, Expression (2.9) indicates that the amount by which $\hat{P}(A)$ will typically differ from $P(A)$ is inversely proportional to the square root of n. So, in particular, as n increases the tendency is for $\hat{P}(A)$ to vary less and less. This speaks to the precision of $\hat{P}(A)$: our estimate becomes more precise as n increases, but not at a very fast rate.

Let's think a bit more about this relationship: suppose your simulation results thus far were too imprecise for your taste. By how much would you have to increase the number of runs to gain one additional decimal place of precision? That's equivalent to reducing the estimated standard error by a factor of 10. Since precision is proportional to $1/\sqrt{n}$, you would need to increase n by a factor of 100 to achieve the desired improvement; e.g., if using $n = 10,000$ runs is insufficient for your purposes, then you'll need 1,000,000 runs to get one additional decimal place of precision. Typically, this will mean running your program 100 times longer—not a big deal if 10,000 runs only take a nanosecond but prohibitive if they require, say, an hour.

Example 2.41 (Example 2.39 continued) Based on $n = 10,000$ runs, we estimated the probability of coming out ahead in a certain game to be $\hat{P}(A) = .1567$. Substituting into (2.9), we get

$$\sqrt{\frac{.1567[1 - .1567]}{10,000}} = .0036$$

This is the (estimated) standard error of our estimate .1567. We interpret as follows: some simulation experiments with $n = 10,000$ will result in an estimated probability that is within .0036 of the actual probability, whereas other such experiments will give an estimated probability that deviates by more than .0036 from the actual $P(A)$; .0036 is roughly the size of a typical deviation between the estimate and what it is estimating. ∎

In Chapter 8, we will return to the notion of standard error and develop a so-called *confidence interval* estimate for $P(A)$: a range of numbers we can be very certain contains $P(A)$.

Exercises Section 2.6 (93–112)

93. Refer to Example 2.38.

    a. Modify the code in Figure 2.16 to estimate the probability that *exactly* one of the two devices functions properly. Then find the exact probability using the techniques from earlier sections of this chapter, and compare it to your estimated probability.

    b. Calculate the estimated standard error for the estimated probability in (a).

94. Imagine you have five independently operating components, each working properly with probability .8. Use simulation to estimate the probability that

    a. All five components work properly.

    b. At least one of the five components works properly.
    [Hints for (a) and (b): You can adapt the code from Example 2.38, but the and/or statements will become tedious. Consider using the max and min functions instead.]

    c. Calculate the estimated standard errors for your answers in (a) and (b).

95. Consider the system depicted in Exercise 89. Assume the seven components operate independently with the following probabilities of functioning properly: .9 for components 1 and 2; .8 for each of components 3, 4, 5, 6; and .95 for component 7. Write a program to estimate the reliability of the system (i.e., the probability the system functions properly).

96. You have an opportunity to answer six trivia questions about your favorite sports team, and you will win a pair of tickets to their next game if you can correctly answer at least three of the questions. Write a simulation program to estimate the chance you win the tickets under each of the following assumptions.

    a. You have a 50–50 chance of getting any question right, independent of all others.

    b. Being a true fan, you have a 75% chance of getting any question right, independent of all others.

    c. The first three questions are fairly easy, so you have a .75 chance of getting each of those right. However, the last three questions are much harder, and you only have a .3 probability of correctly answer each of those.

97. In the game "Now or Then" on the television show *The Price is Right*, the contestant faces a wheel with 6 sectors. Each sector contains a grocery item and a price, and the contestant must decide whether the price is "now" (i.e., the item's price the day of the taping) or "then" (the price at some specified past date, such as September 2003). The contestant wins a prize (bedroom furniture, a Caribbean cruise, etc.) if he/she guesses correctly on three *adjacent* sectors. That is, numbering the sectors 1–6 clockwise, correct guesses on sectors 5, 6, and 1 wins the prize but not on sectors 5, 6, and 3, since the latter are not all adjacent. (The contestant gets to guess on all six sectors, if need be.)
    Write a simulation program to estimate the probability the contestant wins the prize, assuming her/his guesses are independent from item to item. Provide estimated probabilities under of the following assumptions: (1) each guess is "wild" and thus has probability .5 of being correct, and (2) the contestant is a good shopper, with probability .8 of being correct on any item.

98. Refer to the game in Example 2.39. Under the same conditions as in that example, estimate the probability the player is ahead *at any time* during the 25 plays. [*Hint*: This occurs if the player's dollar amount is positive at any of the 25 steps in the loop. So, you will need to keep track of every value of the dollar variable, not just the final result.]

99. Refer again to Example 2.39. Estimate the probability that the player experiences a "swing" of at least $5 during the game. That is, estimate the chance that the difference between the largest and smallest dollar amounts during the game is at least 5. (This would happen, for instance, if the player was at one point ahead at +$2 but later fell behind to –$3.)

100. Teresa and Peter each have a fair coin. Teresa tosses her coin repeatedly until obtaining the sequence HTT. Peter tosses his coin until the sequence HTH is obtained.

   a. Write a program to simulate Teresa's coin tossing and, separately, Peter's. Your program should keep track of the number of tosses each author requires on each simulation run to achieve his target sequence.

   b. Estimate the probability that Peter obtains his sequence with fewer tosses than Teresa requires to obtain her sequence.

101. A 40-question multiple-choice exam is sometimes administered in lower-level statistics courses. The exam has a peculiar feature: 10 of the questions have two options, 13 have three options, 13 have four options, and the other 4 have five options. (FYI, this is completely real!) What is the probability that, purely by guessing, a student could get at least half of these questions correct? Write a simulation program to answer this question.

102. Major League Baseball teams (usually) play a 162-game season, during which fans are often excited by long winning streaks and frustrated by long losing streaks. But how unusual are these streaks, really? How long a streak would you expect if the team's performance were independent from game to game?
   Write a program that simulates a 162-game season, i.e., a string of 162 wins and losses, with $P(\text{win}) = p$ for each game (the value of p to be specified later). Use your program with at least 10,000 runs to answer the following questions.

   a. Suppose you're rooting for a ".500" team—that is, $p = .5$. What is the probability of observing a streak of at least five wins in a 162-game season? Estimate this probability with your program, and include a standard error.

   b. Suppose instead your team is quite good: a .600 team overall, so $p = .6$. Intuitively, should the probability of a winning streak of at least five games be higher or lower? Explain.

   c. Use your program with $p = .6$ to estimate the probability alluded to in (b). Is your answer higher or lower than (a)? Is that what you anticipated?

103. A *derangement* of the numbers 1 through n is a permutation of all n those numbers such that none of them is in the "right place." For example, 34251 is a derangement of 1 through 5, but 24351 is not because 3 is in the 3rd position. We will use simulation to estimate the number of derangements of the numbers 1 through 12.

   a. Write a program that generates random permutations of the integers 1, 2, …, 12. Your program should determine whether or not each permutation is a derangement.

   b. Based on your program, estimate $P(D)$, where $D = \{$a permutation of 1–12 is a derangement$\}$.

   c. From Section 2.3, we know the number of permutations of n items. (How many is that for $n = 12$?) Use this information and your answer to part (b) to estimate the *number* of derangements of the numbers 1 through 12.
   [Hint for (a): Use random sampling without replacement as in Example 2.40.]

104. The famous *Birthday Problem* was presented in Example 2.22. Now suppose you have 500 Facebook friends. Make the same assumptions here as in the Birthday Problem.

a. Write a program to estimate the probability that, on at least one day during the year, Facebook tells you three (or more) of your friends share that birthday. Based on your answer, should you be surprised by this occurrence?

b. Write a program to estimate the probability that, on at least one day during the year, Facebook tells you *five* (or more) of your friends share that birthday. Based on your answer, should you be surprised by this occurrence?

[*Hint*: Generate 500 birthdays *with* replacement, and then determine whether any birthday occurs three or more times (five or more for part (b)). The table function in R may prove useful.]

105. Consider the following game: you begin with $20. You flip a fair coin, winning $10 if the coin lands heads and losing $10 if the coin lands tails. Play continues until you either go broke or have $100 (i.e., a net profit of $80). Write a simulation program to estimate:

a. The probability you win the game.

b. The probability the game ends within ten coin flips.

[*Note:* This is a special case of the *Gambler's Ruin* problem.]

106. Consider the *Coupon Collector's Problem*: 10 different coupons are distributed into cereal boxes, one per box, so that any randomly selected box is equally likely to have any of the 10 coupons inside. Write a program to simulate the process of buying cereal boxes until all 10 distinct coupons have been collected. For each run, keep track of how many cereal boxes you purchased to collect the complete set of coupons. Then use your program to answer the following questions.

a. What is the probability you collect all 10 coupons with just 10 cereal boxes?

b. Use counting techniques to determine the exact probability in (a). [*Hint*: Relate this to the Birthday Problem.]

c. What is the probability you require more than 20 boxes to collect all 10 coupons?

d. Using techniques from Chapters 3 and 5, it can be shown that it takes about 29.3 boxes, on the average, to collect all 10 coupons. What's the probability of collecting all 10 coupons in fewer than average boxes (i.e., less than 29.3)?

107. Consider the following famous puzzle from the early days of probability, investigated by Pascal and Fermat. Which of the following events is more likely: to roll at least one 6 in four rolls of a fair die, or to roll at least one double-6 in 24 rolls of two fair dice?

a. Write a program to simulate a set of four die rolls many times, and use the results to estimate P(at least one 6 in 4 rolls).

b. Now adapt your program to simulate rolling a pair of dice 24 times. Repeat this simulation many times, and use your results to estimate P(at least one double-6 in 24 rolls).

108. *The Problem of the Points.* Pascal and Fermat also explored a question concerning how to divide the stakes in a game that has been interrupted. Suppose two players, Blaise and Pierre, are playing a game where the winner is the first to achieve a certain number of points. The game gets interrupted at a moment when Blaise needs n more points to win and Pierre needs m more to win. How should the game's prize money be divvied up? Fermat argued that Blaise should receive a proportion of the total stake equal to the chance he would have won if the game hadn't been interrupted (and Pierre receives the remainder).

Assume the game is played in rounds, the winner of each round gets 1 point, rounds are independent, and the two players are equally likely to win any particular round.

a. Write a program to simulate the rounds of the game that would have happened after play was interrupted. A single simulation run should terminate as soon as Blaise has n wins or Pierre has m wins (equivalently, Blaise has m losses). Use your program to estimate P(Blaise gets 10 wins before 15 losses), which is the proportion of the total stake Blaise should receive if $n = 10$ and $m = 15$.

b. Use your same program to estimate the relevant probability when $n = m = 10$. Logically, what should the answer be? Is your estimated probability close to that?

c. Finally, let's assume Pierre is actually the better player: P(Blaise wins a round) $= .4$. Again with $n = 10$ and $m = 15$, what proportion of the stake should be awarded to Blaise?

109. Twenty faculty members in a certain department have just participated in a department chair election. Suppose that candidate A has received 12 of the votes and candidate B the other 8 votes. If the ballots are opened one by one in random order and the candidate selected on each ballot is recorded, use simulation to estimate the probability that candidate A remains ahead of candidate B throughout the vote count (which happens if, for example, the result is AA…AB…B but not if the result is AABABB…).

110. Show that the (estimated) standard error for $\hat{P}(A)$ is at most $1/\sqrt{4n}$.

111. Simulation can be used to estimate numerical constants, such as π. Here's one approach: consider the part of a disk of radius 1 that lies in the first quadrant (a quarter-circle). Imagine two random numbers, x and y, both between 0 and 1. The pair (x, y) lies somewhere in the first quadrant; let A denote the event that (x, y) falls inside the quarter-circle.

a. Write a program that simulates pairs (x, y) in order to estimate $P(A)$, the probability that a randomly selected pair of points in the square $[0, 1] \times [0, 1]$ lies in the quarter-circle of radius 1.

b. Using techniques from Chapter 5, it can be shown that the exact probability of A is $\pi/4$ (which makes sense, because that's the ratio of the quarter-circle's area to the square's area). Use that fact to come up with an estimate of π from your simulation. How close is your estimate to 3.14159…?

112. Consider the quadratic equation $ax^2 + bx + c = 0$. Suppose that $a, b,$ and c are random numbers between 0 and 1 (like those produced by an RNG). Estimate the probability that the roots of this quadratic equation are real. [*Hint:* Think about the discriminant.] This probability can be computed exactly using methods from Chapter 5, but a triple integral is required.

Supplementary Exercises: (113–140)

113. The undergraduate statistics club at a certain university has 24 members.

a. All 24 members of the club are eligible to attend a conference next week, but they can only afford to send 4 people. In how many possible ways could 4 attendees be selected?

b. All club members are eligible for any of the four positions of president, VP, secretary, or treasurer. In how many possible ways can these positions be occupied?

c. Suppose it's agreed that two people will be cochairs, one person secretary, and one person treasurer. How many ways are there to fill these positions now?

114. A small manufacturing company will start operating a night shift. There are 20 machinists employed by the company.

a. If a night crew consists of 3 machinists, how many different crews are possible?

b. If the machinists are ranked 1, 2, …, 20 in order of competence, how many of these crews would not have the best machinist?

c. How many of the crews would have at least 1 of the 10 best machinists?

d. If a 3-person crew is selected at random to work on a particular night, what is the probability that the best machinist will not work that night?

115. A factory uses three production lines to manufacture cans of a certain type. The accompanying table gives percentages of nonconforming cans, categorized by type of nonconformance, for each of the three lines during a particular time period.

	Line 1	Line 2	Line 3
Blemish	15	12	20
Crack	50	44	40
Pull Tab Problem	21	28	24
Surface Defect	10	8	15
Other	4	8	2

During this period, line 1 produced 500 nonconforming cans, line 2 produced 400 such cans, and line 3 was responsible for 600 nonconforming cans. Suppose that one of these 1500 cans is randomly selected.

a. What is the probability that the can was produced by line 1? That the reason for nonconformance is a crack?

b. If the selected can come from line 1, what is the probability that it had a blemish?

c. Given that the selected can had a surface defect, what is the probability that it came from line 1?

116. An employee of the records office at a university currently has ten forms on his desk awaiting processing. Six of these are withdrawal petitions, and the other four are course substitution requests.

a. If he randomly selects six of these forms to give to a subordinate, what is the probability that only one of the two types of forms remains on his desk?

b. Suppose he has time to process only four of these forms before leaving for the day. If these four are randomly selected one by one, what is the probability that each succeeding form is of a different type from its predecessor?

117. One satellite is scheduled to be launched from Cape Canaveral in Florida, and another launching is scheduled for Vandenberg Air Force Base in California. Let A denote the event that the Vandenberg launch goes off on schedule, and let B represent the event that the Cape Canaveral launch goes off on schedule. If A and B are independent events with $P(A) > P(B)$ and $P(A \cup B) = .626$, $P(A \cap B) = .144$, determine the values of $P(A)$ and $P(B)$.

118. A transmitter is sending a message by using a binary code, namely a sequence of 0's and 1's. Each transmitted bit (0 or 1) must pass through three relays to reach the receiver. At each relay, the probability is .2 that the bit sent will be different from the bit received (a reversal). Assume that the relays operate independently of one another.

Transmitter → Relay 1 → Relay 2
→ Relay 3 → Receiver

a. If a 1 is sent from the transmitter, what is the probability that a 1 is sent by all three relays?

b. If a 1 is sent from the transmitter, what is the probability that a 1 is received by the receiver? [*Hint*: The eight experimental outcomes can be displayed on a tree diagram with three generations of branches, one generation for each relay.]

c. Suppose 70% of all bits sent from the transmitter are 1's. If a 1 is received by the receiver, what is the probability that a 1 was sent?

119. Individual A has a circle of five close friends (B, C, D, E, and F). A has heard a certain rumor from outside the circle and has invited the five friends to a party to circulate the rumor. To begin, A selects one of the five at random and tells the rumor to the chosen individual. That individual then selects at random one of the four remaining individuals and repeats the rumor. Continuing, a new individual is selected from those not already having heard the rumor by the individual who has just heard it, until everyone has been told.

   a. What is the probability that the rumor is repeated in the order B, C, D, E, and F?
   b. What is the probability that F is the third person at the party to be told the rumor?
   c. What is the probability that F is the last person to hear the rumor?

120. Refer to the previous exercise. If at each stage the person who currently "has" the rumor does not know who has already heard it and selects the next recipient at random from all five possible individuals, what is the probability that F has still not heard the rumor after it has been told ten times at the party?

121. A chemist is interested in determining whether a certain trace impurity is present in a product. An experiment has a probability of .80 of detecting the impurity if it is present. The probability of not detecting the impurity if it is absent is .90. The prior probabilities of the impurity being present and being absent are .40 and .60, respectively. Three separate experiments result in only two detections. What is the posterior probability that the impurity is present?

122. Fasteners used in aircraft manufacturing are slightly crimped so that they lock enough to avoid loosening during vibration. Suppose that 95% of all fasteners pass an initial inspection. Of the 5% that fail, 20% are so seriously defective that they must be scrapped. The remaining fasteners are sent to a re-crimping operation, where 40% cannot be salvaged and are discarded. The other 60% of these fasteners are corrected by the re-crimping process and subsequently pass inspection.

   a. What is the probability that a randomly selected incoming fastener will pass inspection either initially or after re-crimping?
   b. Given that a fastener passed inspection, what is the probability that it passed the initial inspection and did not need re-crimping?

123. One percent of all individuals in a certain population are carriers of a particular disease. A diagnostic test for this disease has a 90% detection rate for carriers and a 5% detection rate for noncarriers. Suppose the test is applied independently to two different blood samples from the same randomly selected individual.

   a. What is the probability that both tests yield the same result?
   b. If both tests are positive, what is the probability that the selected individual is a carrier?

124. A system consists of two components. The probability that the second component functions in a satisfactory manner during its design life is .9, the probability that at least one of the two components does so is .96, and the probability that both components do so is .75. Given that the first component functions in a satisfactory manner throughout its design life, what is the probability that the second one does also?

125. A certain company sends 40% of its overnight mail parcels via express mail service E_1. Of these parcels, 2% arrive after the guaranteed delivery time (denote the event "late delivery" by L). If a record of an overnight mailing is randomly selected from the company's file, what is the

probability that the parcel went via E_1 and was late?

126. Refer to the previous exercise. Suppose that 50% of the overnight parcels are sent via express mail service E_2 and the remaining 10% are sent via E_3. Of those sent via E_2, only 1% arrive late, whereas 5% of the parcels handled by E_3 arrive late.

    a. What is the probability that a randomly selected parcel arrived late?
    b. If a randomly selected parcel has arrived on time, what is the probability that it was not sent via E_1?

127. A company uses three different assembly lines—A_1, A_2, and A_3—to manufacture a particular component. Of those manufactured by line A_1, 5% need rework to remedy a defect, whereas 8% of A_2's components need rework and 10% of A_3's need rework. Suppose that 50% of all components are produced by line A_1, 30% are produced by line A_2, and 20% come from line A_3. If a randomly selected component needs rework, what is the probability that it came from line A_1? From line A_2? From line A_3?

128. Disregarding the possibility of a February 29 birthday, suppose a randomly selected individual is equally likely to have been born on any one of the other 365 days. If ten people are randomly selected, what is the probability that either at least two have the same birthday or at least two have the same last three digits of their Social Security numbers? [*Note*: The article "Methods for Studying Coincidences" (F. Mosteller and P. Diaconis, *J. Amer. Statist. Assoc.* 1989: 853–861) discusses problems of this type.]

129. One method used to distinguish between granitic (G) and basaltic (B) rocks is to examine a portion of the infrared spectrum of the sun's energy reflected from the rock surface. Let R_1, R_2, and R_3 denote measured spectrum intensities at three different wavelengths; typically, for granite $R_1 < R_2 < R_3$, whereas for basalt $R_3 < R_1 < R_2$.

When measurements are made remotely (using aircraft), various orderings of the R_i's may arise whether the rock is basalt or granite. Flights over regions of known composition have yielded the following information:

	Granite	Basalt
$R_1 < R_2 < R_3$	60%	10%
$R_1 < R_3 < R_2$	25%	20%
$R_3 < R_1 < R_2$	15%	70%

Suppose that for a randomly selected rock specimen in a certain region, $P(\text{granite}) = .25$ and $P(\text{basalt}) = .75$.

a. Show that $P(\text{granite} \mid R_1 < R_2 < R_3) > P(\text{basalt} \mid R_1 < R_2 < R_3)$. If measurements yielded $R_1 < R_2 < R_3$, would you classify the rock as granite or basalt?
b. If measurements yielded $R_1 < R_3 < R_2$, how would you classify the rock? Answer the same question for $R_3 < R_1 < R_2$.
c. Using the classification rules indicated in parts (a) and (b), when selecting a rock from this region, what is the probability of an erroneous classification? [*Hint*: Either G could be classified as B or B as G, and $P(B)$ and $P(G)$ are known.]
d. If $P(\text{granite}) = p$ rather than .25, are there values of p (other than 1) for which a rock would always be classified as granite?

130. In a Little League baseball game, team A's pitcher throws a strike 50% of the time and a ball 50% of the time, successive pitches are independent of each other, and the pitcher never hits a batter. Knowing this, team B's manager has instructed the first batter not to swing at anything. Calculate the probability that

a. The batter walks on the fourth pitch.
b. The batter walks on the sixth pitch (so two of the first five must be strikes), using a counting argument or constructing a tree diagram.

c. The batter walks.

d. The first batter up scores while no one is out (assuming that each batter pursues a no-swing strategy).

131. *The Matching Problem.* Four friends— Allison, Beth, Carol, and Diane—who have identical calculators are studying for a statistics exam. They set their calculators down in a pile before taking a study break and then pick them up in random order when they return from the break.

a. What is the probability all four friends pick up the correct calculator?

b. What is the probability that at least one of the four gets her own calculator? [*Hint*: Let A be the event that Alice gets her own calculator, and define events B, C, and D analogously for the other three students. How can the event {at least one gets her own calculator} be expressed in terms of the four events A, B, C, and D? Now use a general law of probability.]

c. Generalize the answer from part (b) to n individuals. Can you recognize the result when n is large (the approximation to the resulting series)?

132. A particular airline has 10 a.m. flights from Chicago to New York, Atlanta, and Los Angeles. Let A denote the event that the New York flight is full and define events B and C analogously for the other two flights. Suppose $P(A) = .6$, $P(B) = .5$, $P(C) = .4$ and the three events are independent. What is the probability that

a. All three flights are full? That at least one flight is not full?

b. Only the New York flight is full? That exactly one of the three flights is full?

133. *The Secretary Problem.* A personnel manager is to interview four candidates for a job. These are ranked 1, 2, 3, and 4 in order of preference and will be interviewed in random order. However, at the conclusion of each interview, the manager will know only how the current candidate compares to those

previously interviewed. For example, the interview order 3, 4, 1, 2 generates no information after the first interview and shows that the second candidate is worse than the first, and that the third is better than the first two. However, the order 3, 4, 2, 1 would generate the same information after each of the first three interviews. The manager wants to hire the best candidate but must make an irrevocable hire/no hire decision after each interview. Consider the following strategy: Automatically reject the first s candidates, and then hire the first subsequent candidate who is best among those already interviewed (if no such candidate appears, the last one interviewed is hired).
For example, with $s = 2$, the order 3, 4, 1, 2 would result in the best being hired, whereas the order 3, 1, 2, 4 would not. Of the four possible s values (0, 1, 2, and 3), which one maximizes P(best is hired)? [*Hint*: Write out the 24 equally likely interview orderings; $s = 0$ means that the first candidate is automatically hired.]

134. Consider four independent events A_1, A_2, A_3, and A_4 and let $p_i = P(A_i)$ for $i = 1, 2, 3, 4$. Express the probability that at least one of these four events occurs in terms of the p_i's, and do the same for the probability that at least two of the events occur.

135. A box contains the following four slips of paper, each having exactly the same dimensions: (1) win prize 1; (2) win prize 2; (3) win prize 3; (4) win prizes 1, 2, and 3. One slip will be randomly selected. Let $A_1 = \{$win prize 1$\}$, $A_2 = \{$win prize 2$\}$, and $A_3 = \{$win prize 3$\}$. Show that A_1 and A_2 are independent, that A_1 and A_3 are independent, and that A_2 and A_3 are also independent (this is *pairwise* independence). However, show that $P(A_1 \cap A_2 \cap A_3) \neq P(A_1) \cdot P(A_2) \cdot P(A_3)$, so the three events are not *mutually* independent.

136. Consider a woman whose brother is afflicted with hemophilia, which implies that the woman's mother has the hemophilia gene

on one of her two X chromosomes (almost surely not both, since that is generally fatal). Thus there is a 50–50 chance that the woman's mother has passed on the bad gene to her. The woman has two sons, each of whom will independently inherit the gene from one of her two chromosomes. If the woman herself has a bad gene, there is a 50–50 chance she will pass this on to a son. Suppose that neither of her two sons is afflicted with hemophilia. What then is the probability that the woman is indeed the carrier of the hemophilia gene? What is this probability if she has a third son who is also not afflicted?

137. Jurors may be a priori biased for or against the prosecution in a criminal trial. Each juror is questioned by both the prosecution and the defense (the *voir dire* process), but this may not reveal bias. Even if bias is revealed, the judge may not excuse the juror for cause because of the narrow legal definition of bias. For a randomly selected candidate for the jury, define events B_0, B_1, and B_2 as the juror being unbiased, biased against the prosecution, and biased against the defense, respectively. Also let C be the event that bias is revealed during the questioning and D be the event that the juror is eliminated for cause. Let $b_i = P(B_i)$ ($i = 0, 1, 2$), $c = P(C \mid B_1) = P(C \mid B_2)$, and $d = P(D \mid B_1 \cap C) = P(D \mid B_2 \cap C)$ ["Fair Number of Peremptory Challenges in Jury Trials," *J. Amer. Statist. Assoc.* 1979: 747–753].

a. If a juror survives the voir dire process, what is the probability that he/she is unbiased (in terms of the b_i's, c, and d)? What is the probability that he/she is biased against the prosecution? What is the probability that he/she is biased against the defense? [*Hint*: Represent this situation using a tree diagram with three generations of branches.]

b. What are the probabilities requested in (a) if $b_0 = .50$, $b_1 = .10$, $b_2 = .40$ (all based on data relating to the famous trial

of the Florida murderer Ted Bundy), $c = .85$ (corresponding to the extensive questioning appropriate in a capital case), and $d = .7$ (a "moderate" judge)?

138. *Gambler's Ruin.* Allan and Beth currently have $2 and $3, respectively. A fair coin is tossed. If the result of the toss is H, Allan wins $1 from Beth, whereas if the coin toss results in T, then Beth wins $1 from Allan. This process is then repeated, with a coin toss followed by the exchange of $1, until one of the two players goes broke (one of the two gamblers is ruined). We wish to determine

$$a_2 = P(\text{Allan is the winner}|\text{he starts with }\$2)$$

To do so, let's also consider probabilities

$$a_i = P(\text{Allan wins}|\text{he starts with }\$i) \text{ for } i = 0, 1, 3, 4, \text{ and } 5.$$

a. What are the values of a_0 and a_5?

b. Use the Law of Total Probability to obtain an equation relating a_2 to a_1 and a_3. [*Hint*: Condition on the result of the first coin toss, realizing that if it is a H, then from that point Allan starts with $3.]

c. Using the logic described in (b), develop a system of equations relating a_i ($i = 1, 2, 3, 4$) to a_{i-1} and a_{i+1}. Then solve these equations. [*Hint*: Write each equation so that $a_i - a_{i-1}$ is on the left-hand side. Then use the result of the first equation to express each other $a_i - a_{i-1}$ as a function of a_1, and add together all four of these expressions ($i = 2, 3, 4, 5$).]

d. Generalize the result to the situation in which Allan's initial fortune is $a and Beth's is $b. [*Note*: The solution is a bit more complicated if $p = P(\text{Allan wins }\$1) \neq .5$.]

139. An event A is said to *attract* event B if $P(B \mid A) > P(B)$ and *repel* B if $P(B \mid A) < P(B)$. (This refines the notion of dependent

events by specifying whether A makes B more likely or less likely to occur.)

a. Show that if A attracts B, then A repels B'.
b. Show that if A attracts B, then A' repels B.
c. Prove the *Law of Mutual Attraction:* event A attracts event B if, and only if, B attracts A.

140. A fair coin is tossed repeatedly until either the sequence TTH or the sequence THT is observed. Let B be the event that stopping occurs because TTH was observed (i.e., that TTH is observed before THT). Calculate $P(B)$. [*Hint*: Consider the following partition of the sample space: $A_1 = \{$1st toss is H$\}$, $A_2 = \{$1st two tosses are TT$\}$, $A_3 = \{$1st three tosses are THT$\}$, and $A_4 = \{$1st three tosses are THH$\}$. Also denote $P(B)$ by p. Apply the Law of Total Probability, and p will appear on both sides in various places. The resulting equation is easily solved for p.]

Discrete Random Variables and Probability Distributions

3

Introduction

Suppose a city's traffic engineering department monitors a certain intersection during a one-hour period in the middle of the day. Many characteristics might be of interest: the number of vehicles that enter the intersection, the largest number of vehicles in the left turn lane during a signal cycle, the speed of the fastest vehicle going through the intersection, the average speed $\bar{x}$ of all vehicles entering the intersection. The value of each one of the foregoing variable quantities is subject to uncertainty—we don't know a priori how many vehicles will enter, what the maximum speed will be, etc. So each of these is referred to as a *random variable*—a variable quantity whose value is determined by what happens in a chance experiment.

The most commonly encountered random variables are one of two fundamentally different types: discrete random variables and continuous random variables. In this chapter, we examine the basic properties and discuss the most important examples of discrete variables. Chapter 4 focuses on continuous random variables.

3.1 Random Variables

In any experiment, numerous characteristics can be observed or measured, but in most cases an experimenter will focus on some specific aspect or aspects of a sample. For example, in a study of commuting patterns in a metropolitan area, each individual in a sample might be asked about commuting distance and the number of people commuting in the same vehicle, but not about IQ, income, family size, and other such characteristics. Alternatively, a researcher may test a sample of components and record only the number that have failed within 1000 h, rather than record the individual failure times.

In general, each outcome of an experiment can be associated with a number by specifying a rule of association, e.g., the number among the sample of ten components that fail to last 1000 h, or the total baggage weight for a sample of 25 airline passengers. Such a rule of association is called a **random variable**—a variable because different numerical values are possible and random because the observed value depends on which of the possible experimental outcomes results (Figure 3.1).
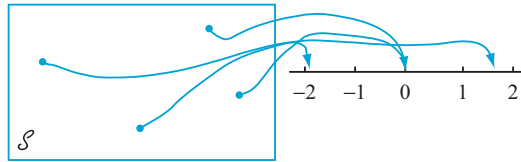
Figure 3.1 A random variable

DEFINITION For a given sample space $\mathscr{S}$ of some experiment, a **random variable (rv)** is any rule that associates a number with each outcome in $\mathscr{S}$. In mathematical language, a random variable is a function whose domain is the sample space and whose range is some subset of real numbers.

Random variables are customarily denoted by uppercase letters, such as X and Y, near the end of our alphabet. In contrast to our previous use of a lowercase letter, such as x, to denote a variable, we will now use lowercase letters to represent some particular value of the corresponding random variable. The notation $X(s) = x$ means that x is the value associated with the outcome s by the rv X.

Example 3.1 When a student attempts to connect to a university's WIFI network, either there is a failure (F) or there is a success (S). With $\mathscr{S} = \{S, F\}$, define a rv X by $X(S) = 1, X(F) = 0$. The rv X indicates whether (1) or not (0) the student can connect. ∎

In Example 3.1, the rv X was specified by explicitly listing each element of $\mathscr{S}$ and the associated number. If $\mathscr{S}$ contains more than a few outcomes, such a listing is tedious, but it can frequently be avoided.

Example 3.2 Consider the experiment in which a telephone number is dialed using a random number dialer (such devices are used extensively by polling organizations), and define a rv Y by

$$Y = \begin{cases} 1 & \text{if the selected number is on the National Do Not Call Registry} \\ 0 & \text{if the selected number is not on the registry} \end{cases}$$

For example, if 916-528-2966 appears on the national registry, then $Y(916\text{-}528\text{-}2966) = 1$, whereas $Y(213\text{-}772\text{-}7350) = 0$ tells us that the number 213-772-7350 is not on the registry. A word description of this sort is more economical than a complete listing, so we will use such a description whenever possible. ∎

In Examples 3.1 and 3.2, the only possible values of the random variable were 0 and 1. Such a random variable arises frequently enough to be given a special name, after the individual who first studied it.

DEFINITION Any random variable whose only possible values are 0 and 1 is called a **Bernoulli random variable**.

We will often want to define and study several different random variables from the same sample space.

Example 3.3 Example 2.3 described an experiment in which the number of pumps in use at each of two gas stations was determined. Define rvs X, Y, and U by

X = the total number of pumps in use at the two stations
Y = the difference between the number of pumps in use at station 1 and the number in use at station 2
U = the maximum of the numbers of pumps in use at the two stations

If this experiment is performed and $s = (2, 3)$ results, then $X((2, 3)) = 2 + 3 = 5$, so we say that the observed value of X is $x = 5$. Similarly, the observed value of Y would be $y = 2 - 3 = -1$, and the observed value of U would be $u = \max(2, 3) = 3$.   ■

Each of the random variables of Examples 3.1–3.3 can assume only a finite number of possible values. This need not be the case.

Example 3.4 Consider any general inspection process, wherein items are examined one by one until we find an item that falls within required specification limits. The sample space of such an experiment is $\mathcal{S} = \{S, FS, FFS, \ldots\}$. Define a rv X by

$$X = \text{the number of items examined until a ``good'' one is found}$$

Then $X(S) = 1, X(FS) = 2, X(FFS) = 3, \ldots, X(FFFFFFS) = 7$, and so on. Any positive integer is a possible value of X, so the set of possible values is infinite.   ■

Example 3.5 Suppose that in some random fashion, a location (latitude and longitude) in the continental USA is selected. Define a rv Y by

$$Y = \text{the height above sea level at the selected location}$$

For example, if the selected location were (39° 50′ N, 98° 35′ W), then it might be the case that $Y((39° 50′ N, 98° 35′ W)) = 1748.26$ ft. The largest possible value of Y is 14,494 (Mt. Whitney), and the smallest possible value is −282 (Death Valley). The set of all possible values of Y is the set of all numbers in the interval between −282 and 14,494—that is,

$$\{y: y \text{ is a number}, -282 \leq y \leq 14{,}494\} = [-282, 14{,}494]$$

and there are infinitely many numbers in this interval (an entire continuum).   ■

Two Types of Random Variables
In Section 1.2 we distinguished between data resulting from observations on a counting variable and data obtained by observing values of a measurement variable. A slightly more formal distinction characterizes two different types of random variables.

DEFINITION	A **discrete** random variable is a rv whose possible values constitute either a finite set or a countably infinite set.[1]

---

[1]For those unfamiliar with the term, a countably infinite set is one for which the elements can be enumerated: a first element, a second element, and so on. The set of all positive integers and the set of all integers are both countably infinite, but an interval like [2, 5] on the number line is not.

A random variable is **continuous** if *both* of the following apply:

i. Its set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from $-\infty$ to ∞) or all numbers in a disjoint union of such intervals (e.g., [0, 10] $\cup$ [20, 30]).

ii. No possible value of the variable has positive probability, that is, $P(X = c) = 0$ for any possible value c.

Although any interval on the number line contains infinitely many numbers, it can be shown that there is no way to create a listing of all these values—there are just too many of them. The second condition describing a continuous random variable is perhaps counterintuitive, since it would seem to imply a total probability of zero for all possible values. But we shall see in Chapter 4 that *intervals* of values have positive probability; the probability of an interval will decrease to zero as the width of the interval shrinks to zero.

Example 3.6 All random variables in Examples 3.1–3.4 are discrete. As another example, suppose we select married couples at random and do a blood test on each person until we find a pair of spouses who have the same Rh factor. With X = the number of blood tests to be performed, possible values of X are 2, 4, 6, 8, …. Since the possible values have been listed in sequence, X is a discrete rv. ∎

To study basic properties of discrete rvs, only the tools of discrete mathematics—summation and differences—are required. The study of continuous variables requires the continuous mathematics of the calculus—integrals and derivatives.

Exercises: Section 3.1 (1–10)

1. A concrete beam may fail either by shear (S) or flexure (F). Suppose that three failed beams are randomly selected and the type of failure is determined for each one. Let X = the number of beams among the three selected that failed by shear. List each outcome in the sample space along with the associated value of X.

2. Give three examples of Bernoulli rvs (other than those in the text).

3. Using the experiment in Example 3.3, define two more random variables and list the possible values of each.

4. Let X = the number of nonzero digits in a randomly selected zip code. What are the possible values of X? Give three possible outcomes and their associated X values.

5. If the sample space $\mathcal{S}$ is an infinite set, does this necessarily imply that any rv X defined from $\mathcal{S}$ will have an infinite set of possible values? If yes, say why. If no, give an example.

6. Starting at a fixed time, each car entering an intersection is observed to see whether it turns left (L), right (R), or goes straight ahead (A). The experiment terminates as soon as a car is observed to turn left. Let X = the number of cars observed. What are possible X values? List five outcomes and their associated X values.

7. For each random variable defined here, describe the set of possible values for the variable, and state whether the variable is discrete.

   a. X = the number of unbroken eggs in a randomly chosen standard egg carton
   b. Y = the number of students on a class list for a particular course who are absent on the first day of classes
   c. U = the number of times a duffer has to swing at a golf ball before hitting it
   d. X = the length of a randomly selected rattlesnake

e. $Z =$ the amount of royalties earned from the sale of a first edition of 10,000 textbooks

f. $Y =$ the pH of a randomly chosen soil sample

g. $X =$ the tension (psi) at which a randomly selected tennis racket has been strung

h. $X =$ the total number of coin tosses required for three individuals to obtain a match (*HHH* or *TTT*)

8. Each time a component is tested, the trial is a success (*S*) or failure (*F*). Suppose the component is tested repeatedly until a success occurs on three *consecutive* trials. Let *Y* denote the number of trials necessary to achieve this. List all outcomes corresponding to the five smallest possible values of *Y*, and state which *Y* value is associated with each one.

9. An individual named Claudius is located at the point 0 in the accompanying diagram.
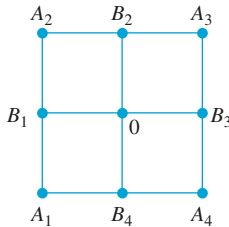
Using an appropriate randomization device (such as a tetrahedral die, one having four sides), Claudius first moves to one of the four locations B_1, B_2, B_3, B_4. Once at one of these locations, he uses another randomization device to decide whether he next returns to 0 or next visits one of the other two adjacent points. This process then continues; after each move, another move to one of the (new) adjacent points is determined by tossing an appropriate die or coin.

a. Let $X =$ the number of moves that Claudius makes before first returning to 0. What are possible values of *X*? Is *X* discrete or continuous?

b. If moves are allowed also along the diagonal paths connecting 0 to $A_1, A_2, A_3,$ and A_4, respectively, answer the questions in part (a).

10. The number of pumps in use at both a six-pump station and a four-pump station will be determined. Give the possible values for each of the following random variables:

a. $T =$ the total number of pumps in use

b. $X =$ the difference between the numbers in use at stations 1 and 2

c. $U =$ the maximum number of pumps in use at either station

d. $Z =$ the number of stations having exactly two pumps in use

3.2   Probability Distributions for Discrete Random Variables

When probabilities are assigned to various outcomes in $\mathscr{S}$, these in turn determine probabilities associated with the values of any particular rv *X*. The *probability distribution of X* says how the total probability of 1 is distributed among (allocated to) the various possible *X* values.

Example 3.7 Six batches of components are ready to be shipped by a supplier. The number of defective components in each batch is as follows:

Batch	1	2	3	4	5	6
Number of defectives	0	2	0	1	2	0

One of these batches is to be randomly selected for shipment to a customer. Let *X* be the number of defectives in the selected batch. The three possible *X* values are 0, 1, and 2. Of the six equally likely simple events, three result in $X = 0$, one in $X = 1$, and the other two in $X = 2$. Let $p(0)$ denote the probability that $X = 0$ and $p(1)$ and $p(2)$ represent the probabilities of the other two possible values of *X*. Then

$$p(0) = P(X = 0) = P(\text{lot 1 or 3 or 6 is sent}) = \frac{3}{6} = .500$$

$$p(1) = P(X = 1) = P(\text{lot 4 is sent}) = \frac{1}{6} = .167$$

$$p(2) = P(X = 2) = P(\text{lot 2 or 5 is sent}) = \frac{2}{6} = .333$$

That is, a probability of .500 is distributed to the X value 0, a probability of .167 is placed on the X value 1, and the remaining probability, .333, is associated with the X value 2. The values of X along with their probabilities collectively specify the probability distribution or *probability mass function* of X. If this experiment were repeated over and over again, in the long run $X = 0$ would occur one-half of the time, $X = 1$ one-sixth of the time, and $X = 2$ one-third of the time.   ∎

DEFINITION   The **probability distribution** or **probability mass function** (pmf) of a discrete rv is defined for every number x by
$$p(x) = P(X = x) = P(\text{all } s \in \mathcal{S}: X(s) = x).[2]$$
The **support** of $p(x)$ consists of all x values for which $p(x) > 0$. We will display a pmf for the values in its support, and it is always understood that $p(x) = 0$ otherwise (i.e., for all other x values).

In words, for every possible value x of the random variable, the pmf specifies the probability of observing that value when the experiment is performed. The conditions $p(x) \geq 0$ and $\sum p(x) = 1$, where the summation is over all possible x, are required of any pmf.

Example 3.8 Consider randomly selecting a student at a large public university, and define a Bernoulli rv by $X = 1$ if the selected student does not qualify for in-state tuition (a success from the university administration's point of view) and $X = 0$ if the student does qualify. If 20% of all students do not qualify, the pmf for X is

$$p(0) = P(X = 0) = P(\text{the selected student does qualify}) = .8$$
$$p(1) = P(X = 1) = P(\text{the selected student does not qualify}) = .2$$
$$p(x) = P(X = x) = 0 \text{ for } x \neq 0 \text{ or } 1.$$

That is,

$$p(x) = \begin{cases} .8 & \text{if } x = 0 \\ .2 & \text{if } x = 1 \end{cases}$$

Figure 3.2 (p. 117) is a picture of this pmf, called a *line graph*.   ∎

Example 3.9 An electronics laboratory has five identical-looking power sources, of which only two are fully charged. The power sources will be tested one by one until a fully charged one is found. Let the rv $Y = $ the number of tests necessary to identify a fully charged source. Let A and B represent the two fully charged power sources and C, D, E the other three. Then the pmf of Y is

---

[2]$P(X = x)$ is read "the probability that the rv X assumes the value x." For example, $P(X = 2)$ denotes the probability that the resulting X value is 2.
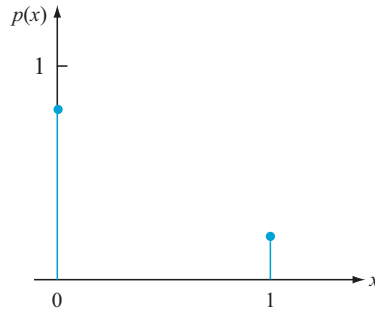
Figure 3.2 The line graph for the pmf in Example 3.8

$$p(1) = P(Y = 1) = P(\text{A or B tested first}) = \frac{2}{5} = .4$$

$$p(2) = P(Y = 2) = P(\text{C, D, or E first, and then A or B})$$

$$= P(\text{C, D, or E first}) \cdot P(\text{A or B next}|\text{ C, D, or E first}) = \frac{3}{5} \cdot \frac{2}{4} = .3$$

$$p(3) = P(Y = 3) = P(\text{C, D, or E first and second, and then A or B}) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = .2$$

$$p(4) = P(Y = 4) = P(\text{C, D, and E all done first}) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} = .1$$

$$p(y) = 0 \quad \text{for } y \neq 1, 2, 3, 4$$

The pmf can be presented compactly in tabular form:

y	1	2	3	4
$p(y)$	.4	.3	.2	.1

where any y value not listed receives zero probability. This pmf can also be displayed in a line graph (Figure 3.3).
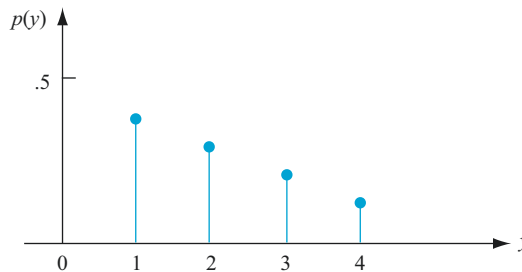
Figure 3.3 The line graph for the pmf in Example 3.9 ∎

The name "probability mass function" is suggested by a model used in physics for a system of "point masses." In this model, masses are distributed at various locations x along a one-dimensional axis. Our pmf describes how the total probability mass of 1 is distributed at various points along the axis of possible values of the random variable (where and how much mass at each x).

Another useful pictorial representation of a pmf, called a **probability histogram**, is similar to histograms discussed in Chapter 1. Above each x in the support of X, construct a rectangle centered at x. The height of each rectangle is proportional to $p(x)$, and the base is the same for all rectangles. When possible values are equally spaced, the base is frequently chosen as the distance between successive x values (though it could be smaller). Figure 3.4 shows two probability histograms.
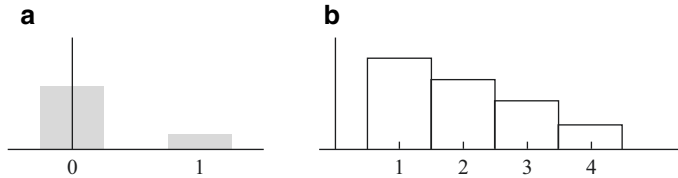
Figure 3.4 Probability histograms: (**a**) Example 3.8; (**b**) Example 3.9

A Parameter of a Probability Distribution

In Example 3.8, we had $p(0) = .8$ and $p(1) = .2$ because 20% of all students did not qualify for in-state tuition. At another university, it may be the case that $p(0) = .9$ and $p(1) = .1$. More generally, the pmf of any Bernoulli rv can be expressed in the form $p(1) = \alpha$ and $p(0) = 1 - \alpha$, where $0 < \alpha < 1$. Because the pmf depends on the particular value of α, we often write $p(x; \alpha)$ rather than just $p(x)$:

$$p(x; \alpha) = \begin{cases} 1 - \alpha & \text{if } x = 0 \\ \alpha & \text{if } x = 1 \end{cases} \tag{3.1}$$

Then each choice of α in Expression (3.1) yields a different pmf.

DEFINITION | Suppose $p(x)$ depends on a quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution. Such a quantity is called a **parameter** of the distribution. The collection of all probability distributions for different values of the parameter is called a **family** of probability distributions.

The quantity α in Expression (3.1) is a parameter. Each different number α between 0 and 1 determines a different member of a family of distributions; two such members are

$$p(x; .6) = \begin{cases} .4 & \text{if } x = 0 \\ .6 & \text{if } x = 1 \end{cases} \quad \text{and} \quad p(x; .5) = \begin{cases} .5 & \text{if } x = 0 \\ .5 & \text{if } x = 1 \end{cases}$$

Every probability distribution for a Bernoulli rv has the form of Expression (3.1), so it is called the *family of Bernoulli distributions*.

Example 3.10 In many communication systems, a receiver will send a short signal back to the transmitter to indicate whether a message has been received correctly or with errors. These signals are often called an *acknowledgement* (*A*) and a *nonacknowledgement* (*N*), respectively. (Bit sum checks and other tools are used by the receiver to determine the absence or presence of errors.) Let $p = P(A)$, assume that successive transmission attempts are independent, and define a rv X = number of attempts required to successfully transmit one message. Then

$$p(1) = P(X = 1) = P(A) = p$$
$$p(2) = P(X = 2) = P(NA) = P(N) \cdot P(A) = (1 - p)p$$

and

$$p(3) = P(X = 3) = P(NNA) = P(N) \cdot P(N) \cdot P(A) = (1 - p)^2 p$$

Continuing this way, a general formula emerges:

$$p(x) = (1 - p)^{x-1} p \quad x = 1, 2, 3, \ldots \tag{3.2}$$

The parameter p can assume any value between 0 and 1. Expression (3.2) describes the family of *geometric* distributions. In most modern communication systems, p is very close to 1, but in a noisy system (such as on a WIFI network with lots of interference and/or intervening walls), p could be considerably lower. ∎

The Cumulative Distribution Function

For some fixed value x, we often wish to compute the probability that the observed value of X will be *at most x*. For example, the pmf in Example 3.7 was

$$p(x) = \begin{cases} .500 & x = 0 \\ .167 & x = 1 \\ .333 & x = 2 \end{cases}$$

The probability that X is at most 1 is then

$$P(X \le 1) = p(0) + p(1) = .500 + .167 = .667$$

In this example, $X \le 1.5$ iff $X \le 1$, so $P(X \le 1.5) = P(X \le 1) = .667$. Similarly, $P(X \le 0) = P(X = 0) = .5$, and $P(X \le .75) = .5$ also. Since 0 is the smallest possible value of X, $P(X \le -1.7) = 0$, $P(X \le -.0001) = 0$, and so on. The largest possible X value is 2, so $P(X \le 2) = 1$. And if x is any number larger than 2, $P(X \le x) = 1$; that is, $P(X \le 5) = 1$, $P(X \le 10.23) = 1$, and so on.

Critically, notice that $P(X < 1) = P(X = 0) = .5 \ne P(X \le 1)$, since the latter probability includes the probability mass at the x value 1, while $P(X < 1)$ does not. When X is a discrete random variable and x is in the support of X, $P(X < x) < P(X \le x)$.

DEFINITION	The **cumulative distribution function (cdf)** $F(x)$ of a discrete rv variable X with pmf $p(x)$ is defined for every number x by

$$F(x) = P(X \le x) = \sum_{y:y \le x} p(y) \tag{3.3}$$

For any number x, $F(x)$ is the probability that the observed value of X will be at most x.

Example 3.11 An online retailer sells flash drives with either 16, 32, 64, 128, or 256 GB of memory. The accompanying table gives the distribution of Y = the amount of memory in a purchased drive:

y	16	32	64	128	256
$p(y)$	.05	.10	.35	.40	.10

Let's first determine $F(y)$ for each of the five possible values of Y:

$$F(16) = P(Y \le 16) = P(Y = 16) = p(16) = .05$$
$$F(32) = P(Y \le 32) = P(Y = 16 \text{ or } 32) = p(16) + p(32) = .15$$
$$F(64) = P(Y \le 64) = P(Y = 16 \text{ or } 32 \text{ or } 64) = p(16) + p(32) + p(64) = .50$$
$$F(128) = P(Y \le 128) = p(16) + p(32) + p(64) + p(128) = .90$$
$$F(256) = P(Y \le 256) = 1$$

Now for any other number y, $F(y)$ will equal the value of F at the closest possible value of y to the left of y. For example,

$$F(48.7) = P(Y \le 48.7) = P(Y \le 32) = F(32) = .15$$
$$F(127.999) = P(Y \le 127.999) = P(Y \le 64) = F(64) = .50$$

If y is less than 16, $F(y) = 0$ [e.g., $F(8) = 0$], and if y is at least 256, $F(y) = 1$ [e.g., $F(512) = 1$]. The cdf is thus

$$F(y) = \begin{cases} 0 & y < 16 \\ .05 & 16 \le y < 32 \\ .15 & 32 \le y < 64 \\ .50 & 64 \le y < 128 \\ .90 & 128 \le y < 256 \\ 1 & 256 \le y \end{cases}$$

A graph of this cdf is shown in Figure 3.5.
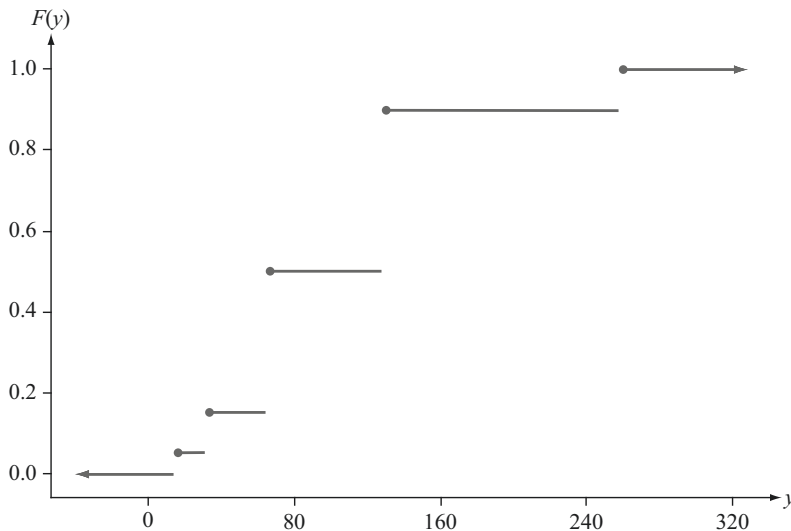
Figure 3.5   A graph of the cdf of Example 3.11   ∎

For a discrete rv X, the graph of $F(x)$ will have a jump at every possible value of X and will be flat between possible values. Such a graph is called a **step function**.

Example 3.12 In Example 3.10, any positive integer was a possible X value, and the pmf was

$$p(x) = (1-p)^{x-1}p \quad x = 1, 2, 3, \ldots$$

For any positive integer x,

$$F(x) = \sum_{y \le x} p(y) = \sum_{y=1}^{x} (1-p)^{y-1}p = p \sum_{y=0}^{x-1} (1-p)^{y} \tag{3.4}$$

To evaluate this sum, we use the fact that the partial sum of a geometric series is

$$\sum_{y=0}^{k} a^y = \frac{1 - a^{k+1}}{1 - a}$$

Using this in Equation (3.4), with $a = 1 - p$ and $k = x - 1$, gives

$$F(x) = p \cdot \frac{1 - (1-p)^x}{1 - (1-p)} = 1 - (1-p)^x \quad x \text{ a positive integer}$$

Since F is constant in between positive integers,

$$F(x) = \begin{cases} 0 & x < 1 \\ 1 - (1-p)^{[x]} & x \ge 1 \end{cases} \tag{3.5}$$

where $[x]$ is the largest integer $\le x$ (e.g., $[2.7] = 2$).

In an extremely noisy channel with $p = .15$, the probability of having to transmit a message at most 5 times to get an acknowledgement is $F(5) = 1 - (1 - .15)^5 = .5563$, whereas $F(50) \approx 1.0000$. This cdf is graphed in Figure 3.6.
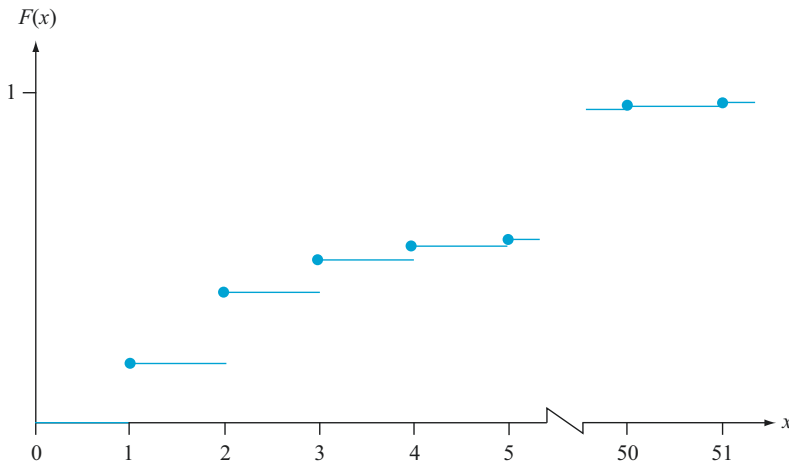
Figure 3.6 A graph of $F(x)$ for Example 3.12

In our examples thus far, the cdf has been derived from the pmf. This process can be reversed to obtain the pmf from the cdf whenever the latter function is available. Suppose, for example, that X represents the number of defective components in a shipment consisting of six components, so that possible X values are 0, 1, …, 6. Then

$$
\begin{aligned}
p(3) &= P(X = 3) \\
&= [p(0) + p(1) + p(2) + p(3)] - [p(0) + p(1) + p(2)] \\
&= P(X \le 3) - P(X \le 2) \\
&= F(3) - F(2)
\end{aligned}
$$

More generally, the probability that X falls in a specified interval is easily obtained from the cdf. For example,

$$
\begin{aligned}
P(2 \le X \le 4) &= p(2) + p(3) + p(4) \\
&= [p(0) + \cdots + p(4)] - [p(0) + p(1)] \\
&= P(X \le 4) - P(X \le 1) \\
&= F(4) - F(1)
\end{aligned}
$$

Notice that $P(2 \le X \le 4) \ne F(4) - F(2)$. This is because the X value 2 is included in $2 \le X \le 4$, so we do not want to subtract out its probability. However, $P(2 < X \le 4) = F(4) - F(2)$ because $X = 2$ is not included in the interval $2 < X \le 4$.

PROPOSITION   For any two numbers a and b with $a \le b$,

$$
P(a \le X \le b) = F(b) - F(a-)
$$

where $F(a-)$ represents the limit of $F(x)$ as x approaches a from the left. In particular, if the only possible values are integers and if a and b are integers, then

$$
\begin{aligned}
P(a \le X \le b) &= P(X = a \text{ or } a+1 \text{ or } \ldots \text{ or } b) \\
&= F(b) - F(a-1)
\end{aligned}
$$

Taking $a = b$ yields $P(X = a) = F(a) - F(a - 1)$ in this case.

The reason for subtracting $F(a-)$ rather than $F(a)$ is that we want to include the probability mass at $X = a$; $F(b) - F(a)$ gives $P(a < X \le b)$. This proposition will be used extensively when computing binomial and Poisson probabilities in Sections 3.5 and 3.6.

Example 3.13 Let X = the number of days of sick leave taken by a randomly selected employee of a large company during a particular year. If the maximum number of allowable sick days per year is 14, possible values of X are 0, 1, …, 14. With $F(0) = .58$, $F(1) = .72$, $F(2) = .76$, $F(3) = .81$, $F(4) = .88$, and $F(5) = .94$,

$$
P(2 \le X \le 5) = P(X = 2, 3, 4, \text{ or } 5) = F(5) - F(1) = .22
$$

and

$$P(X = 3) = F(3) - F(2) = .05$$ ∎

Another View of Probability Mass Functions

It is often helpful to think of a pmf as specifying a mathematical model for a discrete population.

Example 3.14 Consider selecting at random a student who is among the 15,000 registered for the current term at a certain university. Let X = the number of courses for which the selected student is registered, and suppose that X has the following pmf:

x	1	2	3	4	5	6	7
$p(x)$	.01	.03	.13	.25	.39	.17	.02

One way to view this situation is to think of the population as consisting of 15,000 individuals, each having his or her own X value; the proportion with each X value is given by $p(x)$. An alternative viewpoint is to forget about the students and think of the population itself as consisting of the X values: There are some 1's in the population, some 2's, …, and finally some 7's. The population then consists of the numbers 1, 2, …, 7 (so is discrete), and $p(x)$ gives a model for the distribution of population values. ∎

Once we have such a population model, we will use it to compute values of population characteristics (e.g., the mean μ) and make inferences about such characteristics.

Exercises: Section 3.2 (11–27)

11. Let X be the number of students who show up at a professor's office hours on a particular day. Suppose that the only possible values of X are 0, 1, 2, 3, and 4, and that $p(0) = .30$, $p(1) = .25$, $p(2) = .20$, and $p(3) = .15$.
    a. What is $p(4)$?
    b. Draw both a line graph and a probability histogram for the pmf of X.
    c. What is the probability that at least two students come to the office hour? What is the probability that more than two students come to the office hour?
    d. What is the probability that the professor shows up for his office hour?

12. Airlines sometimes overbook flights. Suppose that for a plane with 50 seats, 55 passengers have tickets. Define the random variable Y as the number of ticketed passengers who actually show up for the flight.

The probability mass function of Y appears in the accompanying table.

y	45	46	47	48	49	50	51	52	53	54	55
$p(y)$	.05	.10	.12	.14	.25	.17	.06	.05	.03	.02	.01

a. What is the probability that the flight will accommodate all ticketed passengers who show up?
b. What is the probability that not all ticketed passengers who show up can be accommodated?
c. If you are the first person on the standby list (which means you will be the first one to get on the plane if there are any seats available after all ticketed passengers have been accommodated), what is the probability that you will be able to take the flight? What is this probability if you are the third person on the standby list?

13. A mail-order computer business has six telephone lines. Let X denote the number of lines in use at a specified time. Suppose the pmf of X is as given in the accompanying table.

x	0	1	2	3	4	5	6
$p(x)$	.10	.15	.20	.25	.20	.06	.04

Calculate the probability of each of the following events.

a. {at most three lines are in use}
b. {fewer than three lines are in use}
c. {at least three lines are in use}
d. {between two and five lines, inclusive, are in use}
e. {between two and four lines, inclusive, are not in use}
f. {at least four lines are not in use}

14. A contractor is required by a county planning department to submit one, two, three, four, or five forms (depending on the nature of the project) in applying for a building permit. Let $Y =$ the number of forms required of the next applicant. The probability that y forms are required is known to be proportional to y—that is, $p(y) = ky$ for $y = 1, \ldots, 5$.

a. What is the value of k? [*Hint:* $\sum_{y=1}^{5} p(y) = 1.$]
b. What is the probability that at most three forms are required?
c. What is the probability that between two and four forms (inclusive) are required?
d. Could $p(y) = y^2/50$ for $y = 1, \ldots, 5$ be the pmf of Y?

15. Many manufacturers have quality control programs that include inspection of incoming materials for defects. Suppose a computer manufacturer receives computer boards in lots of five. Two boards are selected from each lot for inspection. We can represent possible outcomes of the selection process by pairs. For example, the pair (1, 2) represents the selection of boards 1 and 2 for inspection.

a. List the ten different possible outcomes.
b. Suppose that boards 1 and 2 are the only defective boards in a lot of five. Two boards are to be chosen at random. Define X to be the number of defective boards observed among those inspected. Determine the probability distribution of X.
c. Let $F(x)$ denote the cdf of X. First determine $F(0) = P(X \leq 0)$, $F(1)$, and $F(2)$, and then obtain $F(x)$ for all other x.

16. Some parts of California are particularly earthquake-prone. Suppose that in one such area, 30% of all homeowners are insured against earthquake damage. Four homeowners are to be selected at random; let X denote the number among the four who have earthquake insurance.

a. Find the probability distribution of X. [*Hint*: Let S denote a homeowner who has insurance and F one who does not. One possible outcome is $SFSS$, with probability $(.3)(.7)(.3)(.3)$ and associated X value 3. There are 15 other outcomes.]
b. Draw the corresponding probability histogram.
c. What is the most likely value for X?
d. What is the probability that at least two of the four selected have earthquake insurance?

17. A new battery's voltage may be acceptable (A) or unacceptable (U). A certain flashlight requires two batteries, so batteries will be independently selected and tested until two acceptable ones have been found. Suppose that 90% of all batteries have acceptable voltages. Let Y denote the number of batteries that must be tested.

a. What is $p(2)$, that is, $P(Y = 2)$?
b. What is $p(3)$? [*Hint:* There are two different outcomes that result in $Y = 3$.]
c. To have $Y = 5$, what must be true of the fifth battery selected? List the four

outcomes for which $Y = 5$ and then determine $p(5)$.

d. Use the pattern in your answers for parts (a)–(c) to obtain a general formula for $p(y)$.

18. Two fair six-sided dice are tossed independently. Let M = the maximum of the two tosses [thus $M(1, 5) = 5$, $M(3, 3) = 3$, etc.].

a. What is the pmf of M? [*Hint:* First determine $p(1)$, then $p(2)$, and so on.]

b. Determine the cdf of M and graph it.

19. Suppose that you read through this year's issues of the *New York Times* and record each number that appears in a news article—the income of a CEO, the number of cases of wine produced by a winery, the total charitable contribution of a politician during the previous tax year, the age of a celebrity, and so on. Now focus on the leading digit of each number, which could be 1, 2, …, 8, or 9. Your first thought might be that the leading digit X of a randomly selected number would be equally likely to be one of the nine possibilities (a discrete uniform distribution). However, much empirical evidence as well as some theoretical arguments suggest an alternative probability distribution called *Benford's law*:

$$p(x) = P(\text{1st digit is } x) = \log_{10}\left(\frac{x+1}{x}\right)$$

$$x = 1, 2, \ldots, 9$$

a. Without computing individual probabilities from this formula, show that it specifies a legitimate pmf.

b. Now compute the individual probabilities and compare to the corresponding discrete uniform distribution.

c. Obtain the cdf of X.

d. Using the cdf, what is the probability that the leading digit is at most 3? At least 5?

[*Note*: Benford's law is the basis for some auditing procedures used to detect fraud in financial reporting—for example, by the Internal Revenue Service.]

20. A library subscribes to two different weekly news magazines, each of which is supposed to arrive in Wednesday's mail. In actuality, each one may arrive on Wednesday, Thursday, Friday, or Saturday. Suppose the two arrive independently of one another, and for each one $P(\text{W}) = .3$, $P(\text{Th}) = .4$, $P(\text{F}) = .2$, and $P(\text{S}) = .1$. Let Y = the number of days beyond Wednesday that it takes for both magazines to arrive (so possible Y values are 0, 1, 2, or 3). Compute the pmf of Y. [*Hint:* There are 16 possible outcomes; $Y(\text{W, W}) = 0$, $Y(\text{F, Th}) = 2$, and so on.]

21. Refer to Exercise 13, and calculate and graph the cdf $F(x)$. Then use it to calculate the probabilities of the events given in parts (a)–(d) of that problem.

22. Let X denote the number of vehicles queued up at a bank's drive-up window at a particular time of day. The cdf of X is as follows:

$$F(x) = \begin{cases} 0 & x < 0 \\ .06 & 0 \le x < 1 \\ .19 & 1 \le x < 2 \\ .39 & 2 \le x < 3 \\ .67 & 3 \le x < 4 \\ .92 & 4 \le x < 5 \\ .97 & 5 \le x < 6 \\ 1 & 6 \le x \end{cases}$$

Calculate the following probabilities directly from the cdf:

a. $p(2)$, that is, $P(X = 2)$

b. $P(X > 3)$

c. $P(2 \le X \le 5)$

d. $P(2 < X < 5)$

23. An insurance company offers its policyholders a number of different premium payment options. For a randomly selected policyholder, let X = the number of months between successive payments. The cdf of X is as follows:

$$F(x) = \begin{cases} 0 & x < 1 \\ .30 & 1 \le x < 3 \\ .40 & 3 \le x < 4 \\ .45 & 4 \le x < 6 \\ .60 & 6 \le x < 12 \\ 1 & 12 \le x \end{cases}$$

a. What is the pmf of X?
b. Using just the cdf, compute
   $P(3 \le X \le 6)$ and $P(4 \le X)$.

24. In Example 3.10, let Y = the number of
*non*acknowledgements before the experi-
ment terminates. With $p = P(A)$ and
$1 - p = P(N)$, what is the pmf of Y? [*Hint*:
First list the possible values of Y, starting
with the smallest, and proceed until you see
a general formula.]

25. Alvie Singer lives at 0 in the accompanying
diagram and has four friends who live at A,
B, C, and D. One day Alvie decides to go
visiting, so he tosses a fair coin twice to
decide which of the four to visit. Once at a
friend's house, he will either return home or
else proceed to one of the two adjacent
houses (such as 0, A, or C when at B), with
each of the three possibilities having prob-
ability $1/3$. In this way, Alvie continues to
visit friends until he returns home.

a. Let X = the number of times that Alvie
visits a friend. Derive the pmf of X.
b. Let Y = the number of straight-line
segments that Alvie traverses (includ-
ing those leading to and from 0). What
is the pmf of Y?
c. Suppose that female friends live at
A and C and male friends at B and D. If
Z = the number of visits to female
friends, what is the pmf of Z?

26. After all students have left the classroom, a
statistics professor notices that four copies
of the text were left under desks. At the
beginning of the next lecture, the professor
distributes the four books in a completely
random fashion to each of the four students
(1, 2, 3, and 4) who claim to have left
books. One possible outcome is that 1
receives 2's book, 2 receives 4's book, 3
receives his or her own book, and 4
receives 1's book. This outcome can be
abbreviated as (2, 4, 3, 1).

a. List the other 23 possible outcomes.
b. Let X denote the number of students
who receive their own book. Determine
the pmf of X.

27. Show that the cdf $F(x)$ is a nondecreasing
function; that is, $x_1 < x_2$ implies that
$F(x_1) \le F(x_2)$. Under what condition will
$F(x_1) = F(x_2)$?
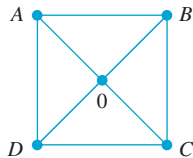
3.3 Expected Values of Discrete Random Variables

In Example 3.14, we considered a university with 15,000 students and let X = the number of courses
for which a randomly selected student is registered. The pmf of X follows. Since $p(1) = .01$, we know
that $(.01) \cdot (15,000) = 150$ of the students are registered for one course, and similarly for the other
x values.

x	1	2	3	4	5	6	7	
$p(x)$	.01	.03	.13	.25	.39	.17	.02	(3.6)
Number registered	150	450	1950	3750	5850	2550	300	

To compute the average number of courses per student, i.e., the average value of X in the population, we should calculate the total number of courses and divide by the total number of students. Since each of 150 students is taking one course, these 150 contribute 150 courses to the total. Similarly, 450 students contribute 2(450) courses, and so on. The population average value of X is then

$$\frac{1(150) + 2(450) + 3(1950) + \cdots + 7(300)}{15{,}000} = 4.57 \tag{3.7}$$

Since $150/15{,}000 = .01 = p(1)$, $450/15{,}000 = .03 = p(2)$, and so on, an alternative expression for (3.7) is

$$1 \cdot p(1) + 2 \cdot p(2) + \cdots + 7 \cdot p(7) \tag{3.8}$$

Expression (3.8) shows that to compute the population average value of X, we need only the possible values of X along with their probabilities (proportions). In particular, the population size is irrelevant as long as the pmf is given by (3.6). The average or mean value of X is then a *weighted* average of the possible values 1, …, 7, where the weights are the probabilities of those values.

The Expected Value of *X*

DEFINITION   Let X be a discrete rv with set of possible values D and pmf $p(x)$. The **expected value** or **mean value** of X, denoted by $E(X)$ or μ_X or just μ, is

$$E(X) = \mu_X = \mu = \sum_{x \in D} x \cdot p(x)$$

This expected value will exist provided that $\sum_{x \in D} |x| \cdot p(x) < \infty$.

Example 3.15 For the pmf in (3.6),

$$\begin{aligned}
\mu &= 1 \cdot p(1) + 2 \cdot p(2) + \cdots + 7 \cdot p(7) \\
&= (1)(.01) + (2)(.03) + \cdots + (7)(.02) \\
&= .01 + .06 + .39 + 1.00 + 1.95 + 1.02 + .14 = 4.57
\end{aligned}$$

If we think of the population as consisting of the X values 1, 2, …, 7, then $\mu = 4.57$ is the population mean (we will often refer to μ as the *population mean* rather than "the mean of X in the population"). Notice that μ here is not 4, the ordinary average of 1, …, 7, because the distribution puts more weight on 4, 5, and 6 than on other X values. ∎

In Example 3.15, the expected value μ was 4.57, which is not a possible value of X. The word *expected* should be interpreted with caution because one would not expect to see an X value of 4.57 when a single student is selected.

Example 3.16 Just after birth, each newborn child is rated on a scale called the Apgar scale. The possible ratings are 0, 1, …, 10, with the child's rating determined by color, muscle tone, respiratory effort, heartbeat, and reflex irritability (the best possible score is 10). Let X be the Apgar score of a randomly selected child born at a certain hospital during the next year, and suppose that the pmf of X is

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	.002	.001	.002	.005	.02	.04	.18	.37	.25	.12	.01

Then the mean value of X is

$$E(X) = \mu = (0)(.002) + (1)(.001) + (2)(.002)$$
$$+ \cdots + (8)(.25) + (9)(.12) + (10)(.01) = 7.15$$

(Again, μ is not a possible value of the variable X.) If the stated model is correct, then the mean Apgar score for the population of all children born at this hospital next year will be 7.15. ∎

Example 3.17 Let $X = 1$ if a randomly selected component needs warranty service and $= 0$ otherwise. If the chance a component needs warranty service is p, then X is a Bernoulli rv with pmf $p(1) = p$ and $p(0) = 1 - p$, from which

$$E(X) = 0 \cdot p(0) + 1 \cdot p(1) = 0(1 - p) + 1(p) = p$$

That is, the expected value of X is just the probability that X takes on the value 1. If we conceptualize a population consisting of 0's in proportion $1 - p$ and 1's in proportion p, then the population average is $\mu = p$. ∎

There is another frequently used interpretation of μ. Consider observing a first value x of our random variable, then observe independently another value, then another, and so on. If after a very large number of x values we average them, the resulting sample average will typically be close to μ; a more rigorous version of this statement is provided by the *Law of Large Numbers* in Chapter 6. That is, μ can be interpreted as the long-run average value of X when the experiment is performed repeatedly. This interpretation is often appropriate for games of chance, where the "population" is not a concrete set of individuals but rather the results of all hypothetical future instances of playing the game.

Example 3.18 A standard American roulette wheel has 38 spaces. Players bet on which space a marble will land in once the wheel has been spun. One of the simplest bets is based on the color of the space: 18 spaces are black, 18 are red, and 2 are green. So, if a player "bets on black," s/he has an 18/38 chance of winning. Casinos consider color bets an "even wager," meaning that a player who wages $1 on black, say, will profit $1 if the marble lands in a black space (and lose the wagered $1 otherwise).

Let $X =$ the return on a $1 wager on black. Then the pmf of X is

x	–$1	+$1
$p(x)$	20/38	18/38

and the expected value of X is $E(X) = (-1)(20/38) + (1)(18/38) = -2/38 = -\$.0526$. If a player makes $1 bets on black on successive spins of the roulette wheel, in the long run s/he can expect to

lose about 5.26 cents per wager. Since players don't necessarily make a large number of wagers, this long-run average interpretation is perhaps more apt from the casino's perspective: in the long run, they will gain an average of 5.26 cents for every $1 wagered on black at the roulette table. ∎

Thus far, we have assumed that the mean of any given distribution exists. If the set of possible values of X is unbounded, so that the sum for μ_X is actually an infinite series, the expected value of X might or might not exist, depending on whether the series converges or diverges.

Example 3.19 From Example 3.10, the general form for the pmf of X = number of attempts required to successfully transmit one message is

$$p(x) = (1 - p)^{x-1}p \quad x = 1, 2, 3, \ldots$$

From the definition,

$$
\begin{aligned}
E(X) &= \sum_D x \cdot p(x) = \sum_{x=1}^{\infty} xp(1 - p)^{x-1} = p\sum_{x=1}^{\infty} x(1 - p)^{x-1} \\
&= p\sum_{x=1}^{\infty} \left[-\frac{d}{dp}(1 - p)^x \right]
\end{aligned}
\tag{3.9}
$$

If we interchange the order of taking the derivative and the summation, the sum is that of a geometric series. A little calculus reveals that the final result is $E(X) = 1/p$. If p is near 1, we expect a successful transmission very soon, whereas if p is near 0, we expect many attempts before the first success. For $p = .5$, $E(X) = 2$. ∎

Example 3.20 Let X, the number of interviews a student has prior to getting a job, have pmf

$$p(x) = k/x^2 \quad x = 1, 2, 3, \ldots$$

where k is chosen so that $\sum_{x=1}^{\infty} (k/x^2) = 1$. [Because $\sum_{x=1}^{\infty} (1/x^2) = \pi^2/6$, the value of k is $6/\pi^2$.] The expected value of X is

$$\mu = E(X) = \sum_{x=1}^{\infty} x\frac{k}{x^2} = k\sum_{x=1}^{\infty} \frac{1}{x} \tag{3.10}$$

The sum on the right of Equation (3.10) is the famous harmonic series of mathematics and can be shown to equal ∞. $E(X)$ is not finite here because $p(x)$ does not decrease sufficiently fast as x increases; statisticians say that the probability distribution of X has "a heavy tail." If a sequence of X values is chosen using this distribution, the sample average will not settle down to some finite number but will tend to grow without bound.

Statisticians use the phrase "heavy tails" in connection with any distribution having a large amount of probability far from μ (so heavy tails do not require $\mu = \infty$). Such heavy tails make it difficult to make inferences about μ. ∎

The Expected Value of a Function

Often we will be interested in the expected value of some function $h(X)$ rather than X itself. An easy way of computing the expected value of $h(X)$ is suggested by the following example.

Example 3.21 The cost of a certain vehicle diagnostic test depends on the number of cylinders X in the vehicle's engine. Suppose the cost function is $h(X) = 20 + 3X + .5X^2$. Since X is a random variable, so is $Y = h(X)$. The pmf of X and the derived pmf of Y are as follows:

x	4	6	8		y	40	56	76
$p(x)$	.5	.3	.2	$\Rightarrow$	$p(y)$	.5	.3	.2

With D^* denoting the possible values of Y,

$$
\begin{aligned}
E[h(X)] = E(Y) = \sum_{y \in D^*} y \cdot p(y) &= (40)(.5) + (56)(.3) + (76)(.2) = \$52 \\
= h(4) \cdot (.5) + h(6) \cdot (.3) + h(8) \cdot (.2) &= \sum_D h(x) \cdot p(x)
\end{aligned}
\tag{3.11}
$$

According to Expression (3.11), it was not necessary to determine the pmf of Y to obtain $E(Y)$; instead, the desired expected value is a weighted average of the possible $h(x)$ (rather than x) values. ∎

LAW OF THE UNCONSCIOUS STATISTICIAN	If the rv X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(X)$, denoted by $E[h(X)]$ or $\mu_{h(X)}$, is computed by $$E[h(X)] = \sum_{x \in D} h(x) \cdot p(x)$$ assuming that $\sum_D |h(x)| \cdot p(x) < \infty$.

According to this proposition, $E[h(X)]$ is computed in the same way that $E(X)$ itself is, except that $h(x)$ is substituted in place of x. That is, $E[h(X)]$ is a weighted average of possible $h(X)$ values, where the weights are the probabilities of the corresponding original X values.

Example 3.22 A computer store has purchased three computers at \$500 apiece. It will sell them for \$1000 apiece. The manufacturer has agreed to repurchase any computers still unsold after a specified period at \$200 apiece. Let X denote the number of computers sold, and suppose that $p(0) = .1$, $p(1) = .2$, $p(2) = .3$, and $p(3) = .4$. With $h(X)$ denoting the profit associated with selling X units, the given information implies that $h(X) = \text{revenue} - \text{cost} = 1000X + 200(3 - X) - 1500 = 800X - 900$. The expected profit is then

$$
\begin{aligned}
E[h(X)] &= h(0) \cdot p(0) + h(1) \cdot p(1) + h(2) \cdot p(2) + h(3) \cdot p(3) \\
&= (-900)(.1) + (-100)(.2) + (700)(.3) + (1500)(.4) \\
&= \$700
\end{aligned}
$$

∎

Because an expected value is a sum, it possesses the same properties as any summation; specifically, the expected value "operator" can be distributed across addition and across multiplication by constants. This important property is known as *linearity of expectation*.

LINEARITY OF EXPECTATION	For any functions $h_1(X)$ and $h_2(X)$ and any constants a_1, a_2, and b,

$$E[a_1 h_1(X) + a_2 h_2(X) + b] = a_1 E[h_1(X)] + a_2 E[h_2(X)] + b$$

In particular, for any linear function $aX + b$,

$$E(aX + b) = a \cdot E(X) + b \tag{3.12}$$

(or, using alternative notation, $\mu_{aX+b} = a \cdot \mu_X + b$).

Proof Let $h(X) = a_1 h_1(X) + a_2 h_2(X) + b$, and apply the Law of the Unconscious Statistician:

$$
\begin{aligned}
E[a_1 h_1(X) + a_2 h_2(X) + b] &= \sum_D (a_1 h_1(x) + a_2 h_2(x) + b) \cdot p(x) \\
&= a_1 \sum_D h_1(x) \cdot p(x) + a_2 \sum_D h_2(x) \cdot p(x) + b \sum_D p(x) \\
&\qquad \text{(distributive property of addition)} \\
&= a_1 E[h_1(X)] + a_2 E[h_2(X)] + b[1] = a_1 E[h_1(X)] + a_2 E[h_2(X)] + b
\end{aligned}
$$

The special case of $aX + b$ is obtained by setting $a_1 = a$, $h_1(X) = X$, and $a_2 = 0$. ∎

By induction, linearity of expectation applies to any finite number of terms. In Example 3.21, straightforward computation gives $E(X) = 4(.5) + 6(.3) + 8(.2) = 5.4$ and $E(X^2) = \sum x^2 \cdot p(x) = 4^2(.5) + 6^2(.3) + 8^2(.2) = 31.6$. Applying linearity of expectation to $Y = h(X) = 20 + 3X + .5X^2$, we obtain

$$\mu_Y = E[20 + 3X + .5X^2] = 20 + 3E(X) + .5E(X^2) = 20 + 3(5.4) + .5(31.6) = \$52,$$

which matches the result of Example 3.21.

The special case (3.12) states that the expected value of a linear function equals the linear function evaluated at the expected value $E(X)$. Because $h(X)$ in Example 3.22 is linear and $E(X) = 2$, $E[h(X)] = 800(2) - 900 = \$700$, as before. Two special cases of (3.12) yield two important rules of expected value:

1. For any constant a, $\mu_{aX} = a \cdot \mu_X$ (take $b = 0$).
2. For any constant b, $\mu_{X+b} = \mu_X + b = E(X) + b$ (take $a = 1$).

Multiplication of X by a constant a changes the unit of measurement (from dollars to cents, where $a = 100$; inches to cm, where $a = 2.54$; etc.). Rule 1 says that the expected value in the new units equals the expected value in the old units multiplied by the conversion factor a. Similarly, if the constant b is added to each possible value of X, then the expected value will be shifted by that same constant amount.

One commonly made error is to substitute μ_X directly into the function $h(X)$ when h is a nonlinear function, in which case (3.12) does not apply. Consider Example 3.21: the mean of X is 5.4, and it's tempting to infer that the mean of $Y = h(X)$ is simply $h(5.4)$. However, since the function $h(X) = 20 + 3X + .5X^2$ is *not* linear in X, this does not yield the correct answer:

$$h(5.4) = 20 + 3(5.4) + .5(5.4)^2 = \$50.78 \neq \$52 = \mu_Y$$

In general, $\mu_{h(X)}$ does not equal $h(\mu_X)$ *unless the function h(x) is linear.*

The Variance and Standard Deviation of *X*

The expected value of *X* describes where the probability distribution is centered. Using the physical analogy of placing point mass $p(x)$ at the value x on a one-dimensional axis, if the axis were then supported by a fulcrum placed at μ, there would be no tendency for the axis to tilt. This is illustrated for two different distributions in Figure 3.7.
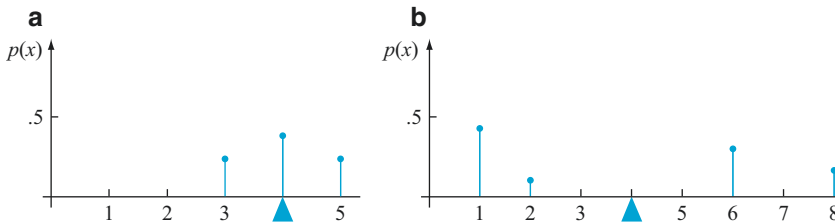
Figure 3.7 Two different probability distributions with $\mu = 4$

Although both distributions pictured in Figure 3.7 have the same mean/fulcrum μ, the distribution of Figure 3.7b has greater spread or variability or dispersion than does that of Figure 3.7a. We will use the variance of *X* to assess the amount of variability in (the distribution of) *X*, just as s^2 was used in Chapter 1 to measure variability in a sample.

DEFINITION

Let *X* have pmf $p(x)$ and expected value μ. Then the **variance** of *X*, denoted by $V(X)$ or σ_X^2 or just σ^2, is

$$V(X) = \sum_D [(x - \mu)^2 \cdot p(x)] = E[(X - \mu)^2]$$

The **standard deviation** of *X*, denoted by SD(*X*) or σ_X or just σ, is

$$\sigma_X = \sqrt{V(X)}$$

The quantity $h(X) = (X - \mu)^2$ is the squared deviation of *X* from its mean, and σ^2 is the expected squared deviation—i.e., a weighted average of the squared deviations from μ. Taking the square root of the variance to obtain the standard deviation returns us to the original units of the variable; e.g., if *X* is measured in dollars, then both μ and σ also have units of dollars. If most of the probability distribution is close to μ, as in Figure 3.7a, then σ will typically be relatively small. However, if there are *x* values far from μ that have large probabilities (as in Figure 3.7b), then σ will be larger. Intuitively, the value of σ describes a typical deviation from μ.

Example 3.23 Consider again the distribution of the Apgar score *X* of a randomly selected newborn described in Example 3.16. The mean value of *X* was calculated as $\mu = 7.15$, so

$$V(X) = \sigma^2 = \sum_{x=0}^{10} (x - 7.15)^2 \cdot p(x)$$

$$= (0 - 7.15)^2 (.002) + \cdots + (10 - 7.15)^2 (.01) = 1.5815$$

The standard deviation of X is $\sigma = \sqrt{1.5815} = 1.26$.   ∎

When the pmf $p(x)$ specifies a mathematical model for the distribution of population values, σ^2 is the population variance, and σ is the population standard deviation.

Properties of Variance
An alternative to the defining formula for $V(X)$ reduces the computational burden.

PROPOSITION $$V(X) = \sigma^2 = E(X^2) - \mu^2$$

This equation is referred to as the *variance shortcut formula*.

In using this formula, $E(X^2)$ is computed first without any subtraction; then μ is computed, squared, and subtracted (once) from $E(X^2)$. This formula is more efficient because it entails only one subtraction, and $E(X^2)$ does not require calculating squared deviations from μ.

Example 3.24   Referring back to the Apgar score scenario of Examples 3.16 and 3.23,

$$E(X^2) = \sum_{x=1}^{10} x^2 \cdot p(x) = (0^2)(.002) + (1^2)(.001) + \cdots + (10^2)(.01) = 52.704$$

Thus, $\sigma^2 = 52.704 - (7.15)^2 = 1.5815$ as before, and again $\sigma = 1.26$.   ∎

Proof of the Variance Shortcut Formula   Expand $(X - \mu)^2$ in the definition of $V(X)$, and then apply linearity of expectation:

$$V(X) = E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2]$$
$$= E(X^2) - 2\mu E(X) + \mu^2$$
$$\text{(by linearity of expectation)}$$
$$= E(X^2) - 2\mu \cdot \mu + \mu^2 = E(X^2) - 2\mu^2 + \mu^2$$
$$= E(X^2) - \mu^2$$ ∎

The quantity $E(X^2)$ in the variance shortcut formula is called the **mean-square value** of the random variable X. Engineers may be familiar with the root-mean-square, or RMS, which is the square root of $E(X^2)$. Do not confuse this with the square of the mean of X, i.e. μ^2! For example, if X has a mean of 7.15, the mean-square value of X is *not* $(7.15)^2$, because $h(x) = x^2$ is not linear. (In Example 3.24, the mean-square value of X is 52.704.) It helps to look at the two formulas side by side:

$$E(X^2) = \sum_D x^2 \cdot p(x) \quad \text{versus} \quad \mu^2 = \left(\sum_D x \cdot p(x) \right)^2$$

The order of operations is clearly different. In fact, it can be shown using the variance shortcut formula that $E(X^2) \geq \mu^2$ for every random variable, with equality if and only if X is constant.

The variance of a function $h(X)$ is the expected value of the squared difference between $h(X)$ and its expected value:

$$V[h(X)] = \sigma^2_{h(X)} = \sum_D \left[(h(x) - \mu_{h(X)})^2 \cdot p(x) \right] = \left[\sum_D h^2(x) \cdot p(x) \right] - \left[\sum_D h(x) \cdot p(x) \right]^2$$

When $h(x)$ is a linear function, $V[h(X)]$ simplifies considerably (see Exercise 42 for a proof).

PROPOSITION $V(aX + b) = \sigma^2_{aX+b} = a^2 \cdot \sigma^2_X \quad \text{and} \quad \sigma_{aX+b} = |a| \cdot \sigma_X$ (3.13)

In particular,

$$\sigma_{aX} = |a| \cdot \sigma_X \quad \text{and} \quad \sigma_{X+b} = \sigma_X$$

The absolute value is necessary because a might be negative, yet a standard deviation cannot be. Usually multiplication by a corresponds to a change in the unit of measurement (e.g., kg to lb or dollars to euros); the sd in the new unit is just the original sd multiplied by the conversion factor. On the other hand, the addition of the constant b does not affect the variance, which is intuitive, because the addition of b changes the location (mean value) but not the spread of values. Together, (3.12) and (3.13) comprise the *rescaling properties* of mean and standard deviation.

Example 3.25   In the computer sales scenario of Example 3.22, $E(X) = 2$ and

$$E(X^2) = (0^2)(.1) + (1^2)(.2) + (2^2)(.3) + (3^2)(.4) = 5$$

so $V(X) = 5 - (2)^2 = 1$. The profit function $Y = h(X) = 800X - 900$ is linear, so (3.13) applies with $a = 800$ and $b = -900$. Hence Y has variance $a^2 \sigma^2_X = (800)^2(1) = 640{,}000$ and standard deviation $\$800$. ∎

Exercises: Section 3.3 (28–45)

28. The pmf for X = the number of major defects on a randomly selected appliance of a certain type is

x	0	1	2	3	4
$p(x)$	.08	.15	.45	.27	.05

Compute the following:

a. $E(X)$
b. $V(X)$ directly from the definition

c.   The standard deviation of X
d. $V(X)$ using the shortcut formula

29. An individual who has automobile insurance from a company is randomly selected. Let Y be the number of moving violations for which the individual was cited during the last 3 years. The pmf of Y is

y	0	1	2	3
$p(y)$	.60	.25	.10	.05

   a. Compute $E(Y)$.

   b. Suppose an individual with Y violations incurs a surcharge of $\$100Y^2$. Calculate the expected amount of the surcharge.

30. Refer to Exercise 12 and calculate $V(Y)$ and σ_Y. Then determine the probability that Y is within 1 standard deviation of its mean value.

31. An appliance dealer sells three different models of upright freezers having 13.5, 15.9, and 19.1 cubic feet of storage space, respectively. Let X = the amount of storage space purchased by the next customer to buy a freezer. Suppose that X has pmf

x	13.5	15.9	19.1
$p(x)$	.2	.5	.3

   a. Compute $E(X)$, $E(X^2)$, and $V(X)$.

   b. If the price of a freezer having capacity X cubic feet is $17X + 180$, what is the expected price paid by the next customer to buy a freezer?

   c. What is the standard deviation of the price $17X + 180$ paid by the next customer?

   d. Suppose that although the rated capacity of a freezer is X, the actual capacity is $h(X) = X - .01X^2$. What is the expected actual capacity of the freezer purchased by the next customer?

32. Let X be a Bernoulli rv with pmf as in Example 3.17.

   a. Compute $E(X^2)$.

   b. Show that $V(X) = p(1 - p)$.

   c. Compute $E(X^{79})$.

33. Suppose that the number of plants of a particular type found in a rectangular region (called a quadrat by ecologists) in a certain geographic area is a rv X with pmf

$$p(x) = c/x^3 \quad \text{for } x = 1, 2, 3, \ldots$$

Is $E(X)$ finite? Justify your answer (this is another distribution that statisticians would call heavy-tailed).

34. A small market orders copies of a certain magazine for its magazine rack each week. Let X = demand for the magazine, with pmf

x	1	2	3	4	5	6
$p(x)$	1/15	2/15	3/15	4/15	3/15	2/15

Suppose the store owner actually pays $\$2.00$ for each copy of the magazine and the price to customers is $\$4.00$. If magazines left at the end of the week have no salvage value, is it better to order three or four copies of the magazine? [Hint: For both three and four copies ordered, express net revenue as a function of demand X, and then compute the expected revenue.]

35. Let X be the damage incurred (in \$) in a certain type of accident during a given year. Possible X values are 0, 1000, 5000, and 10,000, with probabilities .8, .1, .08, and .02, respectively. A particular company offers a $\$500$ deductible policy. If the company wishes its expected profit to be $\$100$, what premium amount should it charge?

36. The n candidates for a job have been ranked 1, 2, 3, …, n. Let X = the rank of a randomly selected candidate, so that X has pmf

$$p(x) = 1/n \quad x = 1, 2, 3, \ldots, n$$

(this is called the *discrete uniform distribution*). Compute $E(X)$ and $V(X)$ using the shortcut formula. [Hint: The sum of the first n positive integers is $n(n + 1)/2$, whereas the sum of their squares is given by $n(n + 1)(2n + 1)/6$.]

37. Let X = the outcome when a fair die is rolled once. If before the die is rolled you are offered either $\$100$ dollars or $h(X) = 350/X$ dollars, would you accept the guaranteed amount or would you gamble? [Hint: Determine $E[h(X)]$, but be careful: the mean of $350/X$ is not $350/\mu$.]

38. A supply company currently has in stock 500 lb of fertilizer, which it sells to customers in 10-pound bags. Let X equal the number of bags purchased by a randomly selected customer. Sales data shows that X has the following pmf:

x	1	2	3	4
$p(x)$	.2	.4	.3	.1

a. Compute the average number of bags bought per customer.
b. Determine the standard deviation for the number of bags bought per customer.
c. Define Y to be the amount of fertilizer left in stock, in pounds, after the first customer. Construct the pmf of Y.
d. Use the pmf of Y to find the expected amount of fertilizer left in stock, in pounds, after the first customer.
e. Write Y as a linear function of X. Then use rescaling properties to find the mean and standard deviation of Y.
f. The supply company offers a discount to each customer based on the formula $W = (X - 1)^2$. Determine the expected discount for a customer.
g. Does your answer to part (f) equal $(\mu_X - 1)^2$? Why or why not?
h. Calculate the standard deviation of W.

39. Refer back to the roulette scenario in Example 3.18. Two other ways to wager at roulette are betting on a single number, or on a four-number "square." The pmfs for the returns on a $1 wager on a number and a square are displayed below. (Payoffs for winning are always based on the odds of losing a wager under the assumption the two green spaces didn't exist.)

Single number:

x	–$1	+$35
$p(x)$	37/38	1/38

Square:

x	–$1	+$8
$p(x)$	34/38	4/38

a. Determine the expected return from a $1 wager on a single number, and then on a square.
b. Compare your answers from (a) to Example 3.18. What can be said about the expected return for a $1 wager? Based on this, does expected return reflect most players' intuition that betting on black is "safer" and betting on a single number is "riskier"?
c. Calculate the standard deviations for the two pmfs above as well as the pmf in Example 3.18.
d. How do the standard deviations of the three betting schemes (color, single number, square) compare? How do these values appear to relate to players' intuitive sense of risk?

40. In the popular game Plinko on *The Price Is Right*, contestants drop a circular disk (a "chip") down a pegged board; the chip bounces down the board and lands in a slot corresponding to one of five dollar mounts. The random variable X = winnings from one chip dropped from the middle slot has roughly the following distribution.

x	$0	$100	$500	$1000	$10,000
$p(x)$	0.39	0.03	0.11	0.24	0.23

a. Graph the probability mass function of X.
b. What is the probability a contestant makes money on a chip?
c. What is the probability a contestant makes at least $1000 on a chip?
d. Determine the expected winnings. Interpret this number.
e. Determine the corresponding standard deviation.

41. a. Draw a line graph of the pmf of X in Exercise 34. Then determine the pmf of $-X$ and draw its line graph. From these two pictures, what can you say about $V(X)$ and $V(-X)$?

  b. Use the proposition involving $V(aX + b)$ to establish a general relationship between $V(X)$ and $V(-X)$.

42. Use the definition of variance to prove that $V(aX + b) = a^2\sigma_X^2$. [*Hint:* With $Y = aX + b$, $E(Y) = a\mu + b$ where $\mu = E(X)$.]

43. Suppose $E(X) = 5$ and $E[X(X - 1)] = 27.5$. What is

  a. $E(X^2)$? [*Hint*: $E[X(X - 1)] = E[X^2 - X] = E(X^2) - E(X)$.]

  b. $V(X)$?

  c. The general relationship among the quantities $E(X)$, $E[X(X - 1)]$, and $V(X)$?

44. Write a general rule for $E(X - c)$ where c is a constant. What happens when you let $c = \mu$, the expected value of X?

45. A result called **Chebyshev's inequality** states that for any probability distribution of a rv X and any number k that is at least 1, $P(|X - \mu| \geq k\sigma) \leq 1/k^2$. In words, the probability that the value of X lies at least k standard deviations from its mean is at most $1/k^2$.

  a. What is the value of the upper bound for $k = 2$? $k = 3$? $k = 4$? $k = 5$? $k = 10$?

  b. Compute μ and σ for the distribution given in Exercise 13. Then evaluate $P(|X - \mu| \geq k\sigma)$ for the values of k given in part (a). What does this suggest about the upper bound relative to the corresponding probability?

  c. Let X have possible values, -1, 0, and 1, with probabilities 1/18, 8/9, and 1/18, respectively. What is $P(|X - \mu| \geq 3\sigma)$, and how does its value compare to the corresponding bound?

  d. Give a distribution for which $P(|X - \mu| \geq 5\sigma) = .04$.

3.4 Moments and Moment Generating Functions

The expected values of integer powers of X and $X - \mu$ are often referred to as *moments*, terminology borrowed from physics. In this section, we'll discuss the general topic of moments and develop a shortcut for computing them.

DEFINITION   The **kth moment** of a random variable X is $E(X^k)$, while the **kth moment about the mean** (or **kth central moment**) of X is $E[(X - \mu)^k]$, where $\mu = E(X)$.

For example, $\mu = E(X)$ is the "first moment" of X and corresponds to the center of mass of the distribution of X. Similarly, $V(X) = E[(X - \mu)^2]$ is the second moment of X about the mean, which is known in physics as the *moment of inertia*.

Example 3.26 A popular brand of dog food is sold in 5, 10, 15, and 20 lb bags. Let X be the weight of the next bag purchased, and suppose the pmf of X is

x	5	10	15	20
$p(x)$	.1	.2	.3	.4

The first moment of X is its mean:

$$\mu = E(X) = \sum_{x \in D} x p(x) = 5(.1) + 10(.2) + 15(.3) + 20(.4) = 15 \text{ lbs}$$

The second moment about the mean is the variance:

$$\sigma^2 = E[(X - \mu)^2] = \sum_{x \in D} (x - \mu)^2 p(x)$$
$$= (5 - 15)^2(.1) + (10 - 15)^2(.2) + (15 - 15)^2(.3) + (20 - 15)^2(.4) = 25,$$

for a standard deviation of 5 lbs. The third central moment of X is

$$E[(X - \mu)^3] = \sum_{x \in D} (x - \mu)^3 p(x)$$
$$= (5 - 15)^3(.1) + (10 - 15)^3(.2) + (15 - 15)^3(.3) + (20 - 15)^3(.4) = -75$$

We'll discuss an interpretation of this last number next.   ∎

It is not difficult to verify that the third moment about the mean is 0 if the pmf of X is symmetric. So, we would like to use $E[(X - \mu)^3]$ as a measure of lack of symmetry, but it depends on the scale of measurement. If we switch the unit of weight in Example 3.26 from pounds to ounces or kilograms, the value of the third moment about the mean (as well as the values of all the other moments) will change. But we can achieve scale independence by dividing the third moment about the mean by σ^3:

$$\frac{E[(X - \mu)^3]}{\sigma^3} = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] \tag{3.14}$$

Expression (3.14) is our measure of departure from symmetry, called the **skewness coefficient**. The skewness coefficient for a symmetric distribution is 0 because its third moment about the mean is 0. However, in the foregoing example the skewness coefficient is $E[(X - \mu)^3]/\sigma^3 = -75/5^3 = -0.6$. When the skewness coefficient is negative, as it is here, we say that the distribution is *negatively skewed* or that it is *skewed to the left*. Generally speaking, it means that the distribution stretches farther to the left of the mean than to the right.

If the skewness coefficient were positive, then we would say that the distribution is *positively skewed* or that it is *skewed to the right*. For example, reverse the order of the probabilities in the pmf of Example 3.26, so the probabilities of the values 5, 10, 15, 20 are now .4, .3, .2, and .1 (customers now favor much smaller bags of dog food). Exercise 61 shows that this changes the sign but not the magnitude of the skewness coefficient, so it becomes +0.6 and the distribution is skewed right. Both distributions are illustrated in Figure 3.8.
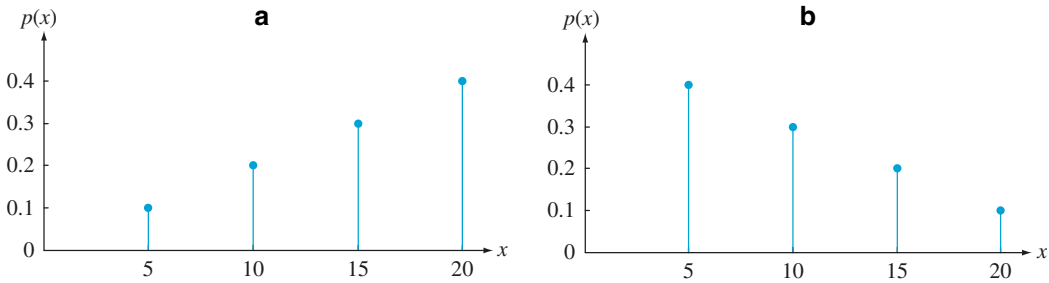
Figure 3.8   Departures from symmetry: (**a**) skewness coefficient < 0 (skewed left);
(**b**) skewness coefficient > 0 (skewed right)

The Moment Generating Function

Calculation of the mean, variance, skewness coefficient, etc., for a particular discrete rv requires extensive, sometimes tedious, summation. Mathematicians have developed a tool, the *moment generating function*, that will allow us to determine the moments of a distribution with less effort. Moreover, this function will allow us to derive properties of several important probability distributions in subsequent sections of the book.

Note first that $e^{1.7X}$ is a particular function of X; its expected value is $E(e^{1.7X}) = \sum e^{1.7x} \cdot p(x)$. The number 1.7 in the foregoing expression can be replaced by any other number—2.5, 179, –3.25, etc. Now consider replacing 1.7 by the letter t. Then the expected value depends on the numerical value of t; that is, $E(e^{tX})$ is a function of t. It is this function that will generate moments for us.

DEFINITION   The **moment generating function (mgf)** of a discrete random variable X is defined to be

$$M_X(t) = E(e^{tX}) = \sum_{x \in D} e^{tx} p(x)$$

where D is the set of possible X values. The moment generating function exists iff $M_X(t)$ is defined for an interval that includes zero as well as positive and negative values of t.

For any random variable X, the mgf evaluated at $t = 0$ is

$$M_X(0) = E(e^{0X}) = \sum_{x \in D} e^{0x} p(x) = \sum_{x \in D} 1 p(x) = 1$$

That is, $M_X(0)$ is the sum of all the probabilities, so it must always be 1. However, in order for the mgf to be useful in generating moments, it will need to be defined for an interval of values of t including 0 in its interior. The moment generating function fails to exist in cases when moments themselves fail to exist (see Example 3.30 below).

Example 3.27 The simplest example of an mgf is for a Bernoulli distribution, where only the X values 0 and 1 receive positive probability. Let X be a Bernoulli random variable with $p(0) = 1/3$ and $p(1) = 2/3$. Then

$$M_X(t) = E(e^{tX}) = \sum_{x \in D} e^{tx} p(x) = e^{t \cdot 0} \frac{1}{3} + e^{t \cdot 1} \frac{2}{3} = \frac{1}{3} + e^t \frac{2}{3}$$

A Bernoulli random variable will always have an mgf of the form $p(0) + p(1)e^t$, a well-defined function for all values of t. ∎

A key property of the mgf is its "uniqueness," the fact that it completely characterizes the underlying distribution.

MGF UNIQUENESS THEOREM	If the mgf exists and is the same for two distributions, then the two distributions are identical. That is, the moment generating function uniquely specifies the probability distribution; there is a one-to-one correspondence between distributions and mgfs.

The proof of this theorem, originally due to Laplace, requires some sophisticated mathematics and is beyond the scope of this textbook.

Example 3.28 Let X, the number of claims submitted on a renter's insurance policy in a given year, have mgf $M_X(t) = .7 + .2e^t + .1e^{2t}$. It follows that X must have the pmf $p(0) = .7$, $p(1) = .2$, and $p(2) = .1$—because if we use this pmf to obtain the mgf, we get $M_X(t)$, and the distribution is uniquely determined by its mgf. ∎

Example 3.29 Consider testing individuals' blood samples one by one in order to find someone whose blood type is Rh+. The rv X = the number of tested samples should follow the pmf specified in Example 3.10 with $p = .85$:

$$p(x) = .85(.15)^{x-1} \quad \text{for } x = 1, 2, 3, \ldots.$$

Determining the moment generating function here requires using the formula for the sum of a geometric series: $1 + r + r^2 + \cdots = 1/(1 - r)$ for $|r| < 1$. The moment generating function is

$$M_X(t) = E(e^{tX}) = \sum_{x \in D} e^{tx} p(x) = \sum_{x=1}^{\infty} e^{tx} .85(.15)^{x-1} = .85e^t \sum_{x=1}^{\infty} e^{t(x-1)} (.15)^{x-1}$$

$$= .85e^t \sum_{x=1}^{\infty} (.15e^t)^{x-1} = .85e^t [1 + .15e^t + (.15e^t)^2 + \cdots] = \frac{.85e^t}{1 - .15e^t}$$

The condition on r requires $|.15e^t| < 1$. Dividing by .15 and taking logs gives $t < -\ln(.15) \approx 1.90$; i.e., this function is defined in the interval $(-\infty, 1.90)$. The result is an interval of values that includes 0 in its interior, so the mgf exists. As a check, $M_X(0) = .85/(1 - .15) = 1$, as required. ∎

Example 3.30 Reconsider Example 3.20, where $p(x) = k/x^2$, $x = 1, 2, 3, \ldots$. Recall that $E(X)$ does not exist for this distribution, portending a problem for the existence of the mgf:

$$M_X(t) = E(e^{tX}) = \sum_{x=1}^{\infty} e^{tx} \frac{k}{x^2}$$

With the help of tests for convergence such as the ratio test, we find that the series converges if and only if $e^t \leq 1$, which means that $t \leq 0$; i.e., the mgf is only defined on the interval $(-\infty, 0]$. Because zero is on the *boundary* of this interval, not the interior of the interval (the interval must include both positive and negative values), the mgf of this distribution does not exist. In any case, it could not be useful for finding moments, because X does not have even a first moment (mean). ∎

Obtaining Moments from the MGF

For any positive integer r, let $M_X^{(r)}(t)$ denote the rth derivative of $M_X(t)$. By computing this and then setting $t = 0$, we get the rth moment about 0.

THEOREM   If the mgf of X exists, then $E(X^r)$ is finite for all positive integers r, and

$$E(X^r) = M_X^{(r)}(0) \tag{3.15}$$

Proof The proof of the finiteness of all moments is beyond the scope of this book. We will show that Expression (3.15) is true for $r = 1$ and $r = 2$. A proof by mathematical induction can be used for general r. The first derivative of the mgf is

$$\frac{d}{dt} M_X(t) = \frac{d}{dt} \sum_{x \in D} e^{xt} p(x) = \sum_{x \in D} \frac{\partial}{\partial t} e^{xt} p(x) = \sum_{x \in D} x e^{xt} p(x)$$

where we have interchanged the order of summation and differentiation. (This is justified inside the interval of convergence, which includes 0 in its interior.) Next set $t = 0$ to obtain the first moment:

$$M_X'(0) = M_X^{(1)}(0) = \sum_{x \in D} x e^{x(0)} p(x) = \sum_{x \in D} x p(x) = E(X)$$

Differentiating a second time gives

$$\frac{d^2}{dt^2} M_X(t) = \frac{d}{dt} \sum_{x \in D} x e^{xt} p(x) = \sum_{x \in D} x \frac{\partial}{\partial t} e^{xt} p(x) = \sum_{x \in D} x^2 e^{xt} p(x)$$

Set $t = 0$ to get the second moment:

$$M_X''(0) = M_X^{(2)}(0) = \sum_{x \in D} x^2 p(x) = E(X^2) \qquad ∎$$

For the pmfs in Examples 3.27 and 3.28, this may seem like needless work—after all, for simple distributions with just a few values, we can quickly determine the mean, variance, etc. The real utility of the mgf arises for more complicated distributions.

Example 3.31 (Example 3.29 continued) Recall that $p = .85$ is the probability of a person having Rh+ blood, and we keep checking people until we find one with this blood type. If X is the number of people we need to check, then $p(x) = .85(.15)^{x-1}$, $x = 1, 2, 3, \ldots$, and the mgf is

$$M_X(t) = E(e^{tX}) = \frac{.85e^t}{1 - .15e^t}$$

Differentiating with the help of the quotient rule,

$$M_X'(t) = \frac{.85e^t}{(1 - .15e^t)^2}$$

Setting $t = 0$ then gives $\mu = E(X) = M_X'(0) = 1/.85 = 1.176$. This corresponds to the formula $\mu = 1/p$ when .85 is replaced by p.

To get the second moment, differentiate again:

$$M_X''(t) = \frac{.85e^t(1 + .15e^t)}{(1 - .15e^t)^3}$$

Setting $t = 0$, $E(X^2) = M_X''(0) = 1.15/.85^2$. Now use the variance shortcut formula:

$$V(X) = \sigma^2 = E(X^2) - \mu^2 = 1.15/.85^2 - \left(\frac{1}{.85}\right)^2 = \frac{.15}{(.85)^2} = .2076 \qquad \blacksquare$$

There is an alternate way of doing the differentiation that can sometimes make the effort easier. Define $R_X(t) = \ln[M_X(t)]$, where $\ln(u)$ is the natural log of u. In Exercise 54 you are requested to verify that if the moment generating function exists,

$$\mu = E(X) = R_X'(0)$$
$$\sigma^2 = V(X) = R_X''(0)$$

Example 3.32 Here we apply $R_X(t)$ to Example 3.31. Using properties of logarithms,

$$R_X(t) = \ln[M_X(t)] = \ln\left(\frac{.85e^t}{1 - .15e^t}\right) = \ln(.85) + t - \ln(1 - .15e^t)$$

The first derivative is

$$R_X'(t) = 0 + 1 - \frac{1}{1 - .15e^t}(-.15e^t) = 1 + \frac{.15e^t}{1 - .15e^t} = \frac{1}{1 - .15e^t}$$

and the second derivative is

$$R_X''(t) = \frac{.15e^t}{(1 - .15e^t)^2}$$

Setting t to 0 gives

$$\mu = E(X) = R_X'(0) = \frac{1}{.85}$$
$$\sigma^2 = V(X) = R_X''(0) = \frac{.15}{(.85)^2}$$

These are in agreement with the results of Example 3.31.   ■

As mentioned in Section 3.3, it is common to transform a rv X using a linear function $Y = aX + b$. What happens to the mgf when we do this?

PROPOSITION   Let X have the mgf $M_X(t)$ and let $Y = aX + b$. Then $M_Y(t) = e^{bt}M_X(at)$.

Example 3.33 Let X be a Bernoulli random variable with $p(0) = 20/38$ and $p(1) = 18/38$. Think of X as the number of wins, 0 or 1, in a single play of roulette. If you play roulette at an American casino and bet on red, then your chance of winning is 18/38 because 18 of the 38 possible outcomes are red. From Example 3.27, $M_X(t) = 20/38 + e^t(18/38)$. Suppose you bet \$5 on red, and let Y be your winnings. If $X = 0$ then $Y = -5$, and if $X = 1$ then $Y = + 5$. The linear equation $Y = 10X - 5$ gives the appropriate relationship.

This equation is of the form $Y = aX + b$ with $a = 10$ and $b = -5$, so by the foregoing proposition

$$M_Y(t) = e^{bt}M_X(at) = e^{-5t}M_X(10t)$$
$$= e^{-5t}\left[\frac{20}{38} + e^{10t}\frac{18}{38}\right] = e^{-5t} \cdot \frac{20}{38} + e^{5t} \cdot \frac{18}{38}$$

This implies that the pmf of Y is $p(-5) = 20/38$ and $p(5) = 18/38$; moreover, we can compute the mean (and other moments) of Y directly from this mgf.   ■

Exercises: Section 3.4 (46–61)

46. Let X be the number of pumps in use at a gas station, and suppose X has the distribution given by the accompanying table. Determine $M_X(t)$ and use it to find $E(X)$ and $V(X)$.

x	0	1	2	3	4	5	6
$p(x)$	.04	.20	.34	.20	.15	.04	.03

47. In flipping a fair coin let X be the number of tosses to get the first head. Then $p(x) = .5^x$ for $x = 1, 2, 3, \ldots$. Determine $M_X(t)$ and use it to get $E(X)$ and $V(X)$.

48. If you toss a fair die with outcome X, $p(x) = \frac{1}{6}$ for $x = 1, 2, 3, 4, 5, 6$. Find $M_X(t)$.

49. For the entry-level employees of a certain fast food chain, the pmf of X = highest grade level completed is specified by $p(9) = .01$, $p(10) = .05$, $p(11) = .16$, and $p(12) = .78$.

    a. Determine the moment generating function of this distribution.
    b. Use (a) to determine the mean and variance of this distribution.

50. Calculate the skewness coefficient for each of the distributions in the previous four exercises. Do those agree with the "shape" of each distribution?

51. Given $M_X(t) = .2 + .3e^t + .5e^{3t}$, find $p(x)$, $E(X)$, $V(X)$.

52. If $M_X(t) = 1/(1 - t^2)$, find $E(X)$ and $V(X)$ by differentiating $M_X(t)$.

53. Show that $g(t) = te^t$ cannot be a moment generating function.

54. Let $M_X(t)$ be the moment generating function of a rv X, and define $R_X(t) = \ln[M_X(t)]$. Show that

    a. $R_X(0) = 0$
    b. $R'_X(0) = \mu_X$
    c. $R''_X(0) = \sigma_X^2$

55. If $M_X(t) = e^{5t + 2t^2}$ then find $E(X)$ and $V(X)$ by differentiating

    a. $M_X(t)$
    b. $R_X(t)$

56. If $M_X(t) = e^{5(e^t - 1)}$ then find $E(X)$ and $V(X)$ by differentiating

    a. $M_X(t)$
    b. $R_X(t)$

57. Using a calculation similar to the one in Example 3.29 show that, if X has the distribution of Example 3.10, then its mgf is

    $$M_X(t) = \frac{pe^t}{1 - (1 - p)e^t}$$

    If Y has mgf $M_Y(t) = .75e^t/(1 - .25e^t)$, determine the probability mass function $p_Y(y)$ with the help of the uniqueness property.

58. Let X have the moment generating function of Example 3.29 and let $Y = X - 1$. Recall that X is the number of people who need to be checked to get someone who is Rh+, so Y is the number of people checked *before* the first Rh+ person is found. Find $M_Y(t)$ using the last proposition in this section.

59. Let $M_X(t) = e^{5t + 2t^2}$ and let $Y = (X - 5)/2$. Find $M_Y(t)$ and use it to find $E(Y)$ and $V(Y)$.

60. a. Prove the result in the last proposition of this section: $M_{aX+b}(t) = e^{bt}M_X(at)$.
    b. Let $Y = aX + b$. Use (a) to establish the relationships between the means and variances of X and Y.

61. Let X be the number of points earned by a randomly selected student on a 10 point quiz, with possible values 0, 1, 2, …, 10 and pmf $p(x)$, and suppose the distribution has a skewness of c. Now consider reversing the probabilities in the distribution, so that $p(0)$ is interchanged with $p(10)$, $p(1)$ is interchanged with $p(9)$, and so on. Show that the skewness of the resulting distribution is $-c$. [*Hint*: Let $Y = 10 - X$ and show that Y has the reversed distribution. Use this fact to determine μ_Y and then the value of skewness for the Y distribution.]

3.5 The Binomial Probability Distribution

Many experiments conform either exactly or approximately to the following list of requirements:

1. The experiment consists of a sequence of n smaller experiments called *trials*, where n is fixed in advance of the experiment.
2. Each trial can result in one of the same two possible outcomes (dichotomous trials), which we denote by success (S) or failure (F).
3. The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.
4. The probability of success is constant from trial to trial (homogeneous trials); we denote this probability by p.

DEFINITION   An experiment for which Conditions 1–4 are satisfied—a fixed number of dichotomous, independent, homogeneous trials—is called **a binomial experiment**.

Example 3.34 The same coin is tossed successively and independently n times. We arbitrarily use S to denote the outcome H (heads) and F to denote the outcome T (tails). Then this experiment satisfies Conditions 1–4. Wagering on n spins of a roulette wheel, with S = win money and F = lose money, also results in a binomial experiment so long as you bet the same way every time (e.g., always on black, so that $P(S)$ remains constant across different spins). Another binomial experiment was

alluded to in Example 3.10: there, a binomial experiment would consist of sending a fixed number n of messages across a communication channel, with S = message received correctly and F = received message contains errors.   ■

Some experiments involve a sequence of independent trials for which there are more than two possible outcomes on any one trial. A binomial experiment can then be created by dividing the possible outcomes into two groups.

Example 3.35 The color of pea seeds is determined by a single genetic locus. If the two alleles at this locus are AA or Aa (the genotype), then the pea will be yellow (the phenotype), and if the allele is aa, the pea will be green. Suppose we pair off 20 Aa seeds and cross the two seeds in each of the ten pairs to obtain ten new genotypes. Call each new genotype a success S if it is aa and a failure otherwise. Then with this identification of S and F, the experiment is binomial with $n = 10$ and $p = P(\text{aa genotype})$. If each member of the pair is equally likely to contribute either a or A, then $p = P(\text{a}) \cdot P(\text{a}) = (1/2)(1/2) = .25$.   ■

Example 3.36 A student has an iPod playlist containing 50 songs, of which 35 were recorded prior to the year 2018 and the other 15 were recorded more recently. Suppose the shuffle function is used to select five from among these 50 songs for listening during a walk between classes. Each selection of a song constitutes a trial; regard a trial as a success if the selected song was recorded before 2018. Then

$$P(S \text{ on first trial}) = \frac{35}{50} = .70$$

and

$$
\begin{aligned}
P(S \text{ on second trial}) &= P(SS) + P(FS) \\
&= P(\text{second } S|\text{first } S)P(\text{first } S) + P(\text{second } S|\text{first } F)P(\text{first } F) \\
&= \frac{34}{49} \cdot \frac{35}{50} + \frac{35}{49} \cdot \frac{15}{50} = \frac{35}{50}\left(\frac{34}{49} + \frac{15}{49}\right) = \frac{35}{50} = .70
\end{aligned}
$$

Similarly, it can be shown that $P(S \text{ on } i\text{th trial}) = .70$ for $i = 3, 4, 5$, so the trials are homogeneous. However,

$$P(S \text{ on fifth trial} \mid SSSS) = \frac{31}{46} = .67$$

whereas

$$P(S \text{ on fifth trial} \mid FFFF) = \frac{35}{46} = .76$$

The experiment is *not* binomial because the trials are not independent. In general, if sampling is without replacement, the experiment will not yield independent trials. If songs had been selected *with* replacement, then trials would have been independent, but this might have resulted in the same song being listened to more than once.   ■

Example 3.37 Suppose a state has 500,000 licensed drivers, of whom 400,000 are insured. A sample of ten drivers is chosen without replacement. The ith trial is labeled S if the ith driver chosen is insured. Although this situation would seem identical to that of Example 3.36, the important

difference is that the size of the population being sampled is very large relative to the sample size. In this case

$$P(S \text{ on } 2|S \text{ on } 1) = \frac{399,999}{499,999} = .7999996 \approx .8$$

and

$$P(S \text{ on } 10|S \text{ on first } 9) = \frac{399,991}{499,991} = .7999964 \approx .8$$

These calculations suggest that although the trials are not exactly independent, the conditional probabilities differ so slightly from one another that for practical purposes the trials can be regarded as independent with constant $P(S) = .8$. Thus, to a very good approximation, the experiment is binomial with $n = 10$ and $p = .8$.   ∎

We will use the following convention in deciding whether a "without-replacement" experiment can be treated as a binomial experiment.

RULE   Consider sampling without replacement from a dichotomous population of size N. If the sample size (number of trials) n is at most 5% of the population size, the experiment can be analyzed as though it were exactly a binomial experiment.

By "analyzed," we mean that probabilities based on the binomial experiment assumptions will be quite close to the actual "without-replacement" probabilities, which are typically more difficult to calculate. In Example 3.36, $n/N = 5/50 = .1 > .05$, so the binomial experiment is not a good approximation, but in Example 3.37, $n/N = 10/500,000 < .05$.

The Binomial Random Variable and Distribution

In most binomial experiments, it is the total number of successes, rather than knowledge of exactly which trials yielded successes, that is of interest.

DEFINITION   Given a binomial experiment consisting of n trials, the **binomial random variable X** associated with this experiment is defined as

X = the number of successes among the n trials

Suppose, for example, that $n = 3$. Then there are eight possible outcomes for the experiment:

$$\mathcal{S} = \{SSS, SSF, SFS, SFF, FSS, FSF, FFS, FFF\}$$

From the definition of X, $X(SSF) = 2$, $X(SFF) = 1$, and so on. Possible values for X in an n-trial experiment are $x = 0, 1, 2, \ldots, n$.

NOTATION   We will write $X \sim \text{Bin}(n, p)$ to indicate that X is a binomial rv based on n trials with success probability p. Because the pmf of a binomial rv X depends on the two parameters n and p, we denote the pmf by $b(x; n, p)$.

Our next goal is to derive a formula for the binomial pmf. Consider first the case $n = 4$ for which each outcome, its probability, and corresponding x value are listed in Table 3.1. For example,

Table 3.1 Outcomes and probabilities for a binomial experiment with four trials

Outcome	x	Probability	Outcome	x	Probability
SSSS	4	p^4	FSSS	3	$p^3(1-p)$
SSSF	3	$p^3(1-p)$	FSSF	2	$p^2(1-p)^2$
SSFS	3	$p^3(1-p)$	FSFS	2	$p^2(1-p)^2$
SSFF	2	$p^2(1-p)^2$	FSFF	1	$p(1-p)^3$
SFSS	3	$p^3(1-p)$	FFSS	2	$p^2(1-p)^2$
SFSF	2	$p^2(1-p)^2$	FFSF	1	$p(1-p)^3$
SFFS	2	$p^2(1-p)^2$	FFFS	1	$p(1-p)^3$
SFFF	1	$p(1-p)^3$	FFFF	0	$(1-p)^4$

$$
\begin{aligned}
P(SSFS) &= P(S) \cdot P(S) \cdot P(F) \cdot P(S) \quad &&\text{independent trials} \\
&= p \cdot p \cdot (1-p) \cdot p \quad &&\text{constant } P(S) \\
&= p^3 \cdot (1-p)
\end{aligned}
$$

In this special case, we wish $b(x; 4, p)$ for $x = 0, 1, 2, 3,$ and 4. For $b(3; 4, p)$, we identify which of the 16 outcomes yield an x value of 3 and sum the probabilities associated with each such outcome:

$$
b(3; 4, \ p) = P(FSSS) + P(SFSS) + P(SSFS) + P(SSSF) = 4p^3(1-p)
$$

There are four outcomes with $x = 3$ and each has probability $p^3(1-p)$ (the probability depends only on the number of S's, *not* the order of S's and F's), so

$$
b(3; 4, p) = \left\{ \begin{array}{c} \text{number of outcomes} \\ \text{with } X = 3 \end{array} \right\} \cdot \left\{ \begin{array}{c} \text{probability of any particular} \\ \text{outcome with } X = 3 \end{array} \right\}
$$

Similarly, $b(2; 4, p) = 6p^2(1-p)^2$, which is also the product of the number of outcomes with $X = 2$ and the probability of any such outcome.

In general,

$$
b(x; n, p) = \left\{ \begin{array}{c} \text{number of sequences of} \\ \text{length } n \text{ consisting of } x \, S's \end{array} \right\} \cdot \left\{ \begin{array}{c} \text{probability of any} \\ \text{particular such sequence} \end{array} \right\}
$$

Since the ordering of S's and F's is not important, the second factor in braces is $p^x(1-p)^{n-x}$ (e.g., the first x trials resulting in S and the last $n - x$ resulting in F). The first factor is the number of ways of choosing x of the n trials to be S's—that is, the number of combinations of size x that can be constructed from n distinct objects (trials here).

THEOREM

$$
b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, 2, \ldots, n
$$

Example 3.38 Each of six randomly selected cola drinkers is given a glass containing cola S and one containing cola F. The glasses are identical in appearance except for a code on the bottom to identify the cola. Suppose there is no tendency among cola drinkers to prefer one cola to the other.

Then $p = P$(a selected individual prefers S) = .5, so with X = the number among the six who prefer S, $X \sim \text{Bin}(6, .5)$.

Thus

$$P(X = 3) = b(3; 6, .5) = \binom{6}{3}(.5)^3(.5)^3 = 20(.5)^6 = .313$$

The probability that at least three prefer S is

$$P(3 \leq X) = \sum_{x=3}^{6} b(x; 6, .5) = \sum_{x=3}^{6} \binom{6}{x}(.5)^x(.5)^{6-x} = .656$$

and the probability that at most one prefers S is

$$P(X \leq 1) = \sum_{x=0}^{1} b(x; 6, .5) = .109 \qquad \blacksquare$$

Computing Binomial Probabilities

Even for a relatively small value of n, the computation of binomial probabilities can be tedious. Software and statistical tables are both available for this purpose; both are often in terms of the cdf $F(x) = P(X \leq x)$ of the distribution, either in lieu of or in addition to the pmf. Various other probabilities can then be calculated using the proposition on cdfs from Section 3.2.

NOTATION   For $X \sim \text{Bin}(n, p)$, the cdf will be denoted by

$$B(x; n, p) = P(X \leq x) = \sum_{y=0}^{x} b(y; n, p) \quad x = 0, 1, \ldots, n$$

Many software packages, including R, have built-in functions to evaluate both the pmf and cdf of the binomial distribution (and many other named distributions). Table 3.2 provides the code for performing binomial calculations in R. In addition, Appendix Table A.1 shows the binomial cdf for $n = 5, 10, 15, 20, 25$ in combination with selected values of p.

Table 3.2 Binomial probability calculations in R

Function:	pmf	cdf
Notation:	$b(x; n, p)$	$B(x; n, p)$
R:	`dbinom(x, n, p)`	`pbinom(x, n, p)`

Example 3.39 Suppose that 20% of all copies of a particular textbook fail a binding strength test. Let X denote the number among 15 randomly selected copies that fail the test. Then X has a binomial distribution with $n = 15$ and $p = .2$: $X \sim \text{Bin}(15, .2)$.

(a) The probability that at most 8 fail the test is

$$P(X \leq 8) = \sum_{y=0}^{8} b(y; 15, .2) = B(8; 15, .2)$$

This is found at the intersection of the $p = .2$ column and $x = 8$ row in the $n = 15$ part of Table A.1: $B(8; 15, .2) = .999$. In R, we may type `pbinom(8,15,.2)`.

(b) The probability that exactly 8 fail is $P(X = 8) = b(8; 15, .2) = \binom{15}{8}(.2)^8(.8)^7 = .0034$. We can evaluate this probability in R with the call `dbinom(8,15,.2)`. To use Table A.1, write

$$P(X = 8) = P(X \leq 8) - P(X \leq 7) = B(8; 15, .2) - B(7; 15, .2)$$

which is the difference between two consecutive entries in the $p = .2$ column. The result is $.999 - .996 = .003$.

(c) The probability that at least 8 fail is $P(X \geq 8) = 1 - P(X \leq 7) = 1 - B(7; 15, .2)$. The cdf may be evaluated using R as above, or by looking up the entry in the $x = 7$ row of the $p = .2$ column in Table A.1. In any case, we find $P(X \geq 8) = 1 - .996 = .004$.

(d) Finally, the probability that between 4 and 7, inclusive, fail is

$$P(4 \leq X \leq 7) = P(X = 4, 5, 6, \text{ or } 7) = P(X \leq 7) - P(X \leq 3)$$
$$= B(7; 15, .2) - B(3; 15, .2) = .996 - .648 = .348$$

Notice that this latter probability is the difference between the cdf values at $x = 7$ and $x = 3$, *not $x = 7$ and $x = 4$.* ∎

Example 3.40 An electronics manufacturer claims that at most 10% of its power supply units need service during the warranty period. To investigate this claim, technicians at a testing laboratory purchase 20 units and subject each one to accelerated testing to simulate use during the warranty period. Let p denote the probability that a power supply unit needs repair during the period (the proportion of all such units that need repair). The laboratory technicians must decide whether the data resulting from the experiment supports the claim that $p \leq .10$. Let X denote the number among the 20 sampled that need repair, so $X \sim \text{Bin}(20, p)$. Consider the following decision rule:

Reject the claim that $p \leq .10$ in favor of the conclusion that $p > .10$ if $x \geq 5$ (where x is the observed value of X), and consider the claim plausible if $x \leq 4$.

The probability that the claim is rejected when $p = .10$ (an incorrect conclusion) is

$$P(X \geq 5 \text{ when } p = .10) = 1 - B(4; 20, .1) = 1 - .957 = .043$$

The probability that the claim is not rejected when $p = .20$ (a different type of incorrect conclusion) is

$$P(X \leq 4 \text{ when } p = .2) = B(4; 20, .2) = .630$$

The first probability is rather small, but the second is intolerably large. When $p = .20$, so that the manufacturer has grossly understated the percentage of units that need service, and the stated decision rule is used, 63% of all samples will result in the manufacturer's claim being judged plausible!

One might think that the probability of this second type of erroneous conclusion could be made smaller by changing the cutoff value 5 in the decision rule to something else. However, although replacing 5 by a smaller number would yield a probability smaller than .630, the other probability would then increase. The only way to make both "error probabilities" small is to base the decision rule on an experiment involving many more units (i.e., to increase n).   ∎

The Mean and Variance of a Binomial Random Variable

For $n = 1$, the binomial distribution becomes the Bernoulli distribution. From Example 3.17, the mean value of a Bernoulli variable is $\mu = p$, so the expected number of S's on any single trial is p. Since a binomial experiment consists of n trials, intuition suggests that for $X \sim \text{Bin}(n, p)$, $E(X) = np$, the product of the number of trials and the probability of success on a single trial. The expression for $V(X)$ is not so intuitive.

PROPOSITION | If $X \sim \text{Bin}(n, p)$, then $E(X) = np$, $V(X) = np(1 - p) = npq$, and $\text{SD}(X) = \sqrt{npq}$ (where $q = 1 - p$).

Thus, calculating the mean and variance of a binomial rv does not necessitate evaluating summations of the sort we employed in Section 3.3. The proof of the result for $E(X)$ is sketched in Exercise 86, and both the mean and the variance are obtained below using the moment generating function.

Example 3.41 If 75% of all purchases at a store are made with a credit card and X is the number among ten randomly selected purchases made with a credit card, then $X \sim \text{Bin}(10, .75)$. Thus $E(X) = np = (10)(.75) = 7.5$, $V(X) = npq = 10(.75)(.25) = 1.875$, and $\sigma = \sqrt{1.875} = 1.37$. Again, even though X can take on only integer values, $E(X)$ need not be an integer. If we perform a large number of independent binomial experiments, each with $n = 10$ trials and $p = .75$, then the average number of S's per experiment will be close to 7.5.   ∎

An important application of the binomial distribution is to estimating the precision of simulated probabilities, as in Section 2.6. The relative frequency definition of probability justified defining an estimate of a probability $P(A)$ by $\hat{P}(A) = X/n$, where n is the number of runs of the simulation program and X equals the number of runs in which event A occurred. Assuming the runs of our simulation are independent (and they usually are), the rv X has a binomial distribution with parameters n and $p = P(A)$. From the preceding proposition and the rescaling properties of mean and standard deviation, we have

$$E(\hat{P}(A)) = E\left(\frac{1}{n}X\right) = \frac{1}{n} \cdot E(X) = \frac{1}{n}(np) = p = P(A)$$

Thus we expect the value of our estimate to coincide with the probability being estimated, in the sense that there is no reason for $\hat{P}(A)$ to be systematically higher or lower than $P(A)$. Also,

$$\text{SD}(\hat{P}(A)) = \text{SD}\left(\frac{1}{n}X\right) = \left|\frac{1}{n}\right| \cdot \text{SD}(X) = \frac{1}{n}\sqrt{np(1-p)} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{P(A)[1 - P(A)]}{n}} \quad (3.16)$$

Expression (3.16) is called the **standard error** of $\hat{P}(A)$ (essentially a synonym for standard deviation) and indicates the amount by which an estimate $\hat{P}(A)$ "typically" varies from the true probability $P(A)$. However, this expression isn't of much use in practice: we most often simulate a probability when

$P(A)$ is unknown, which prevents us from using (3.16). As a solution, we simply substitute the estimate $\hat{P} = \hat{P}(A)$ into this expression and get

$$\text{SD}(\hat{P}(A)) \approx \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}$$

This is the estimated standard error formula (2.9) given in Section 2.6. Very importantly, this estimated standard error gets closer to 0 as the number of runs, n, in the simulation increases.

The Moment Generating Function of a Binomial Random Variable

Determining the mgf of a binomial rv relies on the binomial theorem, which states that $(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$. Using the definition,

$$M_X(t) = E(e^{tX}) = \sum_{x \in D} e^{tx} p(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = (pe^t + 1 - p)^n$$

Notice that the mgf satisfies the property $M_X(0) = 1$ required of all moment generating functions. The mean and variance can be obtained by differentiating $M_X(t)$:

$$M_X'(t) = n(pe^t + 1 - p)^{n-1} pe^t \quad \text{and} \quad \mu = M_X'(0) = np$$

Then the second derivative is

$$M_X''(t) = n(n-1)(pe^t + 1 - p)^{n-2} pe^t pe^t + n(pe^t + 1 - p)^{n-1} pe^t$$

and

$$E(X^2) = M_X''(0) = n(n-1)p^2 + np$$

Therefore,

$$\sigma^2 = V(X) = E(X^2) - [E(X)]^2$$
$$= n(n-1)p^2 + np - n^2 p^2 = np - np^2 = np(1-p)$$

in accord with the foregoing proposition.

Exercises: Section 3.5 (62–88)

62. Determine whether each of the following rvs has a binomial distribution. If it does, identify the values of the parameters n and p (if possible).

    a. X = the number of 4 s in 10 rolls of a fair die

    b. X = the number of multiple-choice questions a student gets right on a 40-question test, when each question has four choices and the student is completely guessing

    c. X = the same as (b), but half the questions have four choices and the other half have three

    d. X = the number of women in a random sample of 8 students, from a class comprised of 20 women and 15 men

e. $X =$ the total weight of 15 randomly selected apples

f. $X =$ the number of apples, out of a random sample of 15, that weigh more than 150 grams

63. Compute the following binomial probabilities directly from the formula for $b(x; n, p)$:

a. $b(3; 8, .6)$

b. $b(5; 8, .6)$

c. $P(3 \leq X \leq 5)$ when $n = 8$ and $p = .6$

d. $P(1 \leq X)$ when $n = 12$ and $p = .1$

64. Use Appendix Table A.1 or software to obtain the following probabilities:

a. $B(4; 10, .3)$

b. $b(4; 10, .3)$

c. $b(6; 10, .7)$

d. $P(2 \leq X \leq 4)$ when $X \sim$ Bin(10, .3)

e. $P(2 \leq X)$ when $X \sim$ Bin(10, .3)

f. $P(X \leq 1)$ when $X \sim$ Bin(10, .7)

g. $P(2 < X < 6)$ when $X \sim$ Bin(10, .3)

65. When circuit boards used in the manufacture of DVD players are tested, the long-run percentage of defectives is 5%. Let $X =$ the number of defective boards in a random sample of size $n = 25$, so $X \sim$ Bin(25, .05).

a. Determine $P(X \leq 2)$.

b. Determine $P(X \geq 5)$.

c. Determine $P(1 \leq X \leq 4)$.

d. What is the probability that none of the 25 boards is defective?

e. Calculate the expected value and standard deviation of X.

66. A company that produces fine crystal knows from experience that 10% of its goblets have cosmetic flaws and must be classified as "seconds."

a. Among six randomly selected goblets, how likely is it that only one is a second?

b. Among six randomly selected goblets, what is the probability that at least two are seconds?

c. If goblets are examined one by one, what is the probability that at most five must be selected to find four that are not seconds?

67. Suppose that only 25% of all drivers come to a complete stop at an intersection having flashing red lights in all directions when no other cars are visible. What is the probability that, of 20 randomly chosen drivers coming to an intersection under these conditions,

a. At most 6 will come to a complete stop?

b. Exactly 6 will come to a complete stop?

c. At least 6 will come to a complete stop?

68. Refer to the previous exercise.

a. What is the expected number of drivers among the 20 that come to a complete stop?

b. What is the standard deviation of the number of drivers among the 20 that come to a complete stop?

c. What is the probability that the number of drivers among these 20 that come to a complete stop differs from the expected number by more than 2 standard deviations?

69. Exercise 29 (Section 3.3) gave the pmf of Y, the number of traffic citations for a randomly selected individual insured by a company. What is the probability that among 15 randomly chosen such individuals

a. At least 10 have no citations?

b. Fewer than half have at least one citation?

c. The number that have at least one citation is between 5 and 10, inclusive?[3]

70. A particular type of tennis racket comes in a midsize version and an oversize version. Sixty percent of all customers at a store want the oversize version.

a. Among ten randomly selected customers who want this type of racket,

---

[3]"Between a and b, inclusive" is equivalent to $(a \leq X \leq b)$.

what is the probability that at least six want the oversize version?

b. Among ten randomly selected customers, what is the probability that the number who want the oversize version is within 1 standard deviation of the mean value?

c. The store currently has seven rackets of each version. What is the probability that all of the next ten customers who want this racket can get the version they want from current stock?

71. Twenty percent of all telephones of a certain type are submitted for service while under warranty. Of these, 60% can be repaired, whereas the other 40% must be replaced with new units. If a company purchases ten of these telephones, what is the probability that exactly two will end up being replaced under warranty?

72. A March 29, 2019, *Washington Post* article reported that (roughly) 5% of all students taking the ACT were granted extra time. Assume that 5% figure is exact, and consider a random sample of 25 students who have recently taken the ACT.

a. What is the probability that exactly 1 was granted extra time?

b. What is the probability that at least 1 was granted extra time?

c. What is the probability that at least 2 were granted extra time?

d. What is the probability that the number among the 25 who were granted extra time is within 2 standard deviations of the number you would expect?

e. Suppose that a student who does not receive extra time is allowed 3 h for the exam, whereas an accommodated student is allowed 4.5 h. What would you expect the average time allowed the 25 selected students to be?

73. Suppose that 90% of all batteries from a supplier have acceptable voltages. A certain type of flashlight requires two type-D batteries, and the flashlight will work only if both its batteries have acceptable voltages. Among ten randomly selected flashlights, what is the probability that at least nine will work? What assumptions did you make in the course of answering the question posed?

74. A *k-out-of-n system* functions provided that at least k of the n components function. Consider independently operating components, each of which functions (for the needed duration) with probability .96.

a. In a 3-component system, what is the probability that exactly two components function?

b. What is the probability a 2-out-of-3 system works?

c. What is the probability a 3-out-of-5 system works?

d. What is the probability a 4-out-of-5 system works?

e. What does the component probability (previously .96) need to equal so that the 4-out-of-5 system will function with probability at least .9999?

75. Bit transmission errors between computers sometimes occur, where one computer sends a 0 but the other computer receives a 1 (or vice versa). Because of this, the computer sending a message repeats each bit three times, so a 0 is sent as 000 and a 1 as 111. The receiving computer "decodes" each triplet by majority rule: whichever number, 0 or 1, appears more often in a triplet is declared to be the intended bit. For example, both 000 and 100 are decoded as 0, while 101 and 011 are decoded as 1. Suppose that 6% of bits are switched (0 to 1, or 1 to 0) during transmission between two particular computers, and that these errors occur independently during transmission.

a. Find the probability that a triplet is decoded incorrectly by the receiving computer.

b. Using your answer to part (a), explain how using triplets reduces communication errors.

c. How does your answer to part (a) change if each bit is repeated five times (instead of three)?

d. Imagine a 25 kilobit message (i.e., one requiring 25,000 bits to send). What is the expected number of errors if there is no bit repetition implemented? If each bit is repeated three times?

76. A very large batch of components has arrived at a distributor. The batch can be characterized as acceptable only if the proportion of defective components is at most .10. The distributor decides to randomly select 10 components and to accept the batch only if the number of defective components in the sample is at most 2.

a. What is the probability that the batch will be accepted when the actual proportion of defectives is .01? .05? .10? .20? .25?

b. Let p denote the actual proportion of defectives in the batch. A graph of P(batch is accepted) as a function of p, with p on the horizontal axis and P(batch is accepted) on the vertical axis, is called the *operating characteristic curve* for the acceptance sampling plan. Use the results of part (a) to sketch this curve for $0 \leq p \leq 1$.

c. Repeat parts (a) and (b) with "1" replacing "2" in the acceptance sampling plan.

d. Repeat parts (a) and (b) with "15" replacing "10" in the acceptance sampling plan.

e. Which of the three sampling plans, that of part (a), (c), or (d), appears most satisfactory, and why?

77. An ordinance requiring that a smoke detector be installed in all previously constructed houses has been in effect in a city for 1 year. The fire department is concerned that many houses remain without detectors. Let p = the true proportion of such houses having detectors, and suppose that a random sample of 25 homes is inspected. If the sample strongly indicates that fewer than 80% of all houses have a detector, the fire department will campaign for a mandatory inspection program. Because of the costliness of the program, the department prefers not to call for such inspections unless sample evidence strongly argues for their necessity. Let X denote the number of homes with detectors among the 25 sampled. Consider rejecting the claim that $p \geq .8$ if $x \leq 15$.

a. What is the probability that the claim is rejected when the actual value of p is .8?

b. What is the probability of not rejecting the claim when $p = .7$? When $p = .6$?

c. How do the "error probabilities" of parts (a) and (b) change if the value 15 in the decision rule is replaced by 14?

78. A toll bridge charges $1.00 for passenger cars and $2.50 for other vehicles. Suppose that during daytime hours, 60% of all vehicles are passenger cars. If 25 vehicles cross the bridge during a particular daytime period, what is the resulting expected toll revenue? [*Hint*: Let X = the number of passenger cars; then the toll revenue $h(X)$ is a linear function of X.]

79. A student who is trying to write a paper for a course has a choice of two topics, A and B. If topic A is chosen, the student will order two books through interlibrary loan, whereas if topic B is chosen, the student will order four books. The student believes that a good paper necessitates receiving and using at least half the books ordered for

either topic chosen. If the probability that a book ordered through interlibrary loan actually arrives in time is .9 and books arrive independently of one another, which topic should the student choose to maximize the probability of writing a good paper? What if the arrival probability is only .5 instead of .9?

80. Twelve jurors are randomly selected from a large population. At least in theory, each juror arrives at a conclusion about the case before the jury independently of the other jurors.

   a. In a criminal case, all 12 jurors must agree on a verdict. Let p denote the probability that a randomly selected member of the population would reach a guilty verdict based on the evidence presented (so a proportion $1 - p$ would reach "not guilty"). What is the probability, in terms of p, that the jury reaches a unanimous verdict one way or the other?

   b. For what values of p is the probability in part (a) the highest? For what value of p is the probability in (a) the lowest? Explain why this makes sense.

   c. In most civil cases, only a nine-person majority is required to decide a verdict. That is, if nine or more jurors favor the plaintiff, then the plaintiff wins; if at least nine jurors side with the defendant, then the defendant wins. Let p denote the probability that someone would side with the plaintiff based on the evidence. What is the probability, in terms of p, that the jury reaches a verdict one way or the other? How does this compare with your answer to part (a)?

81. Customers at a gas station pay with a credit card (A), debit card (B), or cash (C). Assume that successive customers make independent choices, with $P(A) = .5$, $P(B) = .2$, and $P(C) = .3$.

   a. Among the next 100 customers, what are the mean and variance of the number who pay with a debit card? Explain your reasoning.

   b. Answer part (a) for the number among the 100 who don't pay with cash.

82. An airport limousine can accommodate up to four passengers on any one trip. The company will accept a maximum of six reservations for a trip, and a passenger must have a reservation. From previous records, 20% of all those making reservations do not appear for the trip. In the following questions, assume independence, but explain why there could be dependence.

   a. If six reservations are made, what is the probability that at least one individual with a reservation cannot be accommodated on the trip?

   b. If six reservations are made, what is the expected number of available places when the limousine departs?

   c. Suppose the probability distribution of the number of reservations made is given in the accompanying table.

Reservations	3	4	5	6
Probability	.1	.2	.3	.4

   Let X denote the number of passengers on a randomly selected trip. Obtain the probability mass function of X.

83. Let X be a binomial random variable with a specified value of n.

   a. Are there values of p $(0 \le p \le 1)$ for which $V(X) = 0$? Explain why this is so.

   b. For what value of p is $V(X)$ maximized? [*Hint*: Either graph $V(X)$ as a function of p or else take a derivative.]

84. a. Verify the relationship $b(x; n, 1 - p) = b(n - x; n, p)$.

   b. Verify the relationship $B(x; n, 1 - p) = 1 - B(n - x - 1; n, p)$. [*Hint*: At most

x S's is equivalent to at least $(n - x)$ F's.]

c. What do parts (a) and (b) imply about the necessity of including values of $p > .5$ in Appendix Table A.1?

85. Refer to Chebyshev's inequality given in Exercise 45 (Section 3.3). Calculate $P(|X - \mu| \geq k\sigma)$ for $k = 2$ and $k = 3$ when $X \sim \text{Bin}(20, .5)$, and compare to the corresponding upper bounds. Repeat for $X \sim \text{Bin}(20, .75)$.

86. Show that $E(X) = np$ when X is a binomial random variable. [*Hint*: Express $E(X)$ as a sum with lower limit $x = 1$. Then factor out np, let $y = x - 1$ so that the remaining sum is from $y = 0$ to $y = n - 1$, and show that the sum equals 1.]

87. At the end of this section we obtained the mean and variance of a binomial rv using the mgf. Obtain the mean and variance instead from $R_X(t) = \ln[M_X(t)]$.

88. Obtain the moment generating function of the number of failures, $n - X$, in a binomial experiment, and use it to determine the expected number of failures and the variance of the number of failures. Are the expected value and variance intuitively consistent with the expressions for $E(X)$ and $V(X)$? Explain.

3.6   The Poisson Probability Distribution

The binomial distribution was derived by starting with an experiment consisting of trials and applying the laws of probability to various outcomes of the experiment. There is no simple experiment on which the Poisson distribution is based, although we will shortly describe how it can be obtained from the binomial distribution by certain limiting operations.

DEFINITION   A random variable X is said to have a **Poisson distribution** with parameter μ ($\mu > 0$) if the pmf of X is

$$p(x; \mu) = \frac{e^{-\mu}\mu^x}{x!} \quad x = 0, 1, 2, \ldots$$

We shall see shortly that μ is in fact the expected value of X, so the notation here is consistent with our previous use of the symbol μ. Because μ must be positive, $p(x; \mu) > 0$ for all possible x values. The fact that $\sum_{x=0}^{\infty} p(x; \mu) = 1$ is a consequence of the Taylor series expansion of e^{μ}, which appears in most calculus texts:

$$e^{\mu} = 1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + \cdots = \sum_{x=0}^{\infty} \frac{\mu^x}{x!} \tag{3.17}$$

If the two extreme terms in Expression (3.17) are multiplied by $e^{-\mu}$ and then $e^{-\mu}$ is placed inside the summation, the result is

$$1 = \sum_{x=0}^{\infty} \frac{e^{-\mu}\mu^x}{x!}$$

which shows that $p(x; \mu)$ fulfills the second condition necessary for specifying a pmf.

Example 3.42 The article "Detecting *Clostridium difficile* Outbreaks With Ward-Specific Cut-Off Levels Based on the Poisson Distribution" (*Infect. Control Hosp. Epidemiol.* 2019: 265–266) recommends using a Poisson model for X = the number of sporadic *C. difficile* infections (CDIs) in a month in a given hospital ward, as a way to determine when an "outbreak" (that is, an unusually large number of CDIs) has occurred. The article considers several values for μ for different wards in a particular hospital. For a ward in which $\mu = 3$ CDIs per month, the probability of observing exactly 5 CDIs in a particular month is

$$P(X = 5) = \frac{e^{-3}3^5}{5!} = .1008$$

and the chance of observing at least 5 CDIs is

$$P(X \geq 5) = 1 - P(X < 5) = 1 - \sum_{x=0}^{4} \frac{e^{-3}3^x}{x!} = 1 - e^{-3}\left[1 + 3 + \frac{3^2}{2!} + \frac{3^3}{3!} + \frac{3^4}{4!}\right] = .1847$$

These probabilities might not be so low as to convince hospital supervisors that they have an outbreak on their hands. On the other hand, in a ward with a historic mean of $\mu = 1$ CDI per month, the probabilities are $P(X = 5) = .0031$ and $P(X \geq 5) = .0037$, suggesting that five (or more) CDIs in one month would be extremely unusual and should be considered a *C. difficile* outbreak.   ∎

The Poisson Distribution as a Limit
The rationale for using the Poisson distribution in many situations is provided by the following proposition.

PROPOSITION   Suppose that in the binomial pmf $b(x; n, p)$ we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\mu > 0$. Then $b(x; n, p) \to p(x; \mu)$.

Proof Begin with the binomial pmf:

$$b(x; n, p) = \binom{n}{x}p^x(1 - p)^{n-x} = \frac{n!}{x!(n - x)!}p^x(1 - p)^{n-x}$$

$$= \frac{n \cdot (n - 1) \cdot \cdots \cdot (n - x + 1)}{x!}p^x(1 - p)^{n-x}$$

Now multiply both the numerator and denominator by n^x:

$$b(x; n, p) = \frac{n}{n}\frac{n - 1}{n} \cdot \cdots \cdot \frac{n - x + 1}{n} \cdot \frac{(np)^x}{x!} \cdot \frac{(1 - p)^n}{(1 - p)^x}$$

Taking the limit as $n \to \infty$ and $p \to 0$ with $np \to \mu$,

$$\lim_{n \to \infty} b(x; n, p) = 1 \cdot 1 \cdot \cdots \cdot 1 \cdot \frac{\mu^x}{x!} \cdot \left(\lim_{n \to \infty} \frac{(1 - np/n)^n}{1}\right)$$

The limit on the right can be obtained from the calculus theorem that says the limit of $(1 - a_n/n)^n$ is e^{-a} if $a_n \to a$. Because $np \to \mu$,

$$\lim_{n\to\infty} b(x;n,p) = \frac{\mu^x}{x!} \cdot \lim_{n\to\infty} \left(1 - \frac{np}{n}\right)^n = \frac{\mu^x e^{-\mu}}{x!} = p(x;\mu) \qquad\blacksquare$$

According to the proposition, *in any binomial experiment for which the number of trials n is large and the success probability p is small, b(x; n, p) ≈ p(x; μ) where μ = np.* It is interesting to note that Siméon Poisson discovered this eponymous distribution by this approach in the 1830s.

Table 3.3 shows the Poisson distribution for $\mu = 3$ along with three binomial distributions with $np = 3$, and Figure 3.9 (from R) plots the Poisson along with the first two binomial distributions. The approximation is of limited use for $n = 30$, but of course the accuracy is better for $n = 100$ and much better for $n = 300$.

Table 3.3 Comparing the Poisson and three binomial distributions

x	$n = 30$, $p = .1$	$n = 100$, $p = .03$	$n = 300$, $p = .01$	Poisson, $\mu = 3$
0	0.042391	0.047553	0.049041	0.049787
1	0.141304	0.147070	0.148609	0.149361
2	0.227656	0.225153	0.224414	0.224042
3	0.236088	0.227474	0.225170	0.224042
4	0.177066	0.170606	0.168877	0.168031
5	0.102305	0.101308	0.100985	0.100819
6	0.047363	0.049610	0.050153	0.050409
7	0.018043	0.020604	0.021277	0.021604
8	0.005764	0.007408	0.007871	0.008102
9	0.001565	0.002342	0.002580	0.002701
10	0.000365	0.000659	0.000758	0.000810
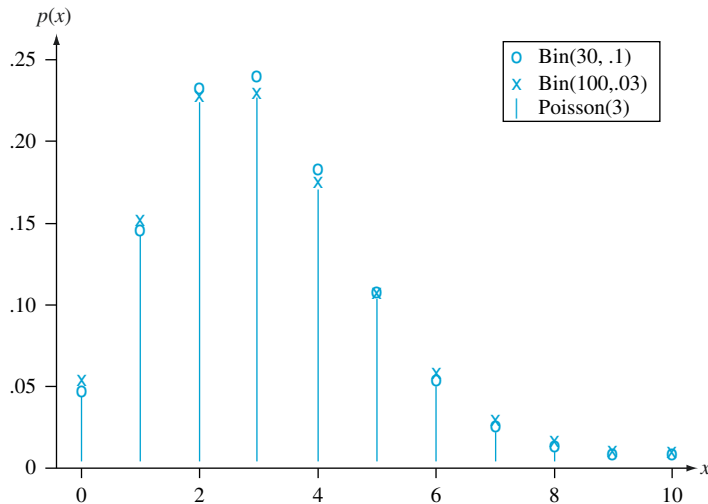
Figure 3.9 Comparing a Poisson and two binomial distributions

Example 3.43

Suppose you have a 4-megabit modem (4,000,000 bits/s) with bit error probability 10^{-8}. Assume bit errors occur independently, and assume your bit rate stays constant at 4 Mbps. What is the probability of exactly 3 bit errors in the next minute? Of at most 3 bit errors in the next minute?

Define a random variable X = the number of bit errors in the next minute. From the description, X satisfies the conditions of a binomial distribution; specifically, since a constant bit rate of 4 Mbps equates to 240,000,000 bits transmitted per minute, $X \sim \text{Bin}(240,000,000, 10^{-8})$. Hence, the probability of exactly three bit errors in the next minute is

$$P(X = 3) = b(3; 240,000,000, 10^{-8}) = \binom{24,000,000}{3}(10^{-8})^3(1 - 10^{-8})^{239,999,997}$$

For a variety of reasons, some calculators will struggle with this computation. The expression for the chance of at most 3 bit errors, $P(X \le 3)$, is even worse. (The inability to compute such expressions in the nineteenth century, even with modest values of n and p, was Poisson's motive to derive an easily computed approximation.)

We may approximate these probabilities using the Poisson distribution. The parameter μ is given by $\mu = np = 240,000,000(10^{-8}) = 2.4$, whence

$$P(X = 3) \approx p(3; 2.4) = \frac{e^{-2.4}2.4^3}{3!} = .20901416$$

Similarly, the probability of at most 3 bit errors in the next minute is approximated by

$$P(X \le 3) \approx \sum_{x=0}^{3} p(x, 2.4) = \sum_{x=0}^{3} \frac{e^{-2.4}2.4^x}{x!} = .77872291$$

Using software, the exact probabilities (i.e., using the binomial model) are .2090141655 and .7787229106, respectively. The Poisson approximations agree to eight decimal places and are clearly more computationally tractable.   ∎

Many software packages will compute both $p(x; \mu)$ and the corresponding cdf $P(x; \mu)$ for specified values of x and μ upon request; the relevant R functions appear in Table 3.4. Appendix Table A.2 exhibits the cdf $P(x; \mu)$ for $\mu = .1, .2, \ldots, 1, 2, \ldots, 10, 15,$ and 20. For example, if $\mu = 2$, then $P(X \le 3) = P(3; 2) = .857$, whereas $P(X = 3) = P(3; 2) - P(2; 2) = .180$.

Table 3.4   Poisson probability calculations in R

Function:	pmf	cdf
Notation:	$p(x; \mu)$	$P(x; \mu)$
R:	dpois(x, μ)	ppois(x, μ)

The Mean, Variance, and MGF of a Poisson Random Variable

Since $b(x; n, p) \to p(x; \mu)$ as $n \to \infty, p \to 0, np \to \mu$, one might guess that the mean and variance of a binomial variable approach those of a Poisson variable. These limits are, respectively, $np \to \mu$ and $np(1 - p) \to \mu$.

| **PROPOSITION** | If X has a Poisson distribution with parameter μ, then $E(X) = V(X) = \mu$. |

These results can also be derived directly from the definitions of mean and variance (see Exercise 104 for the mean).

Example 3.44 (Example 3.42 continued) For the hospital ward with $\mu = 3$, the expected number of CDIs in a month is 3 (obviously), and the standard deviation of the number of monthly CDIs is $\sigma_X = \sqrt{\mu} = \sqrt{3} = 1.73$. So, observing 2–4 CDIs in a month would not be unusual (those values are within one sd of the mean), but a month with 7 CDIs on the ward would be alarming (since that's more than two standard deviations above average). ∎

The moment generating function of the Poisson distribution is easy to derive, and it gives a direct route to the mean and variance (Exercise 106).

| **PROPOSITION** | The Poisson moment generating function is |
| | $$M_X(t) = e^{\mu(e^t - 1)}$$ |

Proof The mgf is by definition

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} e^{-\mu} \frac{\mu^x}{x!} = e^{-\mu} \sum_{x=0}^{\infty} \frac{(\mu e^t)^x}{x!} = e^{-\mu} e^{\mu e^t} = e^{\mu(e^t - 1)}$$

This uses the series expansion $\sum_{x=0}^{\infty} u^x / x! = e^u$.   ∎

The Poisson Process

A very important application of the Poisson distribution arises in connection with the occurrence of events over time. As an example, suppose that starting from a time point that we label $t = 0$, we are interested in counting the number of radioactive pulses recorded by a Geiger counter. We make the following assumptions about the way in which pulses occur:

1. There exists a parameter $\lambda > 0$ such that for any short time interval of length Δt, the probability that exactly one pulse is received is $\lambda \cdot \Delta t + o(\Delta t)$.[4]
2. The probability of more than one pulse being received during Δt is $o(\Delta t)$. [This, along with Assumption 1, implies that the probability of no pulses during Δt is $1 - \lambda \cdot \Delta t - o(\Delta t)$].
3. The number of pulses received during the time interval Δt is independent of the number received prior to this time interval.

Informally, Assumption 1 says that for a short interval of time, the probability of receiving a single pulse is approximately proportional to the length of the time interval, where λ is the constant of proportionality. Now let $P_k(t)$ denote the probability that exactly k pulses will be received by the counter during any particular time interval of length t.

| **PROPOSITION** | $P_k(t) = e^{-\lambda t}(\lambda t)^k / k!$, so that the number of pulses during a time interval of length t is a Poisson rv with parameter $\mu = \lambda t$. The expected number of pulses during any such time interval is then λt, so the expected number during a unit interval of time is λ. |

---

[4] A quantity is $o(\Delta t)$ (read "little o of delta t") if, as Δt approaches 0, so does $o(\Delta t)/\Delta t$. That is, $o(\Delta t)$ is even more negligible than Δt itself. The quantity $(\Delta t)^2$ has this property, but $\sin(\Delta t)$ does not.

Example 3.43 hints at why this might be reasonable: if we "digitize" time—that is, divide time into discrete pieces, such as transmitted bits—and look at the number of the resulting time pieces that include an event, a binomial model is often applicable. If the number of time pieces is very large and the success probability close to zero, which would occur if we divided a fixed time frame into ever-smaller pieces, then we may invoke the Poisson approximation from earlier in this section. See Exercise 105 for a derivation.

Example 3.45 Suppose pulses arrive at the Geiger counter at an average rate of six per minute, so that $\lambda = 6$. To find the probability that in a 30-second interval at least one pulse is received, note that the number of pulses in such an interval has a Poisson distribution with parameter $\lambda t = 6(.5) = 3$ (.5 min is used because λ is expressed as a rate per minute). Then with X = the number of pulses received in the 30-second interval,

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{e^{-3}3^0}{0!} = .950$$

In a one-hour interval ($t = 60$), the expected number of pulses is $\mu = \lambda t = 6(60) = 360$, with a standard deviation of $\sigma = \sqrt{\mu} = \sqrt{360} = 18.97$. According to this model, in a typical hour we will observe 360 ± 19 pulses arrive at the Geiger counter. ∎

If in Assumptions 1–3 we replace "pulse" by "event," then the number of events occurring during a fixed time interval of length t has a Poisson distribution with parameter λt. Any process that has this distribution is called a **Poisson process**, and λ is called the *rate of the process*. Other examples of situations giving rise to a Poisson process include monitoring the status of a computer system over time, with breakdowns constituting the events of interest; recording the number of accidents in an industrial facility over time; logging hits to a website; and observing the number of cosmic-ray showers from an observatory over time.

Instead of observing events over time, consider observing events of some type that occur in a two- or three-dimensional region. For example, we might select on a map a certain region R of a forest, go to that region, and count the number of trees. Each tree would represent an event occurring at a particular point in space. Under assumptions similar to 1–3, it can be shown that the number of events occurring in a region R has a Poisson distribution with parameter $\lambda \cdot a(R)$, where $a(R)$ is the area or volume of R. The quantity λ is the expected number of events per unit area or volume.

Exercises: Section 3.6 (89–107)

89. Let X, the number of flaws on the surface of a randomly selected carpet of a particular type, have a Poisson distribution with parameter $\mu = 5$. Use software or Appendix Table A.2 to compute the following probabilities:

    a. $P(X \leq 8)$
    b. $P(X = 8)$
    c. $P(9 \leq X)$
    d. $P(5 \leq X \leq 8)$
    e. $P(5 < X < 8)$

90. Suppose the number X of tornadoes observed in a particular region during a 1-year period has a Poisson distribution with $\mu = 8$.

    a. Compute $P(X \leq 5)$.
    b. Compute $P(6 \leq X \leq 9)$.
    c. Compute $P(10 \leq X)$.
    d. What is the probability that the observed number of tornadoes exceeds the expected number by more than 1 standard deviation?

91. Suppose that the number of drivers who travel between a particular origin and destination during a designated time period has a Poisson distribution with parameter $\mu = 20$ (suggested in the article "Dynamic Ride Sharing: Theory and Practice," *J. Transp. Engr.* 1997: 308–312). What is the probability that the number of drivers will

    a. Be at most 10?
    b. Exceed 20?
    c. Be between 10 and 20, inclusive? Be strictly between 10 and 20?
    d. Be within 2 standard deviations of the mean value?

92. Consider writing onto a computer disk and then sending it through a certifier that counts the number of missing pulses. Suppose this number X has a Poisson distribution with parameter $\mu = .2$. (Suggested in "Average Sample Number for Semi-Curtailed Sampling Using the Poisson Distribution," *J. Qual. Tech.* 1983: 126–129.)

    a. What is the probability that a disk has exactly one missing pulse?
    b. What is the probability that a disk has at least two missing pulses?
    c. If two disks are independently selected, what is the probability that neither contains a missing pulse?

93. The article "Metal Hips Fail Faster, Raise Other Health Concerns" on the www.arthritis.com website reported that the five-year failure rate of metal-on-plastic implants was 1.7% (rates for metal-on-metal and ceramic implants were significantly higher). Use both a binomial calculation and a Poisson approximation to answer each of the following.

    a. Among 200 randomly selected such implants, what is the probability that exactly three will fail?
    b. Among 200 randomly selected such implants, what is the probability that at most three will fail?

94. Suppose that only .10% of all computers of a certain type experience CPU failure during the warranty period. Consider a sample of 10,000 computers.

    a. What are the expected value and standard deviation of the number of computers in the sample that have the defect?
    b. What is the (approximate) probability that more than 10 sampled computers have the defect?
    c. What is the (approximate) probability that no sampled computers have the defect?

95. If a publisher of nontechnical books takes great pains to ensure that its books are free of typographical errors, so that the probability of any given page containing at least one such error is .005 and errors are independent from page to page, what is the probability that one of its 400-page novels will contain exactly one page with errors? At most three pages with errors?

96. In proof testing of circuit boards, the probability that any particular diode will fail is .01. Suppose a circuit board contains 200 diodes.

    a. How many diodes would you expect to fail, and what is the standard deviation of the number that are expected to fail?
    b. What is the (approximate) probability that at least four diodes will fail on a randomly selected board?
    c. If five boards are shipped to a particular customer, how likely is it that at least four of them will work properly? (A board works properly only if all its diodes work.)

97. Suppose small aircraft arrive at an airport according to a Poisson process with rate $\lambda = 8$ per hour, so that the number of arrivals during a time period of t hours is a Poisson rv with parameter $\mu = 8t$.

    a. What is the probability that exactly 6 small aircraft arrive during a 1-h period? At least 6? At least 10?

b. What are the expected value and standard deviation of the number of small aircraft that arrive during a 90-min period?

c. What is the probability that at least 20 small aircraft arrive during a 2.5-h period? That at most 10 arrive during this period?

98. The number of people arriving for treatment at an emergency room can be modeled by a Poisson process with a rate parameter of 5 per hour.

a. What is the probability that exactly four arrivals occur during a particular hour?

b. What is the probability that at least four people arrive during a particular hour?

c. How many people do you expect to arrive during a 45-min period?

99. The number of requests for assistance received by a towing service is a Poisson process with rate $\lambda = 4$ per hour.

a. Compute the probability that exactly ten requests are received during a particular 2-h period.

b. If the operators of the towing service take a 30-min break for lunch, what is the probability that they do not miss any calls for assistance?

c. How many calls would you expect during their break?

100. The article "Expectation Analysis of the Probability of Failure for Water Supply Pipes" (*J. Pipeline Syst. Engr. Pract.* 2012: 36–46) recommends using a Poisson process to model the number of failures in commercial water pipes. The article also gives estimates of the failure rate λ, in units of failures per 100 miles of pipe per day, for four different types of pipe and for many different years.

a. For PVC pipe in 2008, the authors estimate a failure rate of 0.0081 failures per 100 miles of pipe per day. Consider a 100-mile-long segment of such pipe. What is the expected number of failures in one year (365 days)? Based on this expectation, what is the probability of at least one failure along such a pipe in one year?

b. For cast iron pipe in 2005, the authors' estimate is $\lambda = 0.0864$ failures per 100 miles per day. Suppose a town had 1500 miles of cast iron pipe underground in 2005. What is the probability of at least one failure somewhere along this pipe system on any given day?

101. The article "Reliability-Based Service-Life Assessment of Aging Concrete Structures" (*J. Struct. Engr.* 1993: 1600–1621) suggests that a Poisson process can be used to represent the occurrence of structural loads over time. Suppose the mean time between occurrences of loads (which can be shown to be = $1/\lambda$) is .5 year.

a. How many loads can be expected to occur during a 2-year period?

b. What is the probability that more than five loads occur during a 2-year period?

c. How long must a time period be so that the probability of no loads occurring during that period is at most .1?

102. Automobiles arrive at a vehicle equipment inspection station according to a Poisson process with rate $\lambda = 10$ per hour. Suppose that with probability .5 an arriving vehicle will have no equipment violations.

a. What is the probability that exactly ten arrive during the hour and all ten have no violations?

b. For any fixed $y \geq 10$, what is the probability that y arrive during the hour, of which ten have no violations?

c. What is the probability that ten "no-violation" cars arrive during the next

hour? [*Hint*: Sum the probabilities in part (b) from $y = 10$ to ∞.]

103. Suppose that trees are distributed in a forest according to a two-dimensional Poisson process with parameter λ, the expected number of trees per acre, equal to 80.

    a. What is the probability that in a certain quarter-acre plot, there will be at most 16 trees?
    b. If the forest covers 85,000 acres, what is the expected number of trees in the forest?
    c. Suppose you select a point in the forest and construct a circle of radius.1 mile. Let X = the number of trees within that circular region. What is the pmf of X? [*Hint*: 1 sq mile = 640 acres.]

104. Let X have a Poisson distribution with parameter μ. Show that $E(X) = \mu$ directly from the definition of expected value. [*Hint*: The first term in the sum equals 0, and then x can be canceled. Now factor out μ and show that what is left sums to 1.]

105. a. In a Poisson process, what has to happen in both the time interval $(0, t)$ and the interval $(t, t + \Delta t)$ so that no events occur in the entire interval $(0, t + \Delta t)$? Use this and Assumptions 1–3 to write a relationship between $P_0(t + \Delta t)$ and $P_0(t)$.
    b. Use the result of part (a) to write an expression for the difference $P_0(t + \Delta t) - P_0(t)$. Then divide by Δt and let $\Delta t \to 0$ to obtain an equation involving $(d/dt)P_0(t)$, the derivative of $P_0(t)$ with respect to t.

    c. Verify that $P_0(t) = e^{-\lambda t}$ satisfies the equation of part (b).
    d. It can be shown in a manner similar to parts (a) and (b) that the $P_k(t)$'s must satisfy the system of differential equations

    $$\frac{d}{dt}P_k(t) = \lambda P_{k-1}(t) - \lambda P_k(t)$$
    $$k = 1, 2, 3, \ldots$$

    Verify that $P_k(t) = e^{-\lambda t}(\lambda t)^k/k!$ satisfies the system. (This is actually the only solution.)

106. a. Use derivatives of the moment generating function to obtain the mean and variance for the Poisson distribution.
    b. As discussed in Section 3.4, obtain the Poisson mean and variance from $R_X(t) = \ln[M_X(t)]$. In terms of effort, how does this method compare with the one in part (a)?

107. Show that the binomial moment generating function converges to the Poisson moment generating function if we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\mu > 0$. [*Hint*: Use the calculus theorem that was used in showing that the binomial probabilities converge to the Poisson probabilities.] There is in fact a theorem saying that convergence of the mgf implies convergence of the probability distribution. In particular, convergence of the binomial mgf to the Poisson mgf implies $b(x; n, p) \to p(x; \mu)$.

3.7   Other Discrete Distributions

This section introduces discrete distributions that are closely related to the binomial distribution. Whereas the binomial distribution is the *approximate* probability model for sampling without replacement from a finite dichotomous (*S*/*F*) population, the hypergeometric distribution is the exact probability model for the number of *S*'s in the sample. The binomial rv X is the number of *S*'s when the number n of trials is fixed, whereas the negative binomial distribution arises from fixing the number of *S*'s desired and letting the number of trials be random.

The Hypergeometric Distribution

The assumptions leading to the hypergeometric distribution are as follows:

1. The population or set to be sampled consists of N individuals, objects, or elements (a *finite* population).
2. Each individual can be characterized as a success (S) or a failure (F), and there are M successes in the population.
3. A sample of n individuals is selected without replacement in such a way that each subset of size n is equally likely to be chosen.

The random variable of interest is X = the number of S's in the sample. The probability distribution of X depends on the parameters n, M, and N, so we wish to obtain $P(X = x) = h(x; n, M, N)$.

Example 3.46 During a particular period, a university's information technology office received 20 service orders for problems with laptops, of which 8 were Macs and 12 were PCs. A sample of 5 of these service orders is to be selected for inclusion in a customer satisfaction survey. Suppose that the 5 are selected in a completely random fashion, so that any particular subset of size 5 has the same chance of being selected as does any other subset (think of putting the numbers 1, 2, …, 20 on 20 identical slips of paper, mixing up the slips, and choosing 5 of them). What then is the probability that exactly 2 of the selected service orders were for PC laptops?

In this example, the population size is $N = 20$, the sample size is $n = 5$, and the number of S's (PC = S) and F's (Mac = F) in the population are $M = 12$ and $N - M = 8$, respectively. Let X = the number of PCs among the 5 sampled service orders. Because all outcomes (each consisting of 5 particular orders) are equally likely,

$$P(X = 2) = h(2; 5, 12, 20) = \frac{\text{number of outcomes having } X = 2}{\text{number of possible outcomes}}$$

The number of possible outcomes in the experiment is the number of ways of selecting 5 from the 20 objects without regard to order—that is, $\binom{20}{5}$. To count the number of outcomes having $X = 2$, note that there are $\binom{12}{2}$ ways of selecting 2 of the PC orders, and for each such way there are $\binom{8}{3}$ ways of selecting the 3 Mac orders to fill out the sample. The Fundamental Counting Principle from Section 2.3 then gives $\binom{12}{2} \cdot \binom{8}{3}$ as the number of outcomes with $X = 2$, so

$$h(2; 5, 12, 20) = \frac{\binom{12}{2}\binom{8}{3}}{\binom{20}{5}} = \frac{(66)(56)}{15,504} = \frac{77}{323} = .238 \qquad \blacksquare$$

In general, if the sample size n is smaller than the number of successes in the population (M), then the largest possible X value is n. However, if $M < n$ (e.g., a sample size of 25 and only 15 successes in the population), then X can be at most M. Similarly, whenever the number of population failures ($N - M$) exceeds the sample size, the smallest possible X value is 0 (since all sampled individuals might then be failures). However, if $N - M < n$, the smallest possible X value is $n - (N - M)$. Thus, the possible values of X satisfy the restriction $\max(0, n - N + M) \leq x \leq \min(n, M)$. An argument parallel to that of the previous example gives the pmf of X.

PROPOSITION   If X is the number of S's in a random sample of size n drawn
from a population consisting of M S's and $(N - M)$ F's, then
the probability distribution of X, called the **hypergeometric
distribution**, is given by

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N - M}{n - x}}{\binom{N}{n}} \tag{3.18}$$

for x an integer satisfying $\max(0,\ n - N + M) \leq x \leq \min(n, M)$.[5]

In Example 3.46, $n = 5$, $M = 12$, and $N = 20$, so $h(x; 5, 12, 20)$ for $x = 0, 1, 2, 3, 4, 5$ can be obtained
by substituting these numbers into Equation (3.19).

Example 3.47 *Capture–recapture.* Five individuals from an animal population thought to be near
extinction in a region have been caught, tagged, and released to mix into the population. After they
have had an opportunity to mix, a random sample of ten of these animals is selected. Let $X =$ the
number of tagged animals in the second sample. If there are actually 25 animals of this type in the
region, what is the probability that (a) $X = 2$? (b) $X \leq 2$?

Application of the hypergeometric distribution here requires assuming that every subset of 10
animals has the same chance of being captured. This in turn implies that released animals are no
easier or harder to catch than are those not initially captured. Then the parameter values are $n = 10$,
$M = 5$ (5 tagged animals in the population), and $N = 25$, so

$$h(x;\ 10, 5, 25) = \frac{\binom{5}{x}\binom{20}{10 - x}}{\binom{25}{10}} \quad x = 0, 1, 2, 3, 4, 5$$

For part (a),

$$P(X = 2) = h(2;\ 10, 5, 25) = \frac{\binom{5}{2}\binom{20}{8}}{\binom{25}{10}} = .385$$

For part (b),

$$P(X \leq 2) = P(X = 0, 1,\ \text{or } 2) = \sum_{x=0}^{2} h(x;\ 10, 5, 25)$$
$$= .057 + .257 + .385 = .699 \qquad \blacksquare$$

R and other software packages will easily generate hypergeometric probabilities; see Table 3.5 at the
end of this section. Comprehensive tables of the hypergeometric distribution are available, but

---

[5]If we define $\binom{a}{b} = 0$ for $a < b$, then $h(x; n, M, N)$ may be applied for all integers $0 \leq x \leq n$.

because the distribution has three parameters, these tables require much more space than tables for the binomial or Poisson distributions.

As in the binomial case, there are simple expressions for $E(X)$ and $V(X)$ for hypergeometric rvs.

PROPOSITION   The mean and variance of the hypergeometric rv X having pmf $h(x; n, M, N)$ are

$$E(X) = n \cdot \frac{M}{N} \quad V(X) = \left(\frac{N - n}{N - 1}\right) \cdot n \cdot \frac{M}{N} \left(1 - \frac{M}{N}\right)$$

The proof will be given in Section 6.3. We do not give the moment generating function for the hypergeometric distribution, because the mgf is more trouble than it is worth here.

The ratio M/N is the proportion of S's in the population. Replacing M/N by p in $E(X)$ and $V(X)$ gives

$$E(X) = np \quad V(X) = \left(\frac{N - n}{N - 1}\right) \cdot np(1 - p) \tag{3.19}$$

Expression (3.19) shows that the means of the binomial and hypergeometric rvs are equal, whereas the variances of the two rvs differ by the factor $(N - n)/(N - 1)$, often called the **finite population correction factor**. This factor is < 1, so the hypergeometric variable has smaller variance than does the binomial rv. The correction factor can be written $(1 - n/N)/(1 - 1/N)$, which is approximately 1 when n is small relative to N.

Example 3.48 (Example 3.47 continued) In the animal-tagging example, $n = 10$, $M = 5$, and $N = 25$, so $p = 5/25 = .2$ and

$$E(X) = 10(.2) = 2$$
$$V(X) = \frac{15}{24}(10)(.2)(.8) = (.625)(1.6) = 1$$

If the sampling were carried out with replacement, $V(X) = 1.6$.

Suppose the population size N is not actually known, so the value x is observed and we wish to estimate N. It is reasonable to equate the observed sample proportion of S's, x/n, with the population proportion, M/N, giving the estimate

$$\hat{N} = \frac{M \cdot n}{x}$$

If $M = 100$, $n = 40$, and $x = 16$, then $\hat{N} = 250$.   ■

Our rule in Section 3.5 stated that if sampling is without replacement but n/N is at most .05, then the binomial distribution can be used to compute approximate probabilities involving the number of S's in the sample. A more precise statement is as follows: Let the population size, N, and number of population S's, M, get large with the ratio M/N approaching p. Then $h(x; n, M, N)$ approaches $b(x; n, p)$; so for n/N small, the two are approximately equal provided that p is not too near either 0 or 1. This is the rationale for our rule.

The Negative Binomial and Geometric Distributions

The negative binomial distribution is based on an experiment satisfying the following conditions:

1. The experiment consists of a sequence of independent trials.
2. Each trial can result in either a success (S) or a failure (F).
3. The probability of success is constant from trial to trial, so $P(S$ on trial $i) = p$ for $i = 1, 2, 3 \ldots$.
4. The experiment continues (trials are performed) until a total of r successes have been observed, where r is a specified positive integer.

The random variable of interest is $X =$ the number of trials required to achieve the rth success, and X is called a **negative binomial random variable**. In contrast to the binomial rv, the number of *successes* is fixed and the number of *trials* is random. Possible values of X are $r, r + 1, r + 2, \ldots$, since it takes at least r trials to achieve r successes.

Let $nb(x; r, p)$ denote the pmf of X. The event $\{X = x\}$ is equivalent to $\{r - 1\ S$'s in the first $(x - 1)$ trials and an S on the xth trial$\}$; e.g., if $r = 5$ and $x = 15$, then there must be four S's in the first 14 trials and trial 15 must be an S. Since trials are independent,

$$nb(x; r, p) = P(X = x) = P(r - 1\ S\text{'s on the first } x - 1 \text{ trials}) \cdot P(S) \qquad (3.20)$$

The first probability on the far right of Expression (3.20) is the binomial probability

$$\binom{x-1}{r-1} p^{r-1}(1-p)^{(x-1)-(r-1)} \quad \text{where } p = P(S)$$

Simplifying and then multiplying by the extra factor of p at the end of (3.20) yields the pmf.

PROPOSITION   The pmf of the negative binomial rv X with parameters $r =$ desired number of S's and $p = P(S)$ is

$$nb(x; r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad x = r, r+1, r+2, \ldots$$

Example 3.49 A pediatrician wishes to recruit 4 couples, each of whom is expecting their first child, to participate in a new natural childbirth regimen. Let $p = P$ (a randomly selected couple agrees to participate). If $p = .2$, what is the probability that exactly 15 couples must be asked before 4 are found who agree to participate? Substituting $r = 4$, $p = .2$, and $x = 15$ into $nb(x; r, p)$ gives

$$nb(15; 4, 2) = \binom{15 - 1}{4 - 1} .2^4 .8^{11} = .050$$

The probability that at most 15 couples need to be asked is

$$P(X \leq 15) = \sum_{x=4}^{15} nb(x; 4, .2) = \sum_{x=4}^{15} \binom{x-1}{3} .2^4 .8^{x-4} = .352 \qquad \blacksquare$$

In the special case $r = 1$, the pmf is

$$nb(x; 1, p) = (1 - p)^{x-1}p \quad x = 1, 2, \ldots \tag{3.21}$$

In Example 3.10, we derived the pmf for the number of trials necessary to obtain the first S, and the pmf there is identical to Expression (3.21). The variable X = number of trials required to achieve one success is referred to as a **geometric random variable**, and the pmf in (3.21) is called the **geometric distribution**. The name is appropriate because the probabilities form a geometric series: p, $(1 - p)p$, $(1 - p)^2 p$, ... To see that the sum of the probabilities is 1, recall that the sum of a geometric series is $a + ar + ar^2 + \cdots = a/(1-r)$ if $|r| < 1$, so for $p > 0$,

$$p + (1 - p)p + (1 - p)^2 p + \cdots = \frac{p}{1 - (1 - p)} = 1$$

In Example 3.19, the expected number of trials until the first S was shown to be $1/p$. Intuitively, we would then expect to need $r \cdot 1/p$ trials to achieve the rth S, and this is indeed $E(X)$. There is also a simple formula for $V(X)$ and for the mgf.

PROPOSITION   If X is a negative binomial rv with parameters r and p, then

$$E(X) = \frac{r}{p} \quad V(X) = \frac{r(1 - p)}{p^2} \quad M_X(t) = \left(\frac{pe^t}{1 - (1 - p)e^t}\right)^r$$

See Exercise 123 for a derivation of these formulas. The corresponding formulas for the geometric distribution are obtained by substituting $r = 1$ above.

Example 3.50 (Example 3.49 continued) With $p = .2$, the expected number of couples the doctor must speak to in order to find 4 that will agree to participate is $r/p = 4/.2 = 20$. This makes sense, since with $p = .2 = 1/5$ it will take 5 attempts, on average, to achieve one success. The corresponding variance is $4(1 - .2)/(.2)^2 = 80$, for a standard deviation of about 8.9. ∎

Since they are based on similar experiments, some caution must be taken to distinguish the binomial and negative binomial models, as seen in the next example.

Example 3.51 In many communication systems, a receiver will send a short signal back to the transmitter to indicate whether a message has been received correctly or with errors. (These signals are often called an *acknowledgement* and a *nonacknowledgement*, respectively. Bit sum checks and other tools are used by the receiver to determine the absence or presence of errors.) Assume we are using such a system in a noisy channel, so that each message is sent error-free with probability .86, independent of all other messages. What is the probability that in 10 transmissions, exactly 8 will succeed? What is the probability the system will require exactly 10 attempts to successfully transmit 8 messages?

While these two questions may sound similar, they require two different models for solution. To answer the first question, let X = the number of successful transmissions among the 10. Then $X \sim$ Bin(10, .86), and the answer is

$$P(X = 8) = b(8; 10, .86) = \binom{10}{8}(.86)^8(.14)^2 = .2639$$

However, the event {exactly 10 attempts required to successfully transmit 8 messages} is more restrictive: not only must we observe 8 S's and 2 F's in 10 trials, but *the last trial must be a success*. Otherwise, it took fewer than 10 tries to send 8 messages successfully. Define a variable Y = the number of transmissions (trials) required to successfully transmit 8 messages. Then Y is negative binomial, with $r = 8$ and $p = .86$, and the answer to the second question is

$$P(Y = 10) = nb(10; 8, .86) = \binom{10 - 1}{8 - 1}(.86)^8(.14)^2 = .2111$$

Notice this is smaller than the answer to the first question, which makes sense because (as we noted) the second question imposes an additional constraint. In fact, you can think of the "–1" terms in the negative binomial pmf as accounting for this loss of flexibility in the placement of S's and F's.

Similarly, the expected number of successful transmissions in 10 attempts is $E(X) = np = 10(.86)$ = 8.6, while the expected number of attempts required to successfully transmit 8 messages is $E(Y) = r/p = 8/.86 = 9.3$. In the first case, the number of trials ($n = 10$) is fixed, while in the second case the desired number of successes ($r = 8$) is fixed.   ∎

By expanding the binomial coefficient in front of $p^r(1 - p)^{x-r}$ and doing some cancelation, it can be seen that $nb(x; r, p)$ is well defined even when r is not an integer. This *generalized negative binomial distribution* has been found to fit observed data quite well in a wide variety of applications.

Alternative Definition of the Negative Binomial Distribution

There is not universal agreement on the definition of a negative binomial random variable (or, by extension, a geometric rv). It is not uncommon in the literature, as well as in some textbooks (including previous editions of this book), to see the *number of failures preceding the rth success* called "negative binomial"; in our notation, this simply equals $X - r$. Possible values of this "number of failures" variable are 0, 1, 2, …. Similarly, the geometric distribution is sometimes defined in terms of the number of failures preceding the first success in a sequence of independent and identical trials. If one uses these alternative definitions, then the pmf, mean, and mgf formulas must be adjusted accordingly (the variance, however, will stay the same). See Exercise 124.

The developers of R are among those who have adopted this alternative definition; as a result, we must be careful with our inputs to the relevant software functions. The pmf syntax for the distributions in this section are cataloged in Table 3.5; cdfs may be invoked by changing the initial letter d to p in R. Notice the input argument $x - r$ for the negative binomial functions: R requests the number of *failures*, rather than the number of *trials*.

Table 3.5 R code for hypergeometric and negative binomial calculations

	Hypergeometric	Negative Binomial
Function:	pmf	pmf
Notation:	$h(x; n, M, N)$	$nb(x; r, p)$
R:	dhyper($x, M, N–M, n$)	dnbinom($x–r, r, p$)

For example, suppose X has a hypergeometric distribution with $n = 10$, $M = 5$, $N = 25$ as in Example 3.47. Using R, we may calculate $P(X = 2)$ = dhyper(2,5,20,10) and $P(X \leq 2)$ = phyper(2,5,20,10). If X is the negative binomial variable of Example 3.49 with parameters $r = 4$ and $p = .2$, then the chance of requiring 15 trials to achieve 4 successes (i.e., 11 total failures) can be found in in R with dnbinom(11,4,.2).

Exercises: Section 3.7 (108–124)

108. An electronics store has received a shipment of 20 table radios that have connections for an iPod or iPhone. Twelve of these have two slots (so they can accommodate both devices), and the other eight have a single slot. Suppose that six of the 20 radios are randomly selected to be stored under a shelf where radios are displayed, and the remaining ones are placed in a storeroom. Let $X =$ the number among the radios stored under the display shelf that have two slots.

   a. What kind of a distribution does X have (name and values of all parameters)?
   b. Compute $P(X = 2)$, $P(X \leq 2)$, and $P(X \geq 2)$.
   c. Calculate the mean value and standard deviation of X.

109. Each of 12 refrigerators has been returned to a distributor because of an audible, high-pitched, oscillating noise when the refrigerator is running. Suppose that 7 of these refrigerators have a defective compressor and the other 5 have less serious problems. If the refrigerators are examined in random order, let X be the number among the first 6 examined that have a defective compressor. Compute the following:

   a. $P(X = 5)$
   b. $P(X \leq 4)$
   c. The probability that X exceeds its mean value by more than 1 standard deviation.
   d. Consider a large shipment of 400 refrigerators, of which 40 have defective compressors. If X is the number among 15 randomly selected refrigerators that have defective compressors, describe a less tedious way to calculate (at least approximately) $P(X \leq 5)$ than to use the hypergeometric pmf.

110. An instructor who taught two sections of statistics last term, the first with 20 students and the second with 30, decided to assign a term project. After all projects had been turned in, the instructor randomly ordered them before grading. Consider the first 15 graded projects.

   a. What is the probability that exactly 10 of these are from the second section?
   b. What is the probability that at least 10 of these are from the second section?
   c. What is the probability that at least 10 of these are from the same section?
   d. What are the mean value and standard deviation of the number among these 15 that are from the second section?
   e. What are the mean value and standard deviation of the number of projects not among these first 15 that are from the second section?

111. A geologist has collected 10 specimens of basaltic rock and 10 specimens of granite. The geologist instructs a laboratory assistant to randomly select 15 of the specimens for analysis.

   a. What is the pmf of the number of granite specimens selected for analysis?
   b. What is the probability that all specimens of one of the two types of rock are selected for analysis?
   c. What is the probability that the number of granite specimens selected for analysis is within 1 standard deviation of its mean value?

112. Suppose that 20% of all individuals have an adverse reaction to a particular drug. A medical researcher will administer the drug to one individual after another until the first adverse reaction occurs. Define an appropriate random variable and use its distribution to answer the following questions.

   a. What is the probability that when the experiment terminates, four individuals have not had adverse reactions?
   b. What is the probability that the drug is administered to exactly five individuals?

c. What is the probability that at most four individuals do not have an adverse reaction?

d. How many individuals would you expect to not have an adverse reaction, and to how many individuals would you expect the drug to be given?

e. What is the probability that the number of individuals given the drug is within 1 standard deviation of what you expect?

113. Twenty pairs of individuals playing in a bridge tournament have been seeded 1, …, 20. In the first part of the tournament, the 20 are randomly divided into 10 east–west pairs and 10 north–south pairs.

a. What is the probability that x of the top 10 pairs end up playing east–west?

b. What is the probability that all of the top five pairs end up playing the same direction?

c. If there are $2n$ pairs, what is the pmf of X = the number among the top n pairs who end up playing east–west? What are $E(X)$ and $V(X)$?

114. A second-stage smog alert has been called in an area of Los Angeles County in which there are 50 industrial firms. An inspector will visit 10 randomly selected firms to check for violations of regulations.

a. If 15 of the firms are actually violating at least one regulation, what is the pmf of the number of firms visited by the inspector that are in violation of at least one regulation?

b. If there are 500 firms in the area, of which 150 are in violation, approximate the pmf of part (a) by a simpler pmf.

c. For X = the number among the 10 visited that are in violation, compute $E(X)$ and $V(X)$ both for the exact pmf and the approximating pmf in part (b).

115. Suppose that $p = P(\text{female birth}) = .5$. A couple wishes to have exactly two female children in their family. They will have children until this condition is fulfilled.

a. What is the probability that the family has x male children?

b. What is the probability that the family has four children?

c. What is the probability that the family has at most four children?

d. How many children would you expect this family to have? How many male children would you expect this family to have?

116. A family decides to have children until it has three children of the same sex. Assuming $P(B) = P(G) = .5$, what is the pmf of X = the number of children in the family?

117. Three brothers and their wives decide to have children until each family has two female children. Let X = the total number of male children born to the brothers. What is $E(X)$, and how does it compare to the expected number of male children born to each brother?

118. Individual A has a red die and B has a green die (both fair). If they each roll until they obtain five "doubles" (11, …, 66), what is the pmf of X = the total number of times a die is rolled? What are $E(X)$ and $\text{SD}(X)$?

119. A shipment of 20 integrated circuits (ICs) arrives at an electronics manufacturing site. The site manager will randomly select 4 ICs and test them to see whether they are faulty. Unknown to the site manager, 5 of these 20 ICs are faulty.

a. Suppose the shipment will be accepted if and only if none of the inspected ICs is faulty. What is the probability this shipment of 20 ICs will be accepted?

b. Now suppose the shipment will be accepted if and only if at most one of the inspected ICs is faulty. What is the probability this shipment of 20 ICs will be accepted?

c. How do your answers to (a) and (b) change if the number of faculty ICs in the shipment is 3 instead of 5? Recalculate (a) and (b) to verify your claim.

120. A carnival game consists of spinning a wheel with 10 slots, nine red and one blue. If you land on the blue slot, you win a prize. Suppose your significant other *really* wants that prize, so you will play until you win.

   a. What is the probability you'll win on the first spin?
   b. What is the probability you'll require exactly 5 spins? At least 5 spins? At most five spins?
   c. What is the expected number of spins required for you to win the prize, and what is the corresponding standard deviation?

121. A kinesiology professor, requiring volunteers for her study, approaches students one by one at a campus hub. She will continue until she acquires 40 volunteers. Suppose that 25% of students are willing to volunteer for the study, that the professor's selections are random, and that the student population is large enough that individual "trials" (asking a student to participate) may be treated as independent.

   a. What is the expected number of students the kinesiology professor will need to ask in order to get 40 volunteers? What is the standard deviation?
   b. Determine the probability that the number of students the kinesiology professor will need to ask is within one standard deviation of the mean.

122. Refer back to the communication system of Example 3.51. Suppose a voice packet can be transmitted a maximum of 10 times; i.e., if the 10th attempt fails, no 11th attempt is made to re-transmit the voice packet. Let $X =$ the number of times a message is transmitted. Assuming each transmission succeeds with probability p, determine the pmf of X. Then obtain an expression for the expected number of times a packet is transmitted.

123. Newton's generalization of the binomial theorem can be used to show that, for any positive integer r,

$$(1 - u)^{-r} = \sum_{k=0}^{\infty} \binom{r + k - 1}{r - 1} u^k$$

Use this to derive the negative binomial mgf presented in this section. Then obtain the mean and variance of a negative binomial rv using this mgf.

124. If X is a negative binomial rv, then the rv $Y = X - r$ is the total number of failures preceding the rth success. (As mentioned in this section, Y is also sometimes called a negative binomial rv.)

   a. Use an argument similar to the one presented in this section to derive the pmf of Y.
   b. Obtain the mgf of Y. [*Hint:* Use the mgf of X and the fact that $Y = X - r$.]
   c. Determine the mean and variance of Y. Are these intuitively consistent with the expressions for $E(X)$ and $V(X)$? Explain.

3.8   Simulation of Discrete Random Variables

Probability calculations for complex systems often depend on the behavior of various random variables. When such calculations are difficult or impossible, simulation is the fallback strategy. In this section, we give a general method for simulating an arbitrary discrete random variable and consider implementations in existing software for simulating common discrete distributions.

Example 3.52 Let X = the amount of memory (GB) in a purchased flash drive, and suppose X has the following pmf:

x	16	32	64	128	256
$p(x)$	.05	.10	.35	.40	.10

We wish to simulate X. Recall from Section 2.6 that we begin with a "standard uniform" random number generator, i.e., a software function that generates evenly distributed numbers in the interval $[0, 1)$. Our goal is to convert these decimals into the values of X with the probabilities specified by its pmf: 5% 16's, 10% 32's, 35% 64's, and so on. To that end, we partition the interval $[0, 1)$ according to these percentages: $[0, .05)$ has probability .05; $[.05, .15)$ has probability .1, since the length of the interval is .1; $[.15, .50)$ has probability $.50 - .15 = .35$; etc. Proceed as follows: given a value u from the RNG,

- If $0 \le u < .05$, assign the value 16 to the variable x.
- If $.05 \le u < .15$, assign $x = 32$.
- If $.15 \le u < .50$, assign $x = 64$.
- If $.50 \le u < .90$, assign $x = 128$.
- If $.90 \le u < 1$, assign $x = 256$.

Repeating this algorithm n times gives n simulated values of X. An R program that implements this algorithm appears in Figure 3.10; it returns a vector, x, containing $n = 10{,}000$ simulated values of the specified distribution.

```
x <- NULL
for (i in 1:10000){
      u=runif(1)
      if (u<.05)
            x[i]<-16
      else if (u<.15)
            x[i]<-32
      else if (u<.50)
            x[i]<-64
      else if (u<.90)
            x[i]<-128
      else
            x[i]<-256
}
```

Figure 3.10 R simulation code

Figure 3.11 (p. 175) shows a graph of the results of executing the above code, in the form of a histogram: the height of each rectangle corresponds to the relative frequency of each x value in the simulation (i.e., the number of times that value occurred, divided by 10,000). The exact pmf of X is superimposed for comparison; as expected, simulation results are similar, but not identical, to the theoretical distribution.

Later in this section, we will present a faster, built-in way to simulate discrete distributions in R. The method introduced above will, however, prove useful in adapting to the case of continuous random variables in Chapter 4.
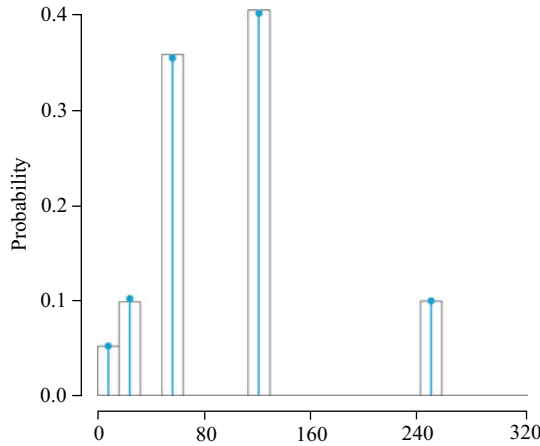
Figure 3.11   Simulation and exact distribution for Example 3.52   ■

In the preceding example, the selected subintervals of [0, 1) were not our only choices—any five intervals with lengths .05, .10, .35, .40, and .10 would produce the desired result. However, those particular five subintervals have one desirable feature: the "cut points" for the intervals (i.e., 0, .05, .15, .50, .90, and 1) are precisely the possible heights of the graph of the cdf, $F(x)$. This permits a geometric interpretation of the algorithm, which can be seen in Figure 3.12. The value u provided by the RNG corresponds to a position on the vertical axis between 0 and 1; we then "invert" the cdf by matching this u-value back to one of the gaps in the graph of $F(x)$, denoted by dashed lines in Figure 3.12. If the gap occurs at horizontal position x, then x is our simulated value of the rv X for that run of the simulation. This is often referred to as the **inverse cdf method** for simulating discrete random variables. The general method is spelled out in the accompanying box.
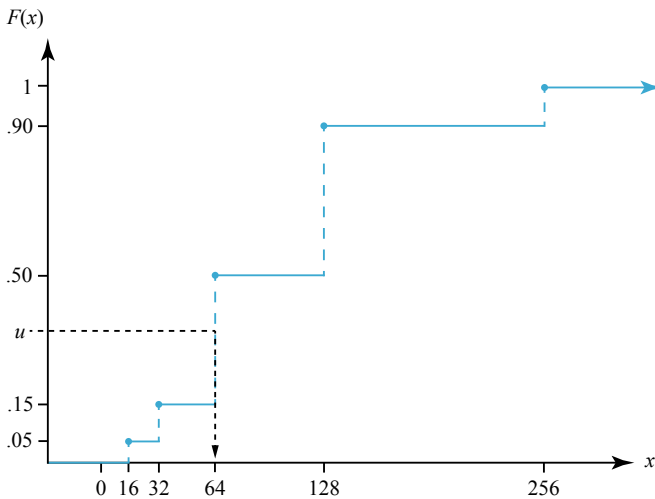
Figure 3.12   The inverse cdf method for Example 3.52

Inverse cdf Method for Simulating Discrete Random Variables	Let X be a discrete random variable taking on values $x_1 < x_2 < \ldots$ with corresponding probabilities p_1, p_2, $\ldots$. Define $F_0 = 0$; $F_1 = F(x_1) = p_1$; $F_2 = F(x_2) = p_1 + p_2$; and, in general, $F_k = F(x_k) = p_1 + \cdots + p_k = F_{k-1} + p_k$. To simulate a value of X, proceed as follows:

    1. Use an RNG to produce a value, u, from [0, 1).
    2. If $F_{k-1} \leq u < F_k$, then assign $x = x_k$.

Example 3.53 (Example 3.52 continued): Suppose the prices for the flash drives, in increasing order of memory size, are \$10, \$15, \$20, \$25, and \$30. If the store sells 80 flash drives in a week, what's the probability they will make a gross profit of at least \$1800?

Let Y = the amount spent on a flash drive, which has the following pmf:

y	10	15	20	25	30
$p(y)$	.05	.10	.35	.40	.10

The gross profit for 80 purchases is the sum of 80 values from this distribution. Let A = {gross profit $\geq$ \$1800}. We can use simulation to estimate $P(A)$, as follows:

    0. Set a counter for the number of times A occurs to zero.
   Repeat n times:
    1. Simulate 80 values $y_1, \ldots, y_{80}$ from the above pmf (using, e.g., an inverse cdf program similar to the one displayed in Figure 3.10).
    2. Compute the week's gross profit, $g = y_1 + \cdots + y_{80}$.
    3. If $g \geq 1800$, add 1 to the count of occurrences for A.

    Once the n runs are complete, then $\hat{P}(A)$ = (count of the occurrences of A)/n.

Figure 3.13 shows the resulting values of g for $n = 10{,}000$ simulations in R. In effect, our program is simulating a random variable $G = Y_1 + \cdots + Y_{80}$ whose pmf is not known (in light of all the possible G values, it would not be worthwhile to attempt to determine its pmf analytically). The highlighted bars in Figure 3.13 correspond to g values of at least \$1800; in our simulation, such values occurred 1940 times. Thus, $\hat{P}(A) = 1940/10{,}000 = .194$, with an estimated standard error of $\sqrt{.194(1 - .194)/10{,}000} = .004$.
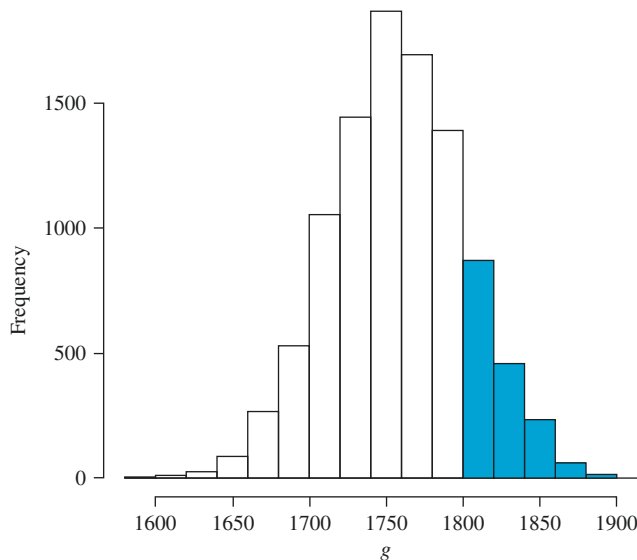
Figure 3.13 Simulated distribution of weekly gross profit for Example 3.53   ∎

Simulations Implemented in R

Earlier in this section, we presented the inverse cdf method as a general way to simulate discrete distributions applicable in any software. In fact, one can simulate generic discrete rvs in R by clever use of the built-in `sample` function. We saw this function in the context of probability simulation in Chapter 2. The `sample` function is designed to generate a random sample from any selected set of values (even including text values, if desired); the "clever" part is that it can accommodate a set of weights. The following short example illustrates their use.

To simulate, say, 35 values from the pmf in Example 3.53, one can use the following code in R: `sample(c(10,15,20,25,30),35,TRUE,c(.05,.10,.35,.40,.10))`. The function takes four arguments: the list of y values, the desired number of simulated values (the "sample size"), whether to sample with replacement (here, `TRUE`), and the list of probabilities in the same order as the y values.

Thanks to the ubiquity of the binomial, Poisson, and other distributions in probability modeling, many software packages have built-in tools for simulating values from these distributions. Table 3.6 summarizes the relevant functions in R; the input argument *size* refers to the desired number of simulated values of the distribution.

Table 3.6   Functions to simulate major discrete distributions in R

Distribution	R code
Binomial	`rbinom(`*size*, n, p`)`
Poisson	`rpois(`*size*, μ`)`
Hypergeometric	`rhyper(`*size*, M, N–M, n`)`
Negative binomial	`rnbinom(`*size*, r, p`)`

A word of warning (really, a reminder) about the way software treats the negative binomial distribution: R defines a negative binomial rv as the number of failures preceding the rth success, which differs from our definition. Assuming you want to simulate the number of *trials* required to achieve r successes, execute the code in the last line of Table 3.6 and then add r to each value.

Example 3.54 The number of customers shipping express mail packages at a certain store during any particular hour of the day is a Poisson rv with mean 5. Each such customer has 1, 2, 3, or 4 packages with probabilities .4, .3, .2, and .1, respectively. Let's carry out a simulation to estimate the probability that at most 10 packages are shipped during any particular hour.

Define an event A = {at most 10 packages shipped in an hour}. Our simulation to estimate $P(A)$ proceeds as follows.

0. Set a counter for the number of times A occurs to zero.
   Repeat n times:
1. Simulate the number of customers in an hour, X, which is Poisson with $\mu = 5$.
2. For each of the X customers, simulate the number of packages shipped according to the pmf above.
3. If the total number of packages shipped is at most 10, add 1 to the counter for A.
   R code to implement this simulation appear in Figure 3.14.

```
A <- 0
for (i in 1:10000){
      x<-rpois(1,5)
      packages <- sample(c(1,2,3,4),x,
                  TRUE,c(.4,.3,.2,.1))
      if (sum(packages)<=10){
            A<-A+1
      }
}
```

Figure 3.14 R simulation code for Example 3.54

In R, 10,000 simulations resulted in 10 or fewer packages 5752 times, for an estimated probability of $\hat{P}(A) = .5752$, with an estimated standard error of $\sqrt{.5752(1 - .5752)/10,000} = .0049$.   ∎

Simulation Mean, Standard Deviation, and Precision

In Section 2.6 and in the preceding examples, we used simulation to estimate the probability of an event. But consider the "gross profit" variable in Example 3.53: since we have 10,000 simulated values of this variable, we should be able to estimate its mean and its standard deviation. In general, suppose we have simulated n values $x_1, \ldots, x_n$ of a random variable X. Then, not surprisingly, we estimate μ_X and σ_X with the *sample* mean $\bar{x}$ and *sample* standard deviation s, respectively, of the n simulated values.

In Section 2.6, we introduced the standard error of an estimated probability, which quantifies the precision of a simulation result $\hat{P}(A)$ as an estimate of a "true" probability $P(A)$. By analogy, it is possible to quantify the amount by which a sample mean, $\bar{x}$, will generally differ from the corresponding expected value μ. For n simulated values of a random variable, with sample standard deviation s, the **(estimated) standard error of the mean** is

$$\text{Estimated standard error of the mean} = \frac{s}{\sqrt{n}} \qquad (3.22)$$

Expression (3.22) will be derived in Chapter 6. As with an estimated probability, (3.22) indicates that the precision of $\bar{x}$ increases (i.e., its standard error *decreases*) as n increases, but not very quickly. To increase the precision of $\bar{x}$ as an estimate of μ by a factor of 10 (one decimal place) requires increasing the number of simulation runs, n, by a factor of 100. Unfortunately, there is no general formula for the standard error of s as an estimate of σ.

Example 3.55 (Example 3.54 continued) The 10,000 simulated values of the random variable G, which we denote by $g_1, \ldots, g_{10000}$, are displayed in the histogram in Figure 3.13. From these simulated values, we can estimate both the expected value and standard deviation of G:

$$\hat{\mu}_G = \bar{g} = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} g_i = 1759.62$$

$$\hat{\sigma}_G = s = \sqrt{\frac{1}{10{,}000 - 1} \sum_{i=1}^{10{,}000} (g_i - \bar{g})^2} = \sqrt{\frac{1}{9999} \sum_{i=1}^{10{,}000} (g_i - 1759.62)^2} = 43.50$$

We estimate that the average weekly gross profit from flash drive sales is \$1759.62, with a standard deviation of \$43.50.

Applying (3.22), the (estimated) standard error of $\bar{g}$ is $s/\sqrt{n} = 43.50/\sqrt{10{,}000} = 0.435$. If 10,000 runs are used to simulate G, it's estimated that the resulting sample mean will differ from $E(G)$ by roughly 0.435. (In contrast, the sample standard deviation, s, estimates that the gross profit for a single week—i.e., a single observation g—typically differs from $E(G)$ by about \$43.50.)

In Chapter 5, we will see how the expected value and variance of random variables like G, that are sums of a fixed number of other rvs, can be obtained analytically. ∎

Example 3.56 The "help desk" at a university's computer center receives both hardware and software queries. Let X and Y be the number of hardware and software queries, respectively, in a given day. Each can be modeled by a Poisson distribution with mean 20. Because computer center employees need to be allocated efficiently, of interest is the *difference* between the sizes of the two queues: $D = |X - Y|$. Let's use simulation to estimate (1) the probability the queue sizes differ by more than 5; (2) the expected difference; (3) the standard deviation of the difference.

Figure 3.15 shows R code to simulate this process. The code exploits the built-in Poisson simulator, as well as the fact that 10,000 simulated values may be called simultaneously.

```
X<-rpois(10000,20)
Y<-rpois(10000,20)
D<-abs(X-Y)
sum((D>5))
mean(D)
sd(D)
```

Figure 3.15 R simulation code for Example 3.56

The line `sum((D>5))` performs two operations: first, `(D>5)` determines if each simulated d value exceeds 5, returning a vector of logical bits; second, `sum()` tallies the "success" bits (1's or TRUEs) and gives a count of the number of times the event $\{D > 5\}$ occurred in the 10,000 simulations. The results from one run were

$$\hat{P}(D > 5) = \frac{3843}{10{,}000} = .3843 \quad \hat{\mu}_D = \bar{d} = 5.0380 \quad \hat{\sigma}_D = s = 3.8436$$

A histogram of the simulated values of D appears in Figure 3.16.
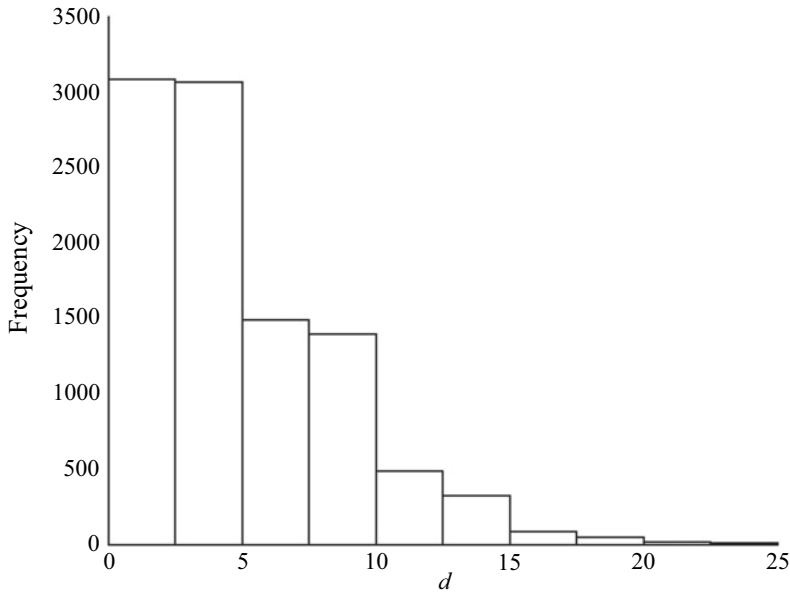
Figure 3.16 Simulation histogram of D in Example 3.56   ∎

Section 3.8 Exercises (125–137)

125. Consider the pmf given in Exercise 29 for the random variable Y = the number of moving violations for which the a randomly selected insured individual was cited during the last 3 years. Write a program to simulate this random variable, then use your simulation to estimate $E(Y)$ and $SD(Y)$. How do these compare to the exact values of $E(Y)$ and $SD(Y)$?

126. Consider the pmf given in Exercise 31 for the random variable X = capacity of a purchased freezer. Write a program to simulate this random variable, then use your simulation to estimate both $E(X)$ and $SD(X)$. How do these compare to the exact values of $E(X)$ and $SD(X)$?

127. Suppose person after person is tested for the presence of a certain characteristic. The probability that any individual tests positive is .75. Let X = the number of people who must be tested to obtain five *consecutive* positive test results. Use simulation to estimate $P(X \leq 25)$.

128. *The matching problem*. Suppose that N items labeled 1, 2, …, N are shuffled so that they are in random order. Of interest is how many of these will be in their "correct" positions (e.g., item #5 situated at the 5th position in the sequence, etc.) after shuffling.

   a. Write program that simulates a permutation of the numbers 1 to N and then records the value of the variable X = number of items in the correct position.
   b. Set $N = 5$ in your program, and use at least 10,000 simulations to estimate $E(X)$, the expected number of items in the correct position.
   c. Set $N = 52$ in your program (as if you were shuffling a deck of cards), and use at least 10,000 simulations to estimate $E(X)$. What do you discover? Is this surprising?

129. Exercise 101 of Chapter 2 referred to a multiple-choice exam in which 10 of the questions have two options, 13 have three options, 13 have four options, and the other 4 have five options. Let X = the number of

questions a student gets right, assuming s/he is completely guessing.

a. Write a program to simulate X, and use your program to estimate the mean and standard deviation of X.

b. Estimate the probability a student will score at least one standard deviation above the mean.

130. Example 3.53 of this section considered the gross profit G resulting from selling flash drives to 80 customers per week. Of course, it isn't realistic for the number of customers to remain fixed from week to week. So, instead, imagine the number of customers buying flash drives in a week follows a Poisson distribution with mean 80, and that the amount paid by each customer follows the distribution for Y provided in that example. Write a program to simulate the random variable G, and use your simulation to estimate

a. The probability that weekly gross sales are at least $1800.

b. The mean of G.

c. The standard deviation of G.

131. Exercise 19 investigated Benford's law, a discrete distribution with pmf given by $p(x) = \log_{10}((x + 1)/x)$ for $x = 1, 2, \ldots, 9$. Use the inverse cdf method to write a program that simulates the Benford's law distribution. Then use your program to estimate the expected value and variance of this distribution.

132. Recall that a geometric rv has pmf $p(x) = p(1-p)^{x-1}$ for $x = 1, 2, 3, \ldots$. In Example 3.12, it was shown that the cdf of this distribution is $F(x) = 1 - (1 - p)^x$ for positive integers x.

a. Write a program that implements the inverse cdf method to simulate a geometric distribution. Your program should have as inputs the numerical value of p and the desired sample size.

b. Use your program to simulate 10,000 values from a geometric rv X with

$p = .85$. From these values, estimate each of the following: $P(X \leq 2)$, $E(X)$, SD(X). How do these compare to the corresponding exact values?

133. Tickets for a particular flight are $250 apiece. The plane seats 120 passengers, but the airline will knowingly overbook (i.e., sell more than 120 tickets), because not every paid passenger shows up. Let t denote the number of tickets the airline sells for this flight, and assume the number of passengers that actually show up for the flight, X, follows a Bin(t, .85) distribution. Let $B =$ the number of paid passengers who show up at the airport but are denied a seat on the plane, so $B = X - 120$ if $X > 120$ and $B = 0$ otherwise. If the airline must compensate these passengers with $500 apiece, then the profit the airline makes on this flight is $250t - 500B$. (Notice that t is fixed, but B is random.)

a. Write a program to simulate this scenario. Specifically, your program should take in t as an input and return many values of the profit variable $250t - 500B$, where B is described above.

b. The airline wishes to determine the optimal value of t, i.e., the number of tickets to sell that will maximize their expected profit. Run your program for $t = 140, 141, \ldots, 150$, and record the average profit from many runs under each of these settings. What value of t appears to return the largest value? [*Note:* If a clear winner does not emerge, you might need to increase the number of runs for each t value!]

134. Imagine the following simple game: flip a fair coin repeatedly, winning $1 for every head and losing $1 for every tail. Your net winnings will potentially oscillate between positive and negative numbers as play continues. How many times do you think net winnings will *change signs* in, say, 1000 coin flips? 5000 flips?

a.   Let X = the number of sign changes in 1000 coin flips. Write a program to simulate X, and use your program to estimate the probability of at least 10 sign changes.

b.   Use your program to estimate both $E(X)$ and SD(X). Does your estimate for $E(X)$ match your intuition for the number of sign changes?

c.   Repeat parts (a)–(b) with 5000 flips.

135.   Exercise 40 describes the game Plinko from *The Price is Right*. Each contestant drops between one and 5 chips down the Plinko board, depending on how well s/he prices several small items. Suppose the random variable C = number of chips earned by a contestant has the following distribution:

c	1	2	3	4	5
$p(c)$	.03	.15	.35	.34	.13

The winnings from each chip follow the distribution presented in Exercise 40. Write a program to simulate Plinko; you will need to consider both the number of chips a contestant earns and how much money is won on each of those chips. Use your simulation to estimate the answers to the following questions:

a.   What is the probability a contestant wins more than $11,000?

b.   What is a contestant's expected winnings?

c.   What is the corresponding standard deviation?

d.   In fact, a player gets one Plinko chip for free and can earn the other four by guessing the prices of small items (waffle irons, alarm clocks, etc.). Assume the player has a 50–50 chance of getting each price correct, so we may write $C = 1 + R$, where $R \sim$ Bin(4, .5). Use this revised model for C to estimate the answers to (a)–(c).

136.   Recall the *Coupon Collector's Problem* described in Exercise 106 of Chapter 2. Let X = the number of cereal boxes purchased in order to obtain all 10 coupons.

a.   Use a simulation program to estimate $E(X)$ and SD(X). Also compute the estimated standard error of your answer.

b.   Repeat (a) with 20 coupons required instead of 10. Does it appear to take roughly twice as long to collect 20 coupons as 10? More than twice as long? Less?

137.   A small high school holds its graduation ceremony in the gym. Because of seating constraints, students are limited to a maximum of four tickets to graduation for family and friends. Suppose 30% of students want four tickets, 25% want three, 25% want two, 15% want one, and 5% want none.

a.   Write a simulation for 150 graduates requesting tickets, where students' requests follow the distribution described above. In particular, keep track of the variable T = the total number of tickets requested by these 150 students.

b.   The gym can seat a maximum of 410 guests. Based on your simulation, estimate the probability that all students' requests can be accommodated.

Supplementary Exercises: (138–169)

138.   Consider a deck consisting of seven cards, marked 1, 2, …, 7. Three of these cards are selected at random. Define a rv W by W = the sum of the resulting numbers, and compute the pmf of W. Then compute μ and σ^2. [*Hint*: Consider outcomes as unordered, so that (1, 3, 7) and (3, 1, 7) are not different outcomes. Then there are 35 outcomes, and they can be listed.] (This type of rv actually arises in connection with *Wilcoxon's rank-sum test*, in which there is

an x sample and a y sample and W is the sum of the ranks of the x's in the combined sample.)

139. After shuffling a deck of 52 cards, a dealer deals out 5. Let $X =$ the number of suits represented in the five-card hand.

a. Show that the pmf of X is

x	1	2	3	4
$p(x)$	.002	.146	.588	.264

[*Hint*: $p(1) = 4P(\text{all spades})$, $p(2) = 6P(\text{only spades and hearts with at least one of each})$, and $p(4) = 4P(2 \text{ spades} \cap \text{one of each other suit})$.]

b. Compute μ, σ^2, and σ.

140. Let X be a rv with mean μ. Show that $E(X^2) \geq \mu^2$, and that $E(X^2) > \mu^2$ unless X is a constant. [*Hint*: Consider variance.]

141. Of all customers purchasing automatic garage-door openers, 75% purchase a chain-driven model. Let $X =$ the number among the next 15 purchasers who select the chain-driven model.

a. What is the pmf of X?
b. Compute $P(X > 10)$.
c. Compute $P(6 \leq X \leq 10)$.
d. Compute μ and σ^2.
e. If the store currently has in stock 10 chain-driven models and 8 shaft-driven models, what is the probability that the requests of these 15 customers can all be met from existing stock?

142. A friend recently planned a camping trip. He had two flashlights, one that required a single 6-V battery and another that used two size-D batteries. He had previously packed two 6-V and four size-D batteries in his camper. Suppose the probability that any particular battery works is p and that batteries work or fail independently of one another. Our friend wants to take just one flashlight. For what values of p should he take the 6-V flashlight?

143. Binary data is transmitted over a noisy communication channel. The probability that a received binary digit is in error due to channel noise is 0.05. Assume that such errors occur independently within the bit stream.

a. What is the probability that the 3rd error occurs on the 50th transmitted bit?
b. On average, how many bits will be transmitted correctly *before* the first error?
c. Consider a 32-bit "word." What is the probability of exactly 2 errors in this word?
d. Consider the next 10,000 bits. What approximating model could we use for $X =$ the number of errors in these 10,000 bits? Give the name of the model and the value(s) of the parameter(s).

144. A manufacturer of flashlight batteries wishes to control the quality of its product by rejecting any lot in which the proportion of batteries having unacceptable voltage appears to be too high. To this end, out of each large lot (10,000 batteries), 25 will be selected and tested. If at least 5 of these generate an unacceptable voltage, the entire lot will be rejected. What is the probability that a lot will be rejected if

a. Five percent of the batteries in the lot have unacceptable voltages?
b. Ten percent of the batteries in the lot have unacceptable voltages?
c. Twenty percent of the batteries in the lot have unacceptable voltages?
d. What would happen to the probabilities in parts (a)–(c) if the critical rejection number were increased from 5 to 6?

145. Of the people passing through an airport metal detector, .5% activate it; let $X =$ the number among a randomly selected group of 500 who activate the detector.

a. What is the (approximate) pmf of X?
b. Compute $P(X = 5)$.
c. Compute $P(5 \leq X)$.

146. An educational consulting firm is trying to decide whether high school students who have never before used a hand-held calculator can solve a certain type of problem more easily with a calculator that uses reverse Polish logic or one that does not use this logic. A sample of 25 students is selected and allowed to practice on both calculators. Then each student is asked to work one problem on the reverse Polish calculator and a similar problem on the other. Let $p = P(S)$, where S indicates that a student worked the problem more quickly using reverse Polish logic than without, and let X = number of S's.

     a. If $p = .5$, what is $P(7 \leq X \leq 18)$?
     b. If $p = .8$, what is $P(7 \leq X \leq 18)$?
     c. If the claim that $p = .5$ is to be rejected when either $X \leq 7$ or $X \geq 18$, what is the probability of rejecting the claim when it is actually correct?
     d. If the decision to reject the claim $p = .5$ is made as in part (c), what is the probability that the claim is not rejected when $p = .6$? When $p = .8$?
     e. What decision rule would you choose for rejecting the claim $p = .5$ if you wanted the probability in part (c) to be at most .01?

147. Consider a disease whose presence can be identified by carrying out a blood test. Let p denote the probability that a randomly selected individual has the disease. Suppose n individuals are independently selected for testing. One way to proceed is to carry out a separate test on each of the n blood samples. A potentially more economical approach, group testing, was introduced during World War II to identify syphilitic men among army inductees. First, take a part of each blood sample, combine these specimens, and carry out a single test. If no one has the disease, the result will be negative, and only the one test is required. If at least one individual is diseased, the test on the combined sample will yield a positive result, in which case the n individual tests are then carried out. If $p = .1$ and $n = 3$, what is the expected number of tests using this procedure? What is the expected number when $n = 5$? [The article "Random Multiple-Access Communication and Group Testing" (*IEEE Trans. Commun.* 1984: 769–774) applied these ideas to a communication system in which the dichotomy was active/idle user rather than diseased/nondiseased.]

148. Let p_1 denote the probability that any particular code symbol is erroneously transmitted through a communication system. Assume that on different symbols, errors occur independently of one another. Suppose also that with probability p_2 an erroneous symbol is corrected upon receipt. Let X denote the number of correct symbols in a message block consisting of n symbols (after the correction process has ended). What is the probability distribution of X?

149. The purchaser of a power-generating unit requires c consecutive successful start-ups before the unit will be accepted. Assume that the outcomes of individual start-ups are independent of one another. Let p denote the probability that any particular start-up is successful. The random variable of interest is X = the number of start-ups that must be made prior to acceptance. Give the pmf of X for the case $c = 2$. If $p = .9$, what is $P(X \leq 8)$? [*Hint*: For $x \geq 5$, express $p(x)$ "recursively" in terms of the pmf evaluated at the smaller values $x - 3, x - 4, \ldots, 2$.] (This problem was suggested by the article "Evaluation of a Start-Up Demonstration Test," *J. Qual. Tech.* 1983: 103–106.)

150. A plan for an executive travelers' club has been developed by an airline on the premise that 10% of its current customers would qualify for membership.

     a. Assuming the validity of this premise, among 25 randomly selected current customers, what is the probability that between 2 and 6 (inclusive) qualify for membership?

b. Again assuming the validity of the premise, what are the expected number of customers who qualify and the standard deviation of the number who qualify in a random sample of 100 current customers?

c. Let X denote the number in a random sample of 25 current customers who qualify for membership. Consider rejecting the company's premise in favor of the claim that $p > .10$ if $x \geq 7$. What is the probability that the company's premise is rejected when it is actually valid?

d. Refer to the decision rule introduced in part (c). What is the probability that the company's premise is not rejected even though $p = .20$ (i.e., 20% qualify)?

151. Forty percent of seeds from maize (modern-day corn) ears carry single spikelets, and the other 60% carry paired spikelets. A seed with single spikelets will produce an ear with single spikelets 29% of the time, whereas a seed with paired spikelets will produce an ear with single spikelets 26% of the time. Consider randomly selecting ten seeds.

a. What is the probability that exactly five of these seeds carry a single spikelet and produce an ear with a single spikelet?

b. What is the probability that exactly five of the ears produced by these seeds have single spikelets? What is the probability that at most five ears have single spikelets?

152. A trial has just resulted in a hung jury because eight members of the jury were in favor of a guilty verdict and the other four were for acquittal. If the jurors leave the jury room in random order and each of the first four leaving the room is accosted by a reporter in quest of an interview, what is the pmf of X = the number of jurors favoring acquittal among those interviewed? How many of those favoring acquittal do you expect to be interviewed?

153. A reservation service employs five information operators who receive requests for information independently of one another, each according to a Poisson process with rate $\lambda = 2/\text{min}$.

a. What is the probability that during a given 1-min period, the first operator receives no requests?

b. What is the probability that during a given 1-min period, exactly four of the five operators receive no requests?

c. Write an expression for the probability that during a given 1-min period, all of the operators receive exactly the same number of requests.

154. Grasshoppers are distributed at random in a large field according to a Poisson distribution with parameter $\lambda = 2$ per square yard. How large should the radius R of a circular sampling region be taken so that the probability of finding at least one in the region equals .99?

155. A newsstand has ordered five copies of a certain issue of a photography magazine. Let X = the number of individuals who come in to purchase this magazine. If X has a Poisson distribution with parameter $\mu = 4$, what is the expected number of copies that are sold?

156. Individuals A and B begin to play a sequence of chess games. Let $S = \{A \text{ wins a game}\}$, and suppose that outcomes of successive games are independent with $P(S) = p$ and $P(F) = 1 - p$ (they never draw). They will play until one of them wins ten games. Let X = the number of games played (with possible values 10, 11, …, 19).

a. For $x = 10$, 11, …, 19, obtain an expression for $p(x) = P(X = x)$.

b. If a draw is possible, with $p = P(S)$, $q = P(F)$, $1 - p - q = P(\text{draw})$, what are the possible values of X? What is $P(20 \leq X)$? [*Hint*: $P(20 \leq X) = 1 - P(X < 20)$.]

157. A test for the presence of a disease has probability .20 of giving a false-positive reading (indicating that an individual has the disease when this is not the case) and probability .10 of giving a false-negative result. Suppose that ten individuals are tested, five of whom have the disease and five of whom do not. Let X = the number of positive readings that result.

   a. Does X have a binomial distribution? Explain your reasoning.
   b. What is the probability that exactly three of the ten test results are positive?

158. The generalized negative binomial pmf, in which r is not necessarily an integer, is

$$nb(x; r, p) = k(r, x) \times p^r (1 - p)^x$$
$$x = 0, 1, 2, \ldots$$

   where

$$k(r, x) = \begin{cases} \dfrac{(x+r-1)(x+r-2)\cdots(x+r-x)}{x!} & x = 1, 2, \ldots \\ 1 & x = 0 \end{cases}$$

   Let X, the number of plants of a certain species found in a particular region, have this distribution with $p = .3$ and $r = 2.5$. What is $P(X = 4)$? What is the probability that at least one plant is found?

159. A small publisher employs two typesetters. The number of errors (in one book) made by the first typesetter has a Poisson distribution mean μ_1, the number of errors made by the second typesetter has a Poisson distribution with mean μ_2, and each typesetter works on the same number of books. Then if one such book is randomly selected, the function

$$p(x; \mu_1, \mu_2) = .5 e^{-\mu_1} \frac{\mu_1^x}{x!} + .5 e^{-\mu_2} \frac{\mu_2^x}{x!}$$
$$x = 0, 1, 2, \ldots$$

   gives the pmf of X = the number of errors in the selected book.

   a. Verify that $p(x; \mu_1, \mu_2)$ is a legitimate pmf (≥ 0 and sums to 1).
   b. What is the expected number of errors in the selected book?
   c. What is the standard deviation of the number of errors in the selected book?
   d. How does the pmf change if the first typesetter works on 60% of all such books and the second typesetter works on the other 40%?

160. The *mode* of a discrete random variable X with pmf $p(x)$ is that value x^* for which $p(x)$ is largest (the most probable x value).

   a. Let $X \sim \text{Bin}(n, p)$. By considering the ratio $b(x + 1; n, p)/b(x; n, p)$, show that $b(x; n, p)$ increases with x as long as $x < np - (1 - p)$. Conclude that the mode x^* is the integer satisfying $(n + 1)p - 1 \leq x^* \leq (n + 1)p$.
   b. Show that if X has a Poisson distribution with parameter μ, the mode is the largest integer less than μ. If μ is an integer, show that both $\mu - 1$ and μ are modes.

161. For a particular insurance policy the number of claims by a policy holder in 5 years is Poisson distributed. If the filing of one claim is four times as likely as the filing of two claims, find the expected number of claims.

162. If X is a hypergeometric rv, show directly from the definition that $E(X) = nM/N$ (consider only the case $n < M$). [*Hint*: Factor nM/N out of the sum for $E(X)$, and show that the terms inside the sum are of the form $h(y; n - 1, M - 1, N - 1)$, where $y = x - 1$.]

163. Use the fact that

$$\sum_{\text{all } x} (x - \mu)^2 p(x) \geq \sum_{x : |x - \mu| \geq k\sigma} (x - \mu)^2 p(x)$$

   to prove Chebyshev's inequality, given in Exercise 45 of this chapter.

164. The simple Poisson process of Section 3.6 is characterized by a constant rate λ at which events occur per unit time. A generalization is to suppose that the probability of exactly one event occurring in the interval $(t, t + \Delta t)$ is $\lambda(t) \cdot \Delta t + o(\Delta t)$ for some function $\lambda(t)$. It can then be shown that the number of events occurring during an interval $[t_1, t_2]$ has a Poisson distribution with parameter

$$\mu = \int_{t_1}^{t_2} \lambda(t)dt$$

The occurrence of events over time in this situation is called a *nonhomogeneous Poisson process*. The article "Inference Based on Retrospective Ascertainment," *J. Amer. Statist. Assoc.* 1989: 360–372, considers the intensity function

$$\lambda(t) = e^{a + bt}$$

as appropriate for events involving transmission of HIV via blood transfusions. Suppose that $a = 2$ and $b = .6$ (close to values suggested in the paper), with time in years.

a. What is the expected number of events in the interval [0, 4]? In [2, 6]?

b. What is the probability that at most 15 events occur in the interval [0, .9907]?

165. Suppose a store sells two different coffee makers of a particular brand, a basic model selling for \$30 and a fancy one selling for \$50. Let X denote the number of people among the next 25 purchasing this brand who choose the more expensive model. Then $h(X)$ = revenue = $50X + 30(25 - X)$ = $20X + 750$, a linear function. If the choices are independent and have the same probability, then how is X distributed? Find the mean and standard deviation of $h(X)$. Explain why the choices might not be independent with the same probability.

166. Let X be a discrete rv with possible values 0, 1, 2,. .. or some subset of these. The function $\psi(s) = E(s^X) = \sum_{x=0}^{\infty} s^x \cdot p(x)$ is called the **probability generating function** (pgf) of X.

a. Suppose X is the number of children born to a family, and $p(0) = .2$, $p(1) = .5$, and $p(2) = .3$. Determine the pgf of X.

b. Determine the pgf when X has a Poisson distribution with parameter μ.

c. Show that $\psi(1) = 1$.

d. Show that $\psi'(0) = p(1)$. (You'll need to assume that the derivative can be brought inside the summation, which is justified.) What results from taking the second derivative with respect to s and evaluating at s = 0? The third derivative? Explain how successive differentiation of $\psi(s)$ and evaluation at $s = 0$ "generates the probabilities in the distribution." Use this to recapture the probabilities of (a) from the pgf. [Note: This shows that the pgf contains all the information about the distribution—knowing $\psi(s)$ is equivalent to knowing $p(x)$.]

167. Three couples and two single individuals have been invited to a dinner party. Assume independence of arrivals to the party, and suppose that the probability of any particular individual or any particular couple arriving late is .4 (the two members of a couple arrive together). Let X = the number of people who show up late for the party. Determine the pmf of X.

168. Consider a sequence of identical and independent trials, each of which will be a success S or failure F. Let $p = P(S)$ and $q = P(F)$.

a. Define a random variable X as the number of trials necessary to obtain the first S, a geometric random variable. Here is an alternative approach to determining $E(X)$. Just as $P(B) = P(B|A)P(A) + P(B|A')P(A')$, it can be shown that

$$E(X) = E(X|A)P(A) + E(X|A')P(A')$$

where $E(X|A)$ denotes the expected value of X given that the event A has occurred. Now let $A = \{S \text{ on 1st trial}\}$. Show again that $E(X) = 1/p$. [*Hint*: Denote $E(X)$ by μ. Then given that the first trial is a failure, one trial has been performed and, starting from the second trial, we are still looking for the first S. This implies that $E(X|A') = E(X|F) = 1 + \mu$.]

b. The expected value property in (a) can be extended to any partition $A_1, A_2, \ldots, A_k$ of the sample space:

$$E(X) = E(X|A_1) \cdot P(A_1) + \\ E(X|A_2) \cdot P(A_2) + \cdots + \\ E(X|A_k) \cdot P(A_k)$$

Now let Y = the number of trials necessary to obtain two *consecutive* S's. It is not possible to determine $E(Y)$ directly from the definition of expected value, because there is no formula for the pmf of Y; the complication is the word *consecutive*. Use the weighted average formula to determine $E(Y)$. [*Hint*: Consider the partition with $k = 3$ and $A_1 = \{F\}$, $A_2 = \{SS\}, A_3 = \{SF\}$.]

169. For a discrete rv X taking values in $\{0, 1, 2, 3, \ldots\}$, we shall derive the following alternative formula for the mean:

$$\mu_X = \sum_{x=0}^{\infty} [1 - F(x)]$$

a. Suppose for now the range of X is $\{0, 1, \ldots, N\}$ for some positive integer N. By re-grouping terms, show that

$$\sum_{x=0}^{N} [x \cdot p(x)] = p(1) + p(2) + p(3) + \cdots + p(N) \\ + p(2) + p(3) + \cdots + p(N) \\ + p(3) + \cdots + p(N) \\ \vdots \\ + p(N)$$

b. Re-write each row in the above expression in terms of the cdf of X, and use this to establish that

$$\sum_{x=0}^{N} [x \cdot p(x)] = \sum_{x=0}^{N-1} [1 - F(x)]$$

c. Let $N \to \infty$ in part (b) to establish the desired result, and explain why the resulting formula works even if the maximum value of X is finite. [*Hint:* If the largest possible value of X is N, what does $1 - F(x)$ equal for $x \geq N$?] (This derivation also implies that a discrete rv X has a finite mean iff the series $\sum [1 - F(x)]$ converges.)

d. Let X have a geometric distribution with parameter p. Use the cdf of X and the alternative mean formula just derived to determine μ_X.

Continuous Random Variables and Probability Distributions

4

Introduction

As mentioned at the beginning of Chapter 3, the two important types of random variables are discrete and continuous. In this chapter, we study the second general type of random variable that arises in many applied problems. Sections 4.1 and 4.2 present the basic definitions and properties of continuous random variables, their probability distributions, and their various expected values. In Section 4.3, we study in detail the normal distribution, arguably the most important and useful in probability and statistics. Sections 4.4 and 4.5 discuss some other continuous distributions that are often used in applied work. In Section 4.6, we introduce a method for assessing whether given sample data is consistent with a specified distribution. Section 4.7 presents methods for obtaining the distribution of a rv Y from the distribution of X when the two are related by some equation $Y = g(X)$. The last section is dedicated to the simulation of continuous rvs.

4.1 Probability Density Functions and Cumulative Distribution Functions

A discrete random variable (rv) is one whose possible values either constitute a finite set or else can be listed in an infinite sequence (a list in which there is a first element, a second element, etc.). A random variable whose set of possible values is an entire interval of numbers is not discrete.

Recall from Chapter 3 that a random variable X is continuous if (1) possible values comprise either a single interval on the number line (for some $A < B$, any number x between A and B is a possible value) or a union of disjoint intervals, and (2) $P(X = c) = 0$ for any number c that is a possible value of X.

Example 4.1 If in the study of the ecology of a lake, we make depth measurements at randomly chosen locations, then X = the depth at such a location is a continuous rv. Here A is the minimum depth in the region being sampled, and B is the maximum depth. ∎

Example 4.2 If a chemical compound is randomly selected and its pH X is determined, then X is a continuous rv because any pH value between 0 and 14 is possible. If more is known about the compound selected for analysis, then the set of possible values might be a subinterval of [0, 14], such as $5.5 \leq x \leq 6.5$, but X would still be continuous. ∎

Example 4.3 Let X represent the amount of time a randomly selected customer spends waiting for a haircut before his/her haircut commences. Your first thought might be that X is a continuous random variable, since a measurement is required to determine its value. However, there are customers lucky enough to have no wait whatsoever before climbing into the barber's chair. So it must be the case that $P(X = 0) > 0$. Conditional on no chairs being empty, though, the waiting time will be continuous since X could then assume any value between some minimum possible time A and a maximum possible time B. This random variable is neither purely discrete nor purely continuous but instead is a mixture of the two types. ∎

One might argue that although in principle variables such as height, weight, and temperature are continuous, in practice the limitations of our measuring instruments restrict us to a discrete (though sometimes very finely subdivided) world. However, continuous models often approximate real-world situations very well, and continuous mathematics (the calculus) is frequently easier to work with than the mathematics of discrete variables and distributions.

Probability Distributions for Continuous Variables

Suppose the variable X of interest is the depth of a lake at a randomly chosen point on the surface. Let M = the maximum depth (in meters), so that any number in the interval $[0, M]$ is a possible value of X. If we "discretize" X by measuring depth to the nearest meter, then possible values are nonnegative integers less than or equal to M. The resulting discrete distribution of depth can be pictured using a probability histogram. If we draw the histogram so that the area of the rectangle above any possible integer k is the proportion of the lake whose depth is (to the nearest meter) k, then the total area of all rectangles is 1. A possible histogram appears in Figure 4.1a.

If depth is measured much more accurately and the same measurement axis as in Figure 4.1a is used, each rectangle in the resulting probability histogram is much narrower, although the total area of all rectangles is still 1. A possible histogram is pictured in Figure 4.1b; it has a much smoother appearance than that of Figure 4.1a. If we continue in this way to measure depth more and more finely, the resulting sequence of histograms approaches a smooth curve, as pictured in Figure 4.1c. Because for each histogram the total area of all rectangles equals 1, the total area under the smooth curve is also 1. The probability that the depth at a randomly chosen point is between a and b is just the area under the smooth curve between a and b. It is exactly a smooth curve of this type that specifies a continuous probability distribution.
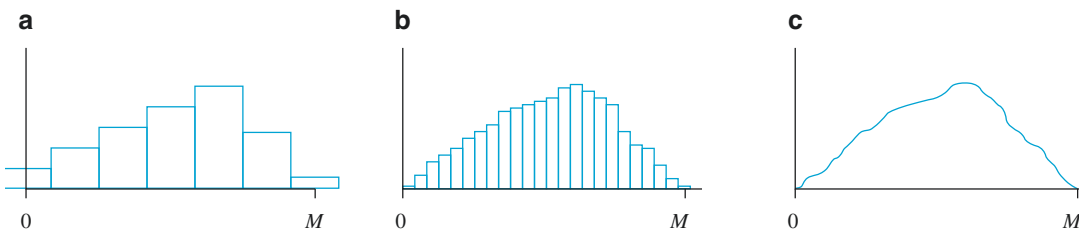
Figure 4.1 (**a**) Probability histogram of depth measured to the nearest meter; (**b**) probability histogram of depth measured to the nearest centimeter; (**c**) a limit of a sequence of discrete histograms

DEFINITION Let X be a continuous rv. Then a **probability distribution** or **probability density function** (pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

That is, the probability that X takes on a value in the interval $[a, b]$ is the area above this interval and under the graph of the density function, as illustrated in Figure 4.2. The graph of $f(x)$ is often referred to as the **density curve**.
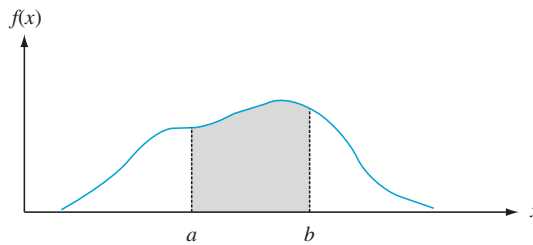
Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

For $f(x)$ to be a legitimate pdf, it must satisfy the following two conditions:

1. $f(x) \geq 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)dx =$ [area under the entire graph of $f(x)$] $= 1$

The **support** of a pdf $f(x)$ consists of all x values for which $f(x) > 0$. Although a pdf is defined for $-\infty < x < \infty$, we will typically display a pdf for the values in its support, and it is always understood that $f(x) = 0$ otherwise.

Example 4.4 The direction of an imperfection with respect to a reference line on a circular object such as a tire, brake rotor, or flywheel is, in general, subject to uncertainty. Consider the reference line connecting the valve stem on a tire to the center point, and let X be the angle measured clockwise to the location of an imperfection. One possible pdf for X is

$$f(x) = \frac{1}{360} \quad 0 \leq x < 360$$

The pdf is graphed in Figure 4.3. Clearly $f(x) \geq 0$. The area under the density curve is just the area of a rectangle: (height)(base) $= \left(\frac{1}{360}\right)(360) = 1$. The probability that the angle is between 90° and 180° is

$$P(90 \leq X \leq 180) = \int_{90}^{180} \frac{1}{360} dx = \frac{x}{360}\Big|_{x=90}^{x=180} = \frac{1}{4} = .25$$
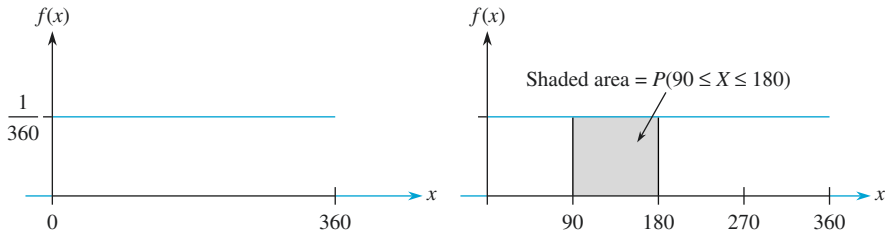
Figure 4.3 The pdf and probability for Example 4.4

The probability that the angle of occurrence is within 90° of the reference line is

$$P(0 \leq X \leq 90) + P(270 \leq X < 360) = .25 + .25 = .50 \qquad \blacksquare$$

Because the pdf in Figure 4.3 is completely "level" (i.e., has a uniform height) on the interval [0, 360), X is said to have a *uniform* distribution.

DEFINITION   A continuous rv X is said to have a **uniform distribution** on the interval $[A, B]$ if the pdf of X is

$$f(x; A, B) = \frac{1}{B - A} \quad A \leq X \leq B$$

The statement that X has a uniform distribution on $[A, B]$ will be denoted $X \sim \text{Unif}[A, B]$.

The graph of any uniform pdf looks like the graph in Figure 4.3 except that the interval of positive density is $[A, B]$ rather than [0, 360).

In the discrete case, a probability mass function (pmf) tells us how little "blobs" of probability mass of various magnitudes are distributed along the measurement axis. In the continuous case, probability density is "smeared" in a continuous fashion along the interval of possible values. When density is smeared uniformly over the interval, a uniform pdf, as in Figure 4.3, results.

When X is a discrete random variable, each possible value is assigned positive probability. This is not true of a continuous random variable, because the area under a density curve that lies above any single value is zero:

$$P(X = c) = P(c \leq X \leq c) = \int_c^c f(x)\,dx = 0$$

The fact that $P(X = c) = 0$ when X is continuous has an important practical consequence: The probability that X lies in some interval between a and b does not depend on whether the lower limit a or the upper limit b is included in the probability calculation:

$$P(a \leq X \leq b) = P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) \qquad (4.1)$$

In contrast, if X were discrete and both a and b were possible values of X (e.g., $X \sim$ Bin(20, .3) and $a = 5$, $b = 10$), then all four of the probabilities in (4.1) would be different. This also means that whether we include the endpoints of the range of values for a continuous rv X is somewhat arbitrary; for example, the pdf in Example 3.4 could be defined to be positive on (0, 360) or [0, 360] rather than [0, 360), and the same applies for a uniform distribution on $[A, B]$ in general.

The zero probability condition has a physical analog. Consider a solid circular rod (with cross-sectional area of 1 in^2 for simplicity). Place the rod alongside a measurement axis and suppose that the density of the rod at any point x is given by the value $f(x)$ of a density function. Then if the rod is sliced at points a and b and this segment is removed, the amount of mass removed is $\int_a^b f(x)dx$; however, if the rod is sliced just at the point c, no mass is removed. Mass is assigned to interval segments of the rod but not to individual points.

So, if $P(X = c) = 0$ when X is a continuous rv, then what does $f(c)$ represent? After all, if X were discrete, its pmf evaluated at $x = c$, $p(c)$, would indicate the probability that X equals c. To help understand what $f(c)$ means, consider a small window near $x = c$—say, $[c, c + \Delta x]$. Using a rectangle to approximate the area under $f(x)$ between c and $c + \Delta x$ (the usual "Riemann approximation" idea from calculus), one obtains $\int_c^{c + \Delta x} f(x)dx \approx \Delta x \cdot f(c)$, from which

$$f(c) \approx \frac{\int_c^{c + \Delta x} f(x)dx}{\Delta x} = \frac{P(c \leq X \leq c + \Delta x)}{\Delta x}$$

This indicates that $f(c)$ is not a probability, but rather roughly the probability of an interval *divided by the length of the chosen interval*. If we associate mass with probability and remember that interval length is the one-dimensional analogue of volume, then f represents their quotient, mass per volume, more commonly known as *density* (hence, the name pdf). The height of the function $f(x)$ at a particular point reflects how "dense" the values of X are near that point—taller sections of $f(x)$ contain more probability within a fixed interval length than do shorter sections.

Example 4.5 Climate change has made effective modeling and management of floodwaters ever more important in coastal areas. One variable of particular importance is the flow rate of water above some minimum threshold (typically where the rate becomes hazardous and requires intervention). The following pdf of $X =$ hazardous flood rate (m^3/s) is suggested under certain conditions by the article "A Framework for Probabilistic Assessment of Clear-Water Scour Around Bridge Piers" (*Structural Safety* 2017: 11–22):

$$f(x) = .04e^{-.04(x-10)} \quad x \geq 10$$

The graph of $f(x)$ is given in Figure 4.4; there is no density associated with flow rates below 10 m^3/s, because such flow rates are deemed nonhazardous under these particular conditions. The flow rate density decreases rapidly (exponentially fast) as x increases from 10. Clearly $f(x) \geq 0$; to show that $\int_{-\infty}^{\infty} f(x)dx = 1$, we use calculus:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{10} 0\,dx + \int_{10}^{\infty} .04e^{-.04(x-10)}dx = .04e^{.4} \int_{10}^{\infty} e^{-.04x}dx = .04e^{.4} \cdot \frac{e^{-.04x}}{-.04}\bigg|_{10}^{\infty}$$

$$= 0 - (-e^{.4} \cdot e^{-.04(10)}) = 1$$
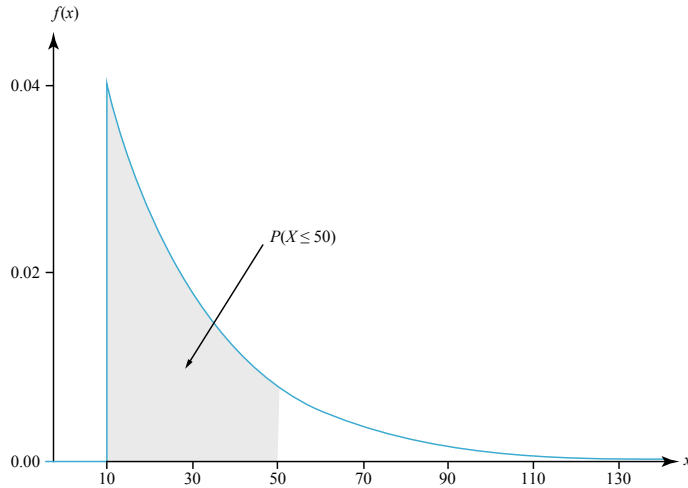
Figure 4.4 The density curve for flood rate in Example 4.5

According to this model, the probability that flood rate is at most 50 m³/s is

$$P(X \leq 50) = \int_{-\infty}^{50} f(x)\,dx = \int_{10}^{50} .04e^{-.04(x-10)}\,dx = .04e^{.4} \int_{10}^{50} e^{-0.4x}\,dx = .04e^{.4} \cdot \frac{e^{-.04x}}{-.04}\Big|_{10}^{50}$$

$$= e^{.4}(-e^{-.04(50)} + e^{-.04(10)}) = .798$$

Similarly, the probability the flood rate hits at least 200 m³/s, the point at which a nearby bridge will collapse, is

$$P(X \geq 200) = \int_{200}^{\infty} .04e^{-.04(x-10)}\,dx = .0005$$

Since X is a continuous rv, .0005 also equals $P(X > 200)$, the probability that the flood rate exceeds 200 m³/s. The difference between these two events is $\{X = 200\}$, i.e., that flood rate is exactly 200, which has probability zero: $P(X = 200) = \int_{200}^{200} f(x)dx = 0$.

This last statement may feel uncomfortable to you: Is there really zero chance that the flood rate is exactly 200 m³/s? If flow rate is treated as continuous, then "exactly 200" means $X = 200.000\ldots$, with an endless repetition of 0s. That is to say, X is not rounded to the nearest tenth or even hundredth; we are asking for the probability that X equals one specific number, 200.000…, out of the (uncountably) infinite collection of possible values of X.   ∎

Unlike discrete distributions such as the binomial, hypergeometric, and negative binomial, the distribution of any given continuous rv cannot usually be derived using simple probabilistic arguments (with a few notable exceptions). Instead, one must make a judicious choice of pdf based on prior knowledge and available data. Fortunately, some general pdf families have been found to fit well in a wide variety of experimental situations; several of these are discussed later in the chapter.

Just as in the discrete case, it is often helpful to think of the population of interest as consisting of X values rather than individuals or objects. The pdf is then a model for the distribution of values in this numerical population, and from this model various population characteristics (such as the mean) can be calculated.

Several of the most important concepts introduced in the study of discrete distributions also play an important role for continuous distributions. Definitions analogous to those in Chapter 3 involve replacing summation by integration.

The Cumulative Distribution Function

The cumulative distribution function (cdf) $F(x)$ for a discrete rv X gives, for any specified number x, the probability $P(X \leq x)$. It is obtained by summing the pmf $p(y)$ over all possible values y satisfying $y \leq x$. The cdf of a continuous rv gives the same probabilities $P(X \leq x)$ and is obtained by integrating the pdf $f(y)$ between the limits $-\infty$ and x.

DEFINITION   The **cumulative distribution function** $F(x)$ for a continuous rv X is defined for every number x by

$$F(x) = P(X \leq x) = \int\limits_{-\infty}^{x} f(y)dy$$

For each x, $F(x)$ is the area under the density curve to the left of x. This is illustrated in Figure 4.5, where $F(x)$ increases smoothly as x increases.
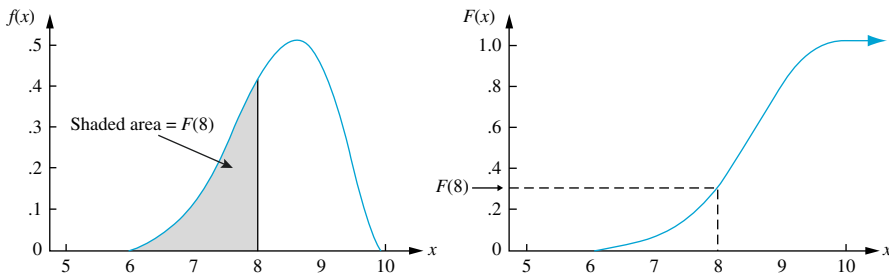
Figure 4.5 A pdf and associated cdf

Example 4.6 Let X, the thickness of a membrane, have a uniform distribution on $[A, B]$. The density function is shown in Figure 4.6. For $x < A$, $F(x) = 0$, since there is no area under the graph of the density function to the left of such an x. For $x \geq B$, $F(x) = 1$, since all the area is accumulated to the left of such an x. Finally, for $A \leq x \leq B$,
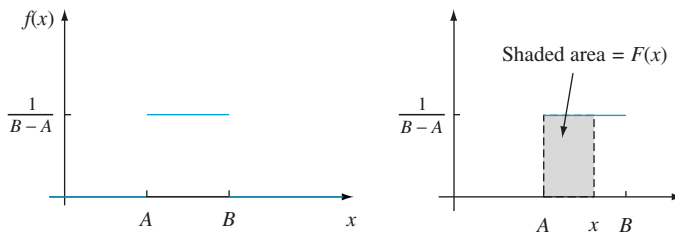
Figure 4.6 The pdf for a uniform distribution

$$F(x) = \int\limits_{-\infty}^{x} f(y)dy = \int\limits_{A}^{x} \frac{1}{B-A}dy = \frac{1}{B-A} \cdot y \bigg|_{y=A}^{y=x} = \frac{x-A}{B-A}$$

The entire cdf is

$$F(x) = \begin{cases} 0 & x < A \\ \dfrac{x-A}{B-A} & A \le x < B \\ 1 & x \ge B \end{cases}$$

The graph of this cdf appears in Figure 4.7.
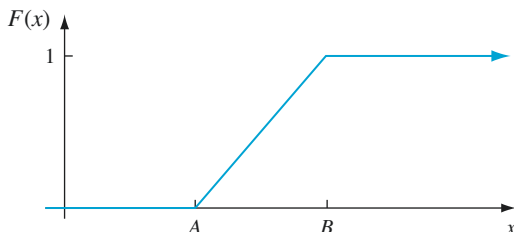
Figure 4.7 The cdf for a uniform distribution ∎

Using *F*(*x*) to Compute Probabilities
The importance of the cdf here, just as for discrete rvs, is that probabilities of various intervals can be computed from a formula or table for $F(x)$.

PROPOSITION   Let X be a continuous rv with pdf $f(x)$ and cdf $F(x)$. Then for any number a,

$$P(X > a) = 1 - F(a)$$

and for any two numbers a and b with $a < b$,

$$P(a \le X \le b) = F(b) - F(a)$$

Figure 4.8 illustrates the second part of this proposition; the desired probability is the shaded area under the density curve between a and b, and it equals the difference between the two shaded cumulative areas. This is different from what is appropriate for a discrete integer-valued rv (e.g., binomial or Poisson): $P(a \leq X \leq b) = F(b) - F(a - 1)$ when a and b are integers.
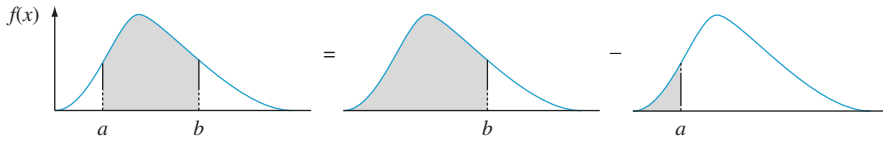
Figure 4.8   Computing $P(a \leq X \leq b)$ from cumulative probabilities

Example 4.7   Suppose the pdf of the magnitude X of a dynamic load on a bridge (in newtons) is given by

$$f(x) = \frac{1}{8} + \frac{3}{8}x \quad 0 \leq x \leq 2$$

For any number x between 0 and 2,

$$F(x) = \int_{-\infty}^{x} f(y)dy = \int_{0}^{x} \left(\frac{1}{8} + \frac{3}{8}y \right)dy = \frac{x}{8} + \frac{3x^2}{16}$$

Thus

$$F(x) = \begin{cases} 0 & x < 0 \\ \dfrac{x}{8} + \dfrac{3x^2}{16} & 0 \leq x \leq 2 \\ 1 & 2 < x \end{cases}$$

The graphs of $f(x)$ and $F(x)$ are shown in Figure 4.9. The probability that the load is between 1 and 1.5 N is
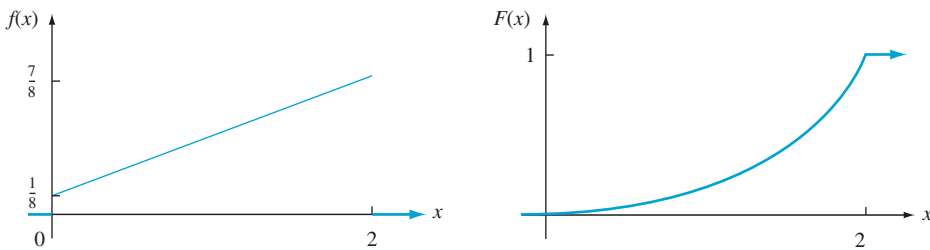
Figure 4.9   The pdf and cdf for Example 4.7

$$P(1 \leq X \leq 1.5) = F(1.5) - F(1) = \left[\frac{1}{8}(1.5) + \frac{3}{16}(1.5)^2\right] - \left[\frac{1}{8}(1) + \frac{3}{16}(1)^2\right] = \frac{19}{64} = .297$$

The probability that the load exceeds 1 N is

$$P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - \left[\frac{1}{8}(1) + \frac{3}{16}(1)^2\right] = \frac{11}{16} = .688 \qquad \blacksquare$$

The beauty of the cdf in the continuous case is that once it is available, any probability involving X can easily be calculated without any further integration.

Obtaining $f(x)$ from $F(x)$

For X discrete, the pmf is obtained from the cdf by taking the difference between two $F(x)$ values. The continuous analog of a difference is a derivative. The following result is a consequence of the Fundamental Theorem of Calculus.

PROPOSITION   If X is a continuous rv with pdf $f(x)$ and cdf $F(x)$, then at every x at which the derivative $F'(x)$ exists, $F'(x) = f(x)$.

Example 4.8   (Example 4.7 continued) The cdf in Example 4.7 is differentiable except at $x = 0$ and $x = 2$, where the graph of $F(x)$ has sharp corners. Since $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x > 2$, $F'(x) = 0 = f(x)$ for such x. For $0 < x < 2$,

$$F'(x) = \frac{d}{dx}\left(\frac{x}{8} + \frac{3x^2}{16}\right) = \frac{1}{8} + \frac{3x}{8} = f(x) \qquad \blacksquare$$

Percentiles of a Continuous Distribution

When we say that an individual's test score was at the 85th percentile of the population, we mean that 85% of all population scores were below that score and 15% were above. Similarly, the 40th percentile is the score that exceeds 40% of all scores and is exceeded by 60% of all scores.

DEFINITION   Let p be a number between 0 and 1. The (**100p**)**th percentile** (equivalently, the **pth quantile**) of the distribution of a continuous rv X, denoted by η_p, is defined by

$$p = F(\eta_p) = \int_{-\infty}^{\eta_p} f(y)dy \qquad (4.2)$$

Assuming we can find the inverse of $F(x)$, this can also be written as

$$\eta_p = F^{-1}(p)$$

In particular, the **median** of a continuous distribution is the 50th percentile, $\eta_{.5}$ or $F^{-1}(.5)$. That is, half the area under the density curve is to the left of the median and half is to the right of the median. We will also denote the median of a distribution by $\tilde{\mu}$.

According to Expression (4.2), η_p is that value on the measurement axis such that $100p\%$ of the area under the graph of $f(x)$ lies to the left of η_p and $100(1-p)\%$ lies to the right. Thus $\eta_{.75}$, the 75th percentile, is such that the area under the graph of $f(x)$ to the left of $\eta_{.75}$ is .75. Figure 4.10 illustrates the definition.
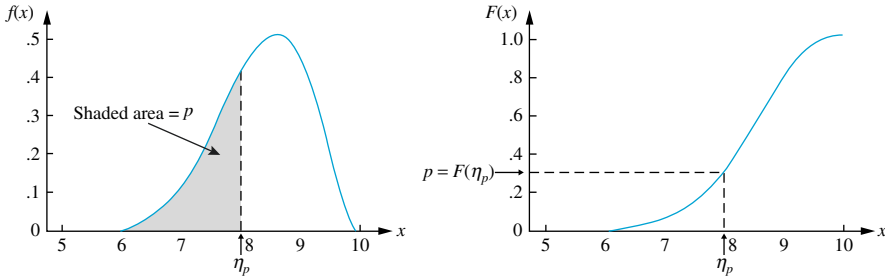
Figure 4.10   The $(100p)$th percentile of a continuous distribution

Example 4.9 The distribution of the amount of gravel (in tons) sold by a construction supply company in a given week is a continuous rv X with pdf

$$f(x) = \frac{3}{2}(1-x^2) \quad 0 \le x \le 1$$

The cdf of sales for any x between 0 and 1 is

$$F(x) = \int_0^x \frac{3}{2}(1-y^2)dy = \frac{3}{2}\left(y - \frac{y^3}{3}\right)\bigg|_{y=0}^{y=x} = \frac{3}{2}\left(x - \frac{x^3}{3}\right)$$

The graphs of both $f(x)$ and $F(x)$ appear in Figure 4.11.
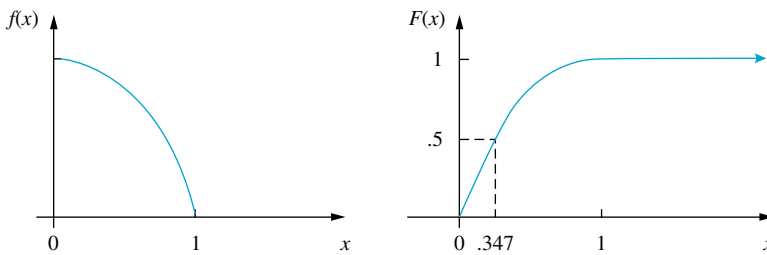
Figure 4.11   The pdf and cdf for Example 4.9

The $(100p)$th percentile of this distribution satisfies the equation

$$p = F(\eta_p) = \frac{3}{2}\left[\eta_p - \frac{\eta_p^3}{3}\right]$$

that is,

$$\eta_p^3 - 3\eta_p + 2p = 0$$

For the median $\tilde{\mu} = \eta_{.5}$, $p = .5$ and the equation to be solved is $\tilde{\mu}^3 - 3\tilde{\mu} + 1 = 0$; the solution is $\tilde{\mu} = .347$. If the distribution remains the same from week to week, then in the long run 50% of all weeks will result in sales of less than .347 tons and 50% in more than .347 tons. ∎

A continuous distribution whose pdf is **symmetric**—which means that the graph of the pdf to the left of some point is a mirror image of the graph to the right of that point—has median $\tilde{\mu}$ equal to the point of symmetry, since half the area under the curve lies to either side of this point. Figure 4.12 gives several examples. The amount of error in a measurement of a physical quantity is often assumed to have a symmetric distribution.
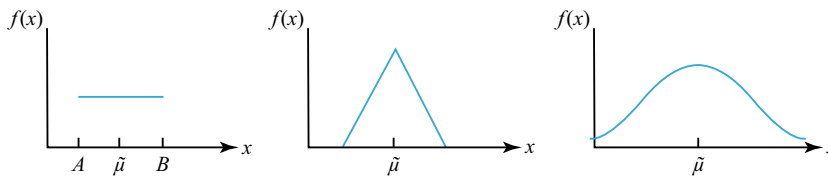
Figure 4.12   Medians of symmetric distributions

Exercises: Section 4.1 (1–17)

1. Let X denote the amount of time for which a book on 2-h reserve at a college library is checked out by a randomly selected student and suppose that X has density function

$$f(x) = .5x \quad 0 \le x \le 2$$

   Calculate the following probabilities:

   a. $P(X \le 1)$
   b. $P(.5 \le X \le 1.5)$
   c. $P(1.5 < X)$

2. Suppose the reaction temperature X (in °C) in a chemical process has a uniform distribution with $A = -5$ and $B = 5$.

   a. Compute $P(X < 0)$.
   b. Compute $P(-2.5 < X < 2.5)$.
   c. Compute $P(-2 \le X \le 3)$.
   d. For k satisfying $-5 < k < k + 4 < 5$, compute $P(k < X < k + 4)$. Interpret this in words.

3. Suppose the error involved in making a measurement is a continuous rv X with pdf

$$f(x) = .09375(4 - x^2) \quad -2 \le x \le 2$$

   a. Sketch the graph of $f(x)$.
   b. Compute $P(X > 0)$.
   c. Compute $P(-1 < X < 1)$.
   d. Compute $P(X < -.5 \text{ or } X > .5)$.

4. Let X denote the power (MW) generated by a wind turbine at a given wind speed. The article "An Investigation of Wind Power Density Distribution at Location With Low and High Wind Speeds Using Statistical Model" (*Appl. Energy* 2018: 442–451) proposes the *Rayleigh* distribution, with pdf

$$f(x; \theta) = \frac{x}{\theta^2} \cdot e^{-x^2/(2\theta^2)} \quad x > 0$$

   as a model for the X distribution. The value of the parameter θ depends upon the prevailing wind speed.

   a. Verify that $f(x; \theta)$ is a legitimate pdf.
   b. Suppose $\theta = 100$. What is the probability that X is at most 200? Less than 200? At least 200?

c. What is the probability that X is between 100 and 200 (again assuming $\theta = 100$)?

d. Give an expression for the cdf of X.

5. A college professor never finishes his lecture before the end of the hour and always finishes his lecture within 2 min after the hour. Let $X =$ the time that elapses between the end of the hour and the end of the lecture and suppose the pdf of X is

$$f(x) = kx^2 \quad 0 \leq x \leq 2$$

a. Find the value of k. [*Hint*: Total area under the graph of $f(x)$ is 1.]

b. What is the probability that the lecture ends within 1 min of the end of the hour?

c. What is the probability that the lecture continues beyond the hour for between 60 and 90 s?

d. What is the probability that the lecture continues for at least 90 s beyond the end of the hour?

6. The grade point averages (GPAs) for graduating seniors at a college are distributed as a continuous rv X with pdf

$$f(x) = k[1 - (x - 3)^2] \quad 2 \leq x \leq 4$$

a. Sketch the graph of $f(x)$.

b. Find the value of k.

c. Find the probability that a GPA exceeds 3.

d. Find the probability that a GPA is within .25 of 3.

e. Find the probability that a GPA differs from 3 by more than .5.

7. The time X (min) for a laboratory assistant to prepare the equipment for a certain experiment is believed to have a uniform distribution with $A = 25$ and $B = 35$.

a. Write the pdf of X and sketch its graph.

b. What is the probability that preparation time exceeds 33 min?

c. What is the probability that preparation time is within 2 min of the median time? [*Hint*: Identify $\tilde{\mu}$ from the graph of $f(x)$.]

d. For any a such that $25 < a < a + 2 < 35$, what is the probability that preparation time is between a and $a + 2$ min?

8. Commuting to work requires getting on a bus near home and then transferring to a second bus. If the waiting time (in minutes) at each stop has a uniform distribution with $A = 0$ and $B = 5$, then it can be shown that the total waiting time Y has the pdf

$$f(y) = \begin{cases} y/25 & 0 \leq y < 5 \\ 2/5 - y/25 & 5 \leq y \leq 10 \end{cases}$$

a. Sketch the pdf of Y.

b. Verify that $\int_{-\infty}^{\infty} f(y)dy = 1$.

c. What is the probability that total waiting time is at most 3 min?

d. What is the probability that total waiting time is at most 8 min?

e. What is the probability that total waiting time is between 3 and 8 min?

f. What is the probability that total waiting time is either less than 2 min or more than 6 min?

9. Consider again the rv $X =$ hazardous flood rate given in Example 4.5. What is the probability that the flood rate is

a. At most 40 m³/s?

b. More than 40 m³/s? At least 40 m³/s?

c. Between 40 and 60 m³/s?

10. A family of pdfs that has been used to approximate the distribution of income, city population size, and size of firms is the *Pareto* family. The family has two parameters, k and θ, both > 0, and the pdf is

$$f(x; k, \theta) = \frac{k \cdot \theta^k}{x^{k+1}} \quad x \geq \theta$$

a. Sketch the graph of $f(x; k, \theta)$.

b. Verify that the total area under the graph equals 1.

c. If the rv X has pdf $f(x; k, \theta)$, for any fixed $b > \theta$, obtain an expression for $P(X \leq b)$.

d. For $\theta < a < b$, obtain an expression for the probability $P(a \leq X \leq b)$.

e. Find an expression for the $(100p)$th percentile η_p.

11. The cdf of checkout duration X as described in Exercise 1 is

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2/4 & 0 \leq x < 2 \\ 1 & 2 \leq x \end{cases}$$

Use this to compute the following:

a. $P(X \leq 1)$
b. $P(.5 \leq X \leq 1)$
c. $P(X > .5)$
d. The median checkout duration [*Hint:* Solve $F(\tilde{\mu}) = .5$.]
e. $F'(x)$ to obtain the density function $f(x)$

12. The cdf for $X = $ measurement error of Exercise 3 is

$$F(x) = \begin{cases} 0 & x < -2 \\ .5 + 3(4x - x^3/3)/32 & -2 \leq x < 2 \\ 1 & 2 \leq x \end{cases}$$

a. Compute $P(X < 0)$.
b. Compute $P(-1 < X < 1)$.
c. Compute $P(.5 < X)$.
d. Verify that $f(x)$ is as given in Exercise 3 by obtaining $F'(x)$.
e. Verify that $\tilde{\mu} = 0$.

13. Suppose that in a certain traffic environment, the distribution of $X = $ the time headway (sec) between two randomly selected consecutive cars has the form

$$f(x) = \frac{k}{x^4} \quad x > 1$$

a. Determine the value of k for which $f(x)$ is a legitimate pdf.
b. Obtain the cumulative distribution function.
c. Use the cdf from (b) to determine the probability that headway exceeds 2 s and also the probability that headway is between 2 and 3 s.

14. Let X denote the amount of space occupied by an article placed in a 1-ft^3 packing container. The pdf of X is

$$f(x) = 90x^8(1 - x) \quad 0 < x < 1$$

a. Graph the pdf. Then obtain the cdf of X and graph it.
b. What is $P(X \leq .5)$ [i.e., $F(.5)$]?
c. Using part (a), what is $P(.25 < X \leq .5)$? What is $P(.25 \leq X \leq .5)$?
d. What is the 75th percentile of the distribution?

15. Answer parts (a)–(d) of Exercise 14 for the random variable X, lecture time past the hour, given in Exercise 5.

16. Let X be a continuous rv with cdf

$$F(x) = \begin{cases} 0 & x \leq 0 \\ x[1 + \ln(4/x)]/4 & 0 < x \leq 4 \\ 1 & x > 4 \end{cases}$$

[This type of cdf is suggested in the article "Variability in Measured Bedload-Transport Rates" (*Water Resources Bull.* 1985:39–48) as a model for a hydrologic variable.] What is

a. $P(X \leq 1)$?
b. $P(1 \leq X \leq 3)$?
c. The pdf of X?

17. Let X be the temperature in °C at which a chemical reaction takes place, and let Y be the temperature in °F (so $Y = 1.8X + 32$).

a. If the median of the X distribution is $\tilde{\mu}$, show that $1.8\tilde{\mu} + 32$ is the median of the Y distribution.
b. How is the 90th percentile of the Y distribution related to the 90th percentile of the X distribution? Verify your conjecture.
c. More generally, if $Y = aX + b$, how is any particular percentile of the Y distribution related to the corresponding percentile of the X distribution?

4.2 Expected Values and Moment Generating Functions

In Section 4.1 we saw that the transition from a discrete cdf to a continuous cdf entails replacing summation by integration. The same thing is true in moving from expected values and mgfs of discrete variables to those of continuous variables.

Expected Values

For a discrete random variable X, $E(X)$ was obtained by summing $x \cdot p(x)$ over possible X values. Here we replace summation by integration and the pmf by the pdf to get a continuous weighted average.

DEFINITION   The **expected** or **mean value** of a continuous rv X with pdf $f(x)$ is

$$\mu = \mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

This expected value will exist provided that $\int_{-\infty}^{\infty} |x| f(x)dx < \infty$. In practice, the limits of integration are specified by the support of the pdf (since $f(x) = 0$ otherwise).

Example 4.10 (Example 4.9 continued) The pdf of weekly gravel sales X was

$$f(x) = \frac{3}{2}(1 - x^2) \quad 0 \le x \le 1$$

so

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx = \int_{0}^{1} x \cdot \frac{3}{2}(1 - x^2)dx$$

$$= \frac{3}{2}\int_{0}^{1}(x - x^3)dx = \frac{3}{2}\left(\frac{x^2}{2} - \frac{x^4}{4}\right)\Big|_{x=0}^{x=1} = \frac{3}{8} = .375$$

If gravel sales are determined week after week according to the given pdf, then the long-run average value of sales per week will be .375 ton. ∎

Similar to the interpretation in the discrete case, the mean value μ can be regarded as the balance point (or fulcrum or center of mass) of a continuous distribution. In Example 4.10, if a piece of cardboard was cut out in the shape of the region under the density curve $f(x)$, then it would balance if supported at $\mu = 3/8$ along the bottom edge. When a pdf $f(x)$ is symmetric, then it will balance at its point of symmetry, which must be the mean μ (assuming μ exists). Recall from Section 4.1 that the median is also the point of symmetry; in general, if a distribution is symmetric and the mean exists, then it is equal to the median.

Often we wish to compute the expected value of some function $h(X)$ of the rv X. If we think of $h(X)$ as a new rv Y, methods from Section 4.7 can be used to derive the pdf of Y, and $E(Y)$ can be computed from the definition. Fortunately, as in the discrete case, there is an easier way to compute $E[h(X)]$.

LAW OF THE
UNCONSCIOUS
STATISTICIAN

If X is a continuous rv with pdf $f(x)$ and $h(X)$ is any function of X, then

$$\mu_{h(X)} = E[h(X)] = \int\limits_{-\infty}^{\infty} h(x) \cdot f(x) dx$$

This expected value will exist provided that $\int_{-\infty}^{\infty} |h(x)| f(x) dx < \infty$.

Importantly, except in the cases where $h(x)$ is a linear function (see later in this section), $E[h(X)]$ is *not* equal to $h(\mu_X)$, the function h evaluated at the mean of X.

Example 4.11 The variation in a certain electrical current source X (in milliamps) can be modeled by the pdf

$$f(x) = 1.25 - .25x \quad 2 \leq x \leq 4$$

The average current from this source is

$$E(X) = \int\limits_{2}^{4} x(1.25 - .25x) dx = \frac{17}{6} = 2.833 \text{ mA}$$

If this current passes through a 220-ohm resistor, the resulting power (in microwatts) is given by the expression $h(X) = (\text{current})^2(\text{resistance}) = 220X^2$. The expected power is given by

$$E(h(X)) = E(220X^2) = \int\limits_{2}^{4} 220x^2(1.25 - .25x) dx = \frac{5500}{3} = 1833.3 \text{ microwatts}$$

Notice that the expected power is *not* equal to $220(2.833)^2$, a common error that results from substituting the mean current μ_X into the power formula. ∎

Example 4.12 Two species are competing in a region for control of a limited amount of a resource. Let X = the proportion of the resource controlled by species 1 and suppose X has pdf

$$f(x) = 1 \quad 0 \leq x \leq 1$$

which is the uniform distribution on [0, 1]. (In her book *Ecological Diversity*, E. C. Pielou calls this the "broken-stick" model for resource allocation, since it is analogous to breaking a stick at a randomly chosen point.) Then the species that controls the majority of this resource controls the amount

$$h(X) = \max(X, 1 - X) = \begin{cases} 1 - X & \text{if} \quad 0 \leq X < .5 \\ X & \text{if} \quad .5 \leq X \leq 1 \end{cases}$$

The expected amount controlled by the species having majority control is then

$$E[h(X)] = \int_{-\infty}^{\infty} \max(x, 1 - x) \cdot f(x)dx = \int_{0}^{1} \max(x, 1 - x) \cdot 1 dx$$

$$= \int_{0}^{.5} (1 - x) \cdot 1 dx + \int_{.5}^{1} x \cdot 1 dx = \frac{3}{4} \qquad \blacksquare$$

In the discrete case, the variance of X was defined as the expected squared deviation from μ and was calculated by summation. Here again integration replaces summation.

DEFINITION   The **variance** of a continuous random variable X with pdf $f(x)$ and mean value μ is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx = E[(X - \mu)^2]$$

The **standard deviation** of X is $\mathrm{SD}(X) = \sigma_X = \sqrt{V(X)}$.

As in the discrete case, σ_X^2 is the expected or average squared deviation about the mean μ, and σ_X can be interpreted roughly as the size of a representative deviation from the mean value μ.

Example 4.13 Let $X \sim \text{Unif}[A, B]$. Since a uniform distribution is symmetric, the mean of X is at the density curve's point of symmetry, which is clearly the midpoint $(A + B)/2$. This can be verified by integration:

$$\mu = \int_{A}^{B} x \cdot \frac{1}{B - A} dx = \frac{1}{B - A} \frac{x^2}{2}\bigg|_{A}^{B} = \frac{1}{B - A} \frac{B^2 - A^2}{2} = \frac{A + B}{2}$$

The variance of X is then given by

$$\sigma^2 = \int_{A}^{B} (x - \mu)^2 \cdot \frac{1}{B - A} dx = \frac{1}{B - A} \int_{A}^{B} \left(x - \frac{A + B}{2}\right)^2 dx$$

$$= \frac{1}{B - A} \int_{-(B-A)/2}^{(B-A)/2} u^2 du \qquad \text{substitute } u = x - \frac{A + B}{2}$$

$$= \frac{2}{B - A} \int_{0}^{(B-A)/2} u^2 du \qquad \text{symmetry}$$

$$= \frac{2}{B - A} \frac{u^3}{3}\bigg|_{0}^{(B-A)/2} = \frac{2}{B - A} \frac{(B - A)^3}{2^3 \cdot 3} = \frac{(B - A)^2}{12}$$

The standard deviation of X is the square root of the variance: $\sigma = (B - A)/\sqrt{12}$. Notice that the standard deviation of a Unif$[A, B]$ distribution is proportional to the length of the interval, $B - A$, which matches our intuitive notion that a larger standard deviation corresponds to greater "spread" in a distribution. ∎

Section 3.3 presented several properties of expected value, variance, and standard deviation for discrete random variables. Those same properties hold for the continuous case; proofs of these results are obtained by replacing summation with integration in the proofs presented in Chapter 3.

PROPOSITION   Let X be a continuous rv with pdf $f(x)$, mean μ, and standard deviation σ. Then the following properties hold.

1. (variance shortcut)

$$V(X) = E(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx - \left(\int_{-\infty}^{\infty} x \cdot f(x)dx \right)^2$$

2. (linearity of expectation) For any functions $h_1(X)$ and $h_2(X)$ and any constants a_1, a_2, and b,

$$E[a_1 h_1(X) + a_2 h_2(X) + b] = a_1 E[h_1(X)] + a_2 E[h_2(X)] + b$$

3. (rescaling) For any constants a and b,

$$E(aX + b) = aE(X) + b \quad V(aX + b) = a^2 \sigma_X^2 \quad \sigma_{aX+b} = |a|\sigma_X$$

Example 4.14 (Example 4.10 continued) For X = weekly gravel sales, we computed $E(X) = 3/8$. Since

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx = \int_{0}^{1} x^2 \cdot \frac{3}{2}(1 - x^2)dx = \frac{3}{2}\int_{0}^{1}(x^2 - x^4)dx = \frac{1}{5},$$

$$V(X) = 1/5 - (3/8)^2 = 19/320 = .059 \quad \text{and} \quad \sigma_X = \sqrt{.059} = .244$$

Suppose the amount of gravel actually received by customers in a week is $h(X) = X - .02X^2$; the second term accounts for the small amount that is lost in transport. Then the average weekly amount received by customers is

$$E(X - .02X^2) = E(X) - .02E(X^2) = .375 - .02(.2) = .371 \text{ tons} \qquad ∎$$

Example 4.15 When a dart is thrown at a circular target, consider the location of the landing point relative to the bull's eye. Let X be the angle in degrees measured from the horizontal, and assume that X is uniformly distributed on [0, 360). By Example 4.13, $E(X) = 180$ and $\sigma_X = 360/\sqrt{12}$. Define Y to be the angle measured in radians between $-\pi$ and π, so $Y = (2\pi/360)X - \pi$. Then, applying the rescaling properties with $a = 2\pi/360$ and $b = -\pi$,

$$E(Y) = \frac{2\pi}{360} E(X) - \pi = \frac{2\pi}{360} 180 - \pi = 0.$$

and

$$\sigma_Y = \left| \frac{2\pi}{360} \right| \cdot \sigma_X = \frac{2\pi}{360} \frac{360}{\sqrt{12}} = \frac{2\pi}{\sqrt{12}} \qquad \blacksquare$$

As a special case of the result $E(aX + b) = aE(X) + b$, set $a = 1$ and $b = -\mu$, giving $E(X - \mu) = E(X) - \mu = 0$. This can be interpreted as saying that the expected deviation from μ is 0; $\int_{-\infty}^{\infty} (x - \mu)f(x)dx = 0$. The integral suggests a physical interpretation: With $(x - \mu)$ as the lever arm and $f(x)$ as the weight function, the total torque is 0. Using a seesaw as a model with weight distributed in accord with $f(x)$, the seesaw will balance at μ.

Approximating the Mean Value and Standard Deviation

Let X be a random variable with mean value μ and variance σ^2. Then we have already seen that the new random variable $Y = h(X) = aX + b$, a linear function of X, has mean value $a\mu + b$ and variance $a^2\sigma^2$. But what can be said about the mean and variance of Y if $h(x)$ is a nonlinear function?

PROPOSITION
(The Delta Method)

Suppose $h(x)$ is differentiable and that its derivative evaluated at μ satisfies $h'(\mu) \neq 0$. Then if the variance of X is small, so that the distribution of X is largely concentrated on an interval of values close to μ, the mean value and variance of $Y = h(X)$ can be approximated as follows:

$$E[h(X)] \approx h(\mu), \quad V[h(X)] \approx [h'(\mu)]^2 \sigma^2$$

The justification for these approximations is a first-order Taylor series expansion of $h(X)$ about μ; that is, we approximate the function for values near μ by the tangent line to the function at the point $(\mu, h(\mu))$:

$$Y = h(X) \approx h(\mu) + h'(\mu)(X - \mu)$$

Taking the expected value of this gives $E[h(X)] \approx h(\mu)$. Since $h(\mu)$ and $h'(\mu)$ are numerical constants, the variance of the linear approximation is $V[h(X)] \approx 0 + [h'(\mu)]^2 V(X - \mu) = [h'(\mu)]^2 \sigma^2$.

Example 4.16 A chemistry student determined the mass m and volume X of an aluminum chunk and took the ratio to obtain the density $Y = h(X) = m/X$. The mass is measured much more accurately, so for an approximate calculation it can be regarded as a constant. The derivative of $h(X)$ is $-m/X^2$, so

$$\sigma_Y^2 \approx \left[\frac{-m}{\mu_X^2} \right]^2 \sigma_X^2$$

The standard deviation is then $\sigma_Y \approx [m/\mu_X^2] \sigma_X$. A particular aluminum chunk had measurements $m = 18.19$ g and $X = 6.6$ cm^3, which gives an estimated density $Y = m/X = 18.19/6.6 = 2.76$. A rough value for the standard deviation of X is $\sigma_X = .3$ cm^3. Our best guess for the mean of the

X distribution is the measured value, so $\mu_Y \approx h(\mu_X) = 18.19/6.6 = 2.76$, and the estimated standard deviation for the estimated density is

$$\sigma_Y \approx \frac{m}{\mu_X^2}\sigma_X = \frac{18.19}{6.6^2}(.3) = .125$$

Compare the estimate of 2.76, standard deviation .125, with the official value 2.70 for the density of aluminum. ∎

Moment Generating Functions

Moments and moment generating functions for discrete random variables were introduced in Section 3.4. These concepts carry over to the continuous case.

DEFINITION   The **moment generating function** (mgf) of a continuous random variable X is

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx}f(x)dx.$$

As in the discrete case, the moment generating function exists if $M_X(t)$ is defined for an interval that includes zero as well as positive and negative values of t.

Just as before, when $t = 0$ the value of the mgf is always 1:

$$M_X(0) = E(e^{0X}) = \int_{-\infty}^{\infty} e^{0x}f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$$

Example 4.17 At a store, the checkout time X in minutes has the pdf $f(x) = 2e^{-2x}$, $x \geq 0$. Then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx}f(x)dx = \int_{0}^{\infty} e^{tx}(2e^{-2x})dx = \int_{0}^{\infty} 2e^{-(2-t)x}dx$$

$$= -\frac{2}{2-t}e^{-(2-t)x}\Big|_{0}^{\infty} = \frac{2}{2-t} - \frac{2}{2-t}\lim_{x\to\infty} e^{-(2-t)x}$$

The limit above exists (in fact, it equals zero) provided the coefficient on x is negative, i.e., $-(2-t) < 0$. This is equivalent to $t < 2$. The mgf exists because it is defined for an interval of values including 0 in its interior, specifically $(-\infty, 2)$. For t in that interval, the mgf of X is $M_X(t) = 2/(2-t)$. Notice that $M_X(0) = 2/(2-0) = 1$, as it must be. ∎

The uniqueness property for the mgf of a discrete rv is equally valid in the continuous case. Two distributions have the same pdf if and only if they have the same moment generating function, assuming that the mgf exists. For example, if a random variable X is known to have mgf $M_X(t) = 2/(2-t)$ for $t < 2$, then from Example 4.17 it must necessarily be the case that the pdf of X is $f(x) = 2e^{-2x}$ for $x \geq 0$ and $f(x) = 0$ otherwise.

In the discrete case we had a theorem on how to get moments from the mgf, and this theorem applies also in the continuous case: the rth moment of a continuous rv with mgf $M_X(t)$ is given by

$$E(X^r) = M_X^{(r)}(0),$$

the rth derivative of the mgf with respect to t evaluated at $t = 0$, if the mgf exists.

Example 4.18 (Example 4.17 continued) The mgf of the rv X = checkout time at the store was found to be $M_X(t) = 2/(2 - t) = 2(2 - t)^{-1}$ for $t < 2$. To find the mean and standard deviation, first compute the derivatives:

$$M_X'(t) = -2(2 - t)^{-2}(-1) = \frac{2}{(2 - t)^2}$$

$$M_X''(t) = \frac{d}{dt}\left[2(2 - t)^{-2}\right] = -4(2 - t)^{-3}(-1) = \frac{4}{(2 - t)^3}$$

Setting t to 0 in the first derivative gives the expected checkout time as

$$E(X) = M_X^{(1)}(0) = M_X'(0) = .5 \text{ min.}$$

Setting t to 0 in the second derivative gives the second moment

$$E(X^2) = M_X^{(2)}(0) = M_X''(0) = .5,$$

from which $V(X) = E(X^2) - [E(X)]^2 = .5 - .5^2 = .25$ and $\sigma = \sqrt{.25} = .5$ min. ∎

As in the discrete case, if X has the mgf $M_X(t)$ then the mgf of the linear function $Y = aX + b$ is $M_Y(t) = e^{bt}M_X(at)$.

Example 4.19 Let X have a uniform distribution on the interval $[A, B]$, so its pdf is $f(x) = 1/(B - A)$, $A \le x \le B$; $f(x) = 0$ otherwise. As verified in Exercise 32, the moment generating function of X is

$$M_X(t) = \begin{cases} \dfrac{e^{Bt} - e^{At}}{(B - A)t} & t \ne 0 \\ 1 & t = 0 \end{cases}$$

In particular, consider the situation in Example 4.15. Let X, the angle measured in degrees, be uniform on [0, 360], so $A = 0$ and $B = 360$. Then

$$M_X(t) = \frac{e^{360t} - 1}{360t} \quad t \ne 0, \quad M_X(0) = 1$$

Now let $Y = (2\pi/360)X - \pi$, so Y is the angle measured in radians and Y is between $-\pi$ and π. Using the mgf rule for linear transformations with $a = 2\pi/360$ and $b = -\pi$,

$$M_Y(t) = e^{bt}M_X(at) = e^{-\pi t}M_X\left(\frac{2\pi}{360}t\right)$$

$$= e^{-\pi t}\frac{e^{360(2\pi/360)t} - 1}{360\left(\frac{2\pi}{360}t\right)}$$

$$= \frac{e^{\pi t} - e^{-\pi t}}{2\pi t} \quad t \neq 0, \quad M_Y(0) = 1$$

This matches the general form of the moment generating function for a uniform random variable with $A = -\pi$ and $B = \pi$. Thus, by the uniqueness principle, $Y \sim \text{Unif}[-\pi, \pi]$. ∎

Exercises: Section 4.2 (18–38)

18. Reconsider the distribution of checkout duration X described in Exercises 1 and 11. Compute the following:

    a. $E(X)$
    b. $V(X)$ and σ_X
    c. If the borrower is charged an amount $h(X) = X^2$ when checkout duration is X, compute the expected charge $E[h(X)]$.

19. Recall the distribution of hazardous flood rate used in Example 4.5.

    a. Obtain the mean and standard deviation of this distribution.
    b. What is the probability that the flood rate is within 1 standard deviation of the mean value?

20. The article "Forecasting Postflight Hip Fracture Probability Using Probabilistic Modeling" (*J. Biomech. Engr.* 2019) examines the risk of bone breaks for astronauts returning from space, who typically lose density during missions. One quantity the article's authors model is the *midpoint fracture risk index* (mFRI), the ratio of applied load to bone strength at which the chance of a fracture is 50–50. The article suggests a uniform distribution on [0.55, 1.45] to model this unitless index value.

    a. Calculate the mean and standard deviation of mFRI using the specified model.
    b. Determine the cdf of mFRI.

c. What is the probability that mFRI is less than 1? Between 0.75 and 1.25?
d. What is the probability that mFRI is within one standard deviation of its expected value? Within two standard deviations?

21. For the distribution of Exercise 14,

    a. Compute $E(X)$ and σ_X.
    b. What is the probability that X is more than 1 standard deviation from its mean value?

22. Consider the pdf of X = grade point average given in Exercise 6.

    a. Obtain and graph the cdf of X.
    b. From the graph of $f(x)$, what is $\tilde{\mu}$?
    c. Compute $E(X)$ and $V(X)$.

23. Let X have a uniform distribution on the interval $[A, B]$.

    a. Obtain an expression for the $(100p)$th percentile.
    b. Obtain an expression for the median, $\tilde{\mu}$. How does this compare to the mean μ, and why does that make sense for this distribution?
    c. For n a positive integer, compute $E(X^n)$.

24. Consider the pdf for total waiting time Y for two buses

$$f(y) = \begin{cases} .04y & 0 \leq y < 5 \\ .4 - .04y & 5 \leq y \leq 10 \end{cases}$$

introduced in Exercise 8.

a. Compute and sketch the cdf of Y. [*Hint*: Consider separately $0 \leq y < 5$ and $5 \leq y \leq 10$ in computing $F(y)$. A graph of the pdf should be helpful.]

b. Obtain an expression for the $(100p)$th percentile. [*Hint*: Consider separately $0 < p < .5$ and $.5 \leq p < 1$.]

c. Compute $E(Y)$ and $V(Y)$. How do these compare with the expected waiting time and variance for a single bus when the time is uniformly distributed on $[0, 5]$?

25. An ecologist wishes to mark off a circular sampling region having radius 10 m. However, the radius of the resulting region is actually a random variable R with pdf

$$f(r) = \frac{3}{4}[1 - (10 - r)^2] \quad 9 \leq r \leq 11$$

What is the expected area of the resulting circular region?

26. The weekly demand for propane gas (in 1000s of gallons) from a particular facility is an rv X with pdf

$$f(x) = 2\left(1 - \frac{1}{x^2}\right) \quad 1 \leq x \leq 2$$

a. Compute the cdf of X.

b. Obtain an expression for the $(100p)$th percentile. What is the value of $\tilde{\mu}$?

c. Compute $E(X)$. How do the mean and median of this distribution compare?

d. Compute $V(X)$ and σ_X.

e. If 1.5 thousand gallons are in stock at the beginning of the week and no new supply is due in during the week, how much of the 1.5 thousand gallons is expected to be left at the end of the week? [*Hint*: Let $h(x)$ = amount left when demand is x.]

27. If the temperature at which a compound melts is a random variable with mean value 120 °C and standard deviation 2 °C, what are the mean temperature and standard deviation measured in °F? [*Hint*: °F = 1.8 °C + 32.]

28. Let X have the Pareto pdf introduced in Exercise 10.

$$f(x; k, \theta) = \frac{k \cdot \theta^k}{x^{k+1}} \quad x \geq \theta$$

a. If $k > 1$, compute $E(X)$.

b. What can you say about $E(X)$ if $k = 1$?

c. If $k > 2$, show that $V(X) = k\theta^2(k-1)^{-2}(k-2)^{-1}$.

d. If $k = 2$, what can you say about $V(X)$?

e. What conditions on k are necessary to ensure that $E(X^n)$ is finite?

29. The time (min) between successive visits to a particular website has pdf $f(x) = 4e^{-4x}$, $x \geq 0$; $f(x) = 0$ otherwise. Use integration by parts to obtain $E(X)$ and $V(X)$.

30. Suppose that the pdf of X is

$$f(x) = .5 - \frac{x}{8} \quad 0 \leq x \leq 4$$

a. Show that $E(X) = 4/3$ and $V(X) = 8/9$.

b. The coefficient of skewness is defined as $E[(X - \mu)^3]/\sigma^3$. Show that its value for the given pdf is .566. What would the skewness be for a perfectly symmetric pdf? Explain your reasoning.

31. a. If the voltage v across a medium is fixed but current I is random, then resistance will also be a random variable related to I by $R = v/I$. If $\mu_I = 20$ and $\sigma_I = .5$, use the delta method to calculate approximations to μ_R and σ_R.

b. Let R have the distribution in Exercise 25, whose mean and variance are 10 and 1/5, respectively. Let $h(R) = \pi R^2$, the area of the ecologist's sampling region. How does $E[h(R)]$ from Exercise 25 compare to the delta method approximation $h(10)$?

c. The variance of the region's area is $V[h(R)] = 14008\pi^2/175$. Compute the delta method approximation to $V[h(R)]$. How good is the approximation?

32. Let X have a uniform distribution on the interval $[A, B]$, so its pdf is $f(x) = 1/(B - A)$, $A \leq x \leq B, f(x) = 0$ otherwise. Show that the moment generating function of X is

$$M_X(t) = \frac{e^{Bt} - e^{At}}{(B - A)t} \qquad t \neq 0$$

33. Let $X \sim \text{Unif}[0, 1]$. Find a linear function $Y = g(X)$ such that the interval $[0, 1]$ is transformed into $[-5, 5]$. Use the relationship for linear functions $M_{aX+b}(t) = e^{bt}M_X(at)$ to obtain the mgf of Y from the mgf of X. Compare your answer with the result of Exercise 32, and use this to obtain the pdf of Y.

34. If the pdf of a measurement error X is $f(x) = .5e^{-|x|}$, $-\infty < x < \infty$ show that $M_X(t) = \dfrac{1}{1 - t^2}$ for $|t| < 1$.

35. Consider the rv X = hazardous flood rate in Example 4.5.

    a. Find the moment generating function and use it to find the mean and variance.
    b. Now consider a random variable whose pdf is

    $$f(x) = .04e^{-.04x} \quad x \geq 0$$

    Find the moment generating function and use it to find the mean and variance. Compare with (a), and explain the similarities and differences.
    c. Let $Y = X - 10$ and use the relationship for linear functions $M_{aX+b}(t) = e^{bt}M_X(at)$ to obtain the mgf of Y from (a). Compare with the result of (b) and explain.

36. Define $R_X(t) = \ln[M_X(t)]$. It was shown in Chapter 3 that $R'_X(t) = E(X)$ and $R''_X(t) = V(X)$.

    a. Determine $M_X(t)$ for the pdf in Exercise 29, and use this mgf to obtain $E(X)$ and $V(X)$. How does this compare, in terms of difficulty, with the integration by parts required in that exercise?
    b. Determine $R_X(t)$ for this same distribution, and use $R_X(t)$ to obtain $E(X)$ and $V(X)$. How does the computational effort here compare with that of (a)?

37. Let X be a nonnegative, continuous rv with pdf $f(x)$ and cdf $F(x)$.

    a. Show that, for any constant $t > 0$,
    $$\int_t^\infty x \cdot f(x)dx \geq t \cdot P(X > t) = t \cdot [1 - F(t)]$$
    b. Assume the mean of X is finite (i.e., the integral defining μ converges). Use part (a) to show that

    $$\lim_{t \to \infty} t \cdot [1 - F(t)] = 0$$

38. Let X be a nonnegative, continuous rv with cdf $F(x)$.

    a. Assuming the mean μ of X is finite, show that

    $$\mu = \int_0^\infty [1 - F(x)]dx$$

    [*Hint*: Apply integration by parts to the integral above, and use the result of the previous exercise.]
    b. A similar argument can be used to show that the kth moment of X is given by

    $$E(X^k) = k \int_0^\infty x^{k-1}[1 - F(x)]dx$$

    and that $E(X^k)$ exists iff $t^k[1 - F(t)] \to 0$ as $t \to \infty$. (This was the topic of a 2012 article in *The American Statistician*.) Suppose the lifetime X, in weeks, of a low-grade transistor under continuous use has cdf $F(x) = 1 - (x + 1)^{-3}$ for $x > 0$. Without finding the pdf of X, determine its mean and its standard deviation.

4.3   The Normal Distribution

The normal distribution is the most important one in all of probability and statistics. Many numerical populations have distributions that can be fit very closely by an appropriate normal curve. Examples include heights, weights, and other physical characteristics, measurement errors in scientific experiments, measurements on fossils, reaction times in psychological experiments, scores on various tests, and numerous economic measures and indicators. Even when the underlying distribution is discrete, the normal curve often gives an excellent approximation. In addition, even when individual variables themselves are not normally distributed, sums and averages of the variables will under suitable conditions have approximately a normal distribution; this is the content of the Central Limit Theorem discussed in Chapter 6.

DEFINITION   A continuous rv X is said to have a **normal distribution** with parameters μ and σ, where $-\infty < \mu < \infty$ and $\sigma > 0$, if the pdf of X is

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty \tag{4.3}$$

The statement that X is normally distributed with parameters μ and σ will be denoted by $X \sim N(\mu, \sigma)$.

Figure 4.13 presents graphs of $f(x; \mu, \sigma)$ for several different (μ, σ) pairs. Each resulting density curve is symmetric about μ and bell-shaped, so the center of the bell (point of symmetry) is both the mean of the distribution and the median. The value of σ is the distance from μ to the inflection points of the curve (the points at which the curve changes between turning downward to turning upward). Large values of σ yield density curves that are quite spread out about μ, whereas small values of σ yield density curves with a high peak above μ and most of the area under the density curve quite close to μ. Thus a large σ implies that a value of X far from μ may well be observed, whereas such a value is quite unlikely when σ is small.
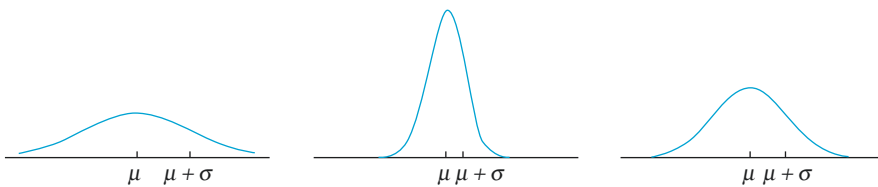
Figure 4.13   Normal density curves

Clearly $f(x; \mu, \sigma) \geq 0$, but a clever calculus argument is required to prove that $\int_{-\infty}^{\infty} f(x; \mu, \sigma)dx = 1$ (see Exercise 66). It can be shown using calculus (Exercise 67) or moment generating functions (Exercise 68) that $E(X) = \mu$ and $V(X) = \sigma^2$, so the parameters μ and σ are the mean and the standard deviation, respectively, of X.

The Standard Normal Distribution

To compute $P(a \leq X \leq b)$ when $X \sim N(\mu, \sigma)$, we must evaluate

$$\int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx \tag{4.4}$$

None of the standard integration techniques can be used to evaluate (4.4), and there is no closed-form expression for (4.4). Table 4.2 at the end of this section provides the code for performing such normal calculations in R. For the purpose of hand calculation, we now introduce a special normal distribution.

DEFINITION   The normal distribution with parameter values $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution**. A random variable that has a standard normal distribution is called a **standard normal random variable** and will be denoted by Z. The pdf of Z, denoted $\phi(z)$, is

$$\phi(z) = f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty < z < \infty$$

The cdf of Z is $P(Z \leq z) = \int_{-\infty}^z \phi(y)dy = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$, which we will denote by $\Phi(z)$.

The standard normal distribution does not frequently serve as a model for a naturally arising population, since few variables have mean 0 and standard deviation 1. Instead, it is a reference distribution from which information about other normal distributions can be obtained. Appendix Table A.3 gives values of $\Phi(z)$ for $z = -3.49, -3.48, \ldots, 3.48, 3.49$ and is referred to as the *standard normal table* or *z table*. Figure 4.14 illustrates the type of cumulative area (probability) tabulated in Table A.3. From this table, various other probabilities involving Z can be calculated.
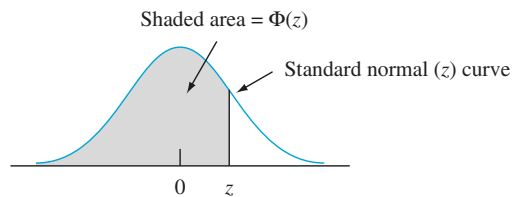
Figure 4.14 Standard normal cumulative areas tabulated in Appendix Table A.3

Example 4.20 Here we demonstrate how the z table is used to calculate various probabilities involving a standard normal rv.

a. $P(Z \leq 1.25) = \Phi(1.25)$, a probability that is tabulated in Table A.3 at the intersection of the row marked 1.2 and the column marked .05. The number there is .8944, so $P(Z \leq 1.25) = .8944$. See Figure 4.15a. In R, we may type pnorm(1.25,0,1) or just pnorm(1.25).

a

Shaded area = $\Phi(1.25)$

z curve

0   1.25

b

z curve

0   1.25
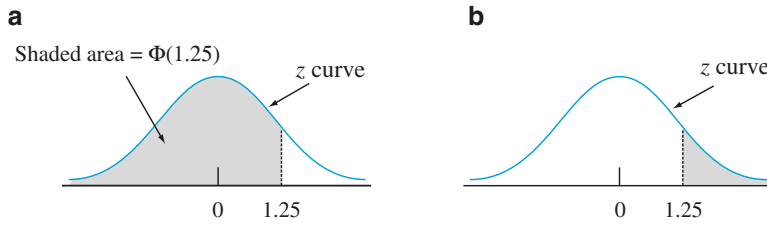
Figure 4.15 Normal curve areas (probabilities) for Example 4.20

b. $P(Z > 1.25) = 1 - P(Z \le 1.25) = 1 - \Phi(1.25)$, the area under the standard normal curve to the right of 1.25 (an upper-tail area). Since $\Phi(1.25) = .8944$, it follows that $P(Z > 1.25) = .1056$. Since Z is a continuous rv, $P(Z \ge 1.25)$ also equals .1056. See Figure 4.15b.

c. $P(Z \le -1.25) = \Phi(-1.25)$, a lower-tail area. Directly from the z table, $\Phi(-1.25) = .1056$. By symmetry of the normal curve, this is identical to the probability in (b).

d. $P(-.38 \le Z \le 1.25)$ is the area under the standard normal curve above the interval [–.38, 1.25]. From Section 4.1, if Z is a continuous rv with cdf $F(z)$, then $P(a \le Z \le b) = F(b) - F(a)$. This gives $P(-.38 \le Z \le 1.25) = \Phi(1.25) - \Phi(-.38) = .8944 - .3520 = .5424$ (see Figure 4.16). To evaluate this probability in R, type `pnorm(1.25, 0,1)-pnorm(-.38, 0, 1)` or just `pnorm(1.25)-pnorm(-.38)`.

z curve

=   −

−.38 0   1.25   0   1.25   −.38 0
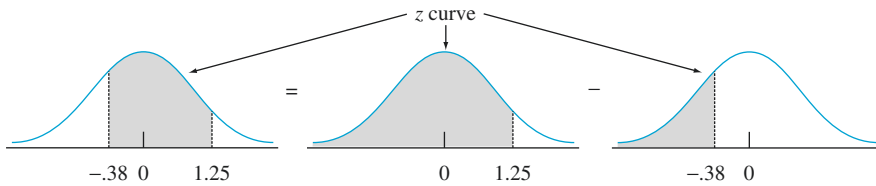
Figure 4.16 $P(-.38 \le Z \le 1.25)$ as the difference between two cumulative areas   ∎

From Section 4.1, we have that the $(100p)$th percentile of the standard normal distribution, for any p between 0 and 1, is the solution to the equation $\Phi(z) = p$. So, we may write the $(100p)$th percentile of the standard normal distribution as $\eta_p = \Phi^{-1}(p)$. Software or the z table can be used to obtain this percentile.

Example 4.21 The 99th percentile of the standard normal distribution is that value on the horizontal axis such that the area under the curve to the left of the value is .9900. Appendix Table A.3 gives for fixed z the area under the standard normal curve to the left of z, whereas here we have the area and want the value of z. This is the "inverse" problem to $P(Z \le z) = ?$ so the table is used in an inverse fashion: Find in the middle of the table .9900; the row and column in which it lies identify the 99th z percentile. Here .9901 lies in the row marked 2.3 and column marked .03, so the 99th percentile is (approximately) $z = 2.33$ (see Figure 4.17). By symmetry, the first percentile is the negative of the 99th percentile, so it equals −2.33 (1% lies below the first and above the 99th). See Figure 4.18.
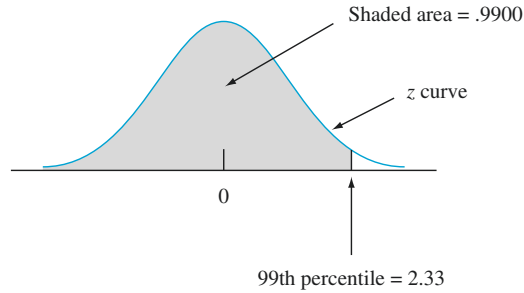
Figure 4.17 Finding the 99th percentile
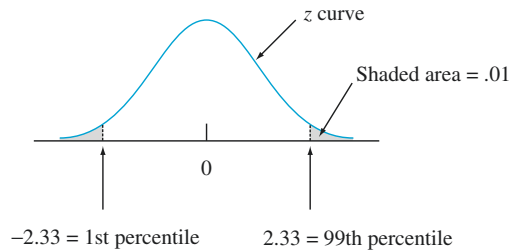
Figure 4.18 The relationship between the 1st and 99th percentiles

To determine the 99th percentile of the standard normal distribution in R, use the command `qnorm(.99,0,1)` or just `qnorm(.99)`.   ∎

In general, the $(100p)$th percentile is identified by the row and column of Appendix Table A.3 in which the entry p is found (e.g., the 67th percentile is obtained by finding .6700 in the body of the table, which gives $z = .44$). If p does not appear, the number closest to it is often used, although linear interpolation gives a more accurate answer. For example, to find the 95th percentile, we look for .9500 inside the table. Although .9500 does not appear, both .9495 and .9505 do, corresponding to $z = 1.64$ and 1.65, respectively. Since .9500 is halfway between the two probabilities that do appear, we will use 1.645 as the 95th percentile and -1.645 as the 5th percentile.

z_α Notation
In statistical inference, we will need the values on the measurement axis that capture certain small tail areas under the standard normal curve.

NOTATION z_α will denote the value on the measurement axis for which α of the area under the z curve lies to the right of z_α. That is, $z_\alpha = \Phi^{-1}(1 - \alpha)$ (see Figure 4.19).
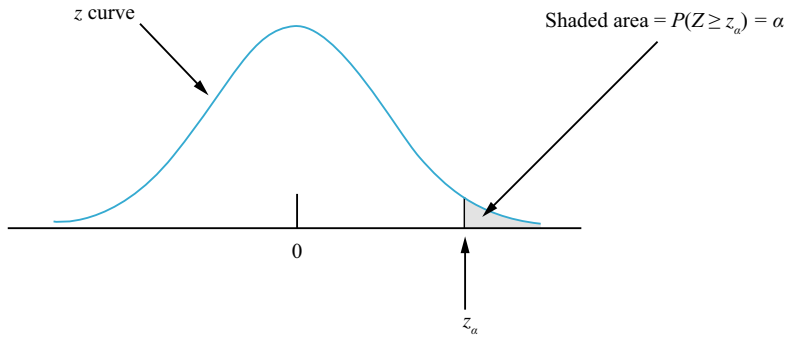
Figure 4.19 z_α notation illustrated

For example, $z_{.10}$ captures upper-tail area .10 and $z_{.01}$ captures upper-tail area .01.

Since α of the area under the standard normal curve lies to the right of z_α, $1 - \alpha$ of the area lies to the left of z_α. Thus z_α *is the* $100(1 - \alpha)th$ *percentile of the standard normal distribution*. By symmetry the area under the standard normal curve to the left of $-z_\alpha$ is also α. The z_α's are usually referred to as **z critical values**. Table 4.1 lists the most useful standard normal percentiles and z_α values.

Table 4.1 Standard normal percentiles and critical values

Percentile	90	95	97.5	99	99.5	99.9	99.95
α (tail area)	.1	.05	.025	.01	.005	.001	.0005
$z_\alpha = 100(1 - \alpha)$th percentile	1.28	1.645	1.96	2.33	2.58	3.08	3.27

Example 4.22 The $100(1 - .05)$th = 95th percentile of the standard normal distribution is $z_{.05}$, so $z_{.05} = 1.645$. The area under the standard normal curve to the left of $-z_{.05}$ is also .05. See Figure 4.20.
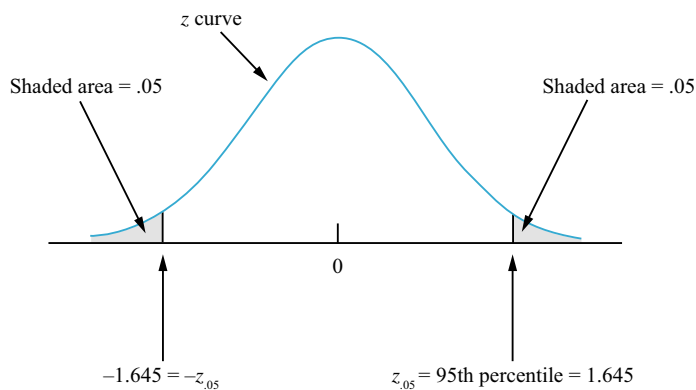
Figure 4.20 Finding $z_{.05}$ ∎

Nonstandardized Normal Distributions

When $X \sim N(\mu, \sigma)$, probabilities involving X may be computed by "standardizing." A **standardized variable** has the form $(X - \mu)/\sigma$. Subtracting μ shifts the mean from μ to zero; dividing by σ scales the variable so that the standard deviation is 1 rather than σ.

Standardizing amounts to calculating a distance from the mean value and then re-expressing the distance as some number of standard deviations. For example, if $\mu = 100$ and $\sigma = 15$, then $x = 130$ corresponds to $z = (130 - 100)/15 = 30/15 = 2.00$. Thus 130 is 2 standard deviations above (i.e., to the right of) the mean value. Similarly, standardizing 85 gives $(85 - 100)/15 = -1.00$, so 85 is 1 standard deviation below the mean. According to the next proposition, the z table applies to *any* normal distribution provided that we think in terms of number of standard deviations away from the mean value.

PROPOSITION   If $X \sim N(\mu, \sigma)$, then the "standardized" rv Z defined by

$$Z = \frac{X - \mu}{\sigma}$$

has a standard normal distribution. Thus

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right),$$

$$P(X \leq a) = \Phi\left(\frac{a - \mu}{\sigma}\right), \quad P(X \geq b) = 1 - \Phi\left(\frac{b - \mu}{\sigma}\right),$$

and the $(100p)$th percentile of the $N(\mu, \sigma)$ distribution is given by

$$\eta_p = \mu + \Phi^{-1}(p) \cdot \sigma.$$

Conversely, if $Z \sim N(0, 1)$ and μ and σ are constants (with $\sigma > 0$), then the "unstandardized" rv $X = \mu + \sigma Z$ has a normal distribution with mean μ and standard deviation σ.

Proof Let $X \sim N(\mu, \sigma)$. Then the cdf of $Z = (X - \mu)/\sigma$ is given by

$$\begin{aligned}
F_Z(z) &= P(Z \leq z) \\
&= P\left(\frac{X - \mu}{\sigma} \leq z\right) \\
&= P(X \leq \mu + z\sigma) = \int_{-\infty}^{\mu + z\sigma} f(x; \mu, \sigma)dx = \int_{-\infty}^{\mu + z\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}dx
\end{aligned}$$

Now make the substitution $u = (x - \mu)/\sigma$. The new limits of integration become $-\infty$ to z, and the differential dx is replaced by $\sigma\, du$, resulting in

$$F_Z(z) = \int\limits_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} e^{-u^2/2} \sigma du = \int\limits_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \Phi(z)$$

Thus, the cdf of $(X - \mu)/\sigma$ is the standard normal cdf, so $(X - \mu)/\sigma \sim N(0, 1)$.

The probability formulas in the statement of the proposition follow directly from this main result, as does the formula for the $(100p)$th percentile:

$$p = P(X \le \eta_p) = P\left(\frac{X - \mu}{\sigma} \le \frac{\eta_p - \mu}{\sigma}\right) = \Phi\left(\frac{\eta_p - \mu}{\sigma}\right) \Rightarrow \frac{\eta_p - \mu}{\sigma} = \Phi^{-1}(p) \Rightarrow \eta_p = \mu + \Phi^{-1}(p) \cdot \sigma$$

The converse statement $Z \sim N(0, 1) \Rightarrow \mu + \sigma Z \sim N(\mu, \sigma)$ is derived similarly. ∎

The key idea of this proposition is that by standardizing, any probability involving X can be expressed as a probability involving a standard normal rv Z, so that the z table can be used. This is illustrated in Figure 4.21.
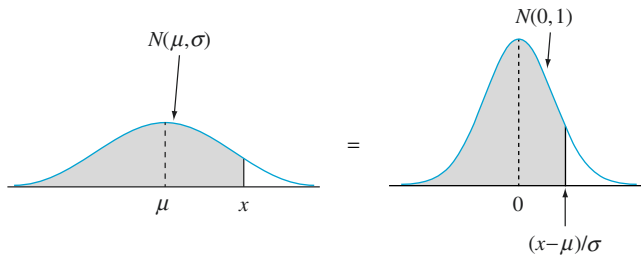
Figure 4.21 Equality of nonstandard and standard normal curve areas

Software eliminates the need for standardizing X, although the standard normal distribution is still important in its own right. Table 4.2 at the end of this section details the relevant R commands, which are also illustrated in the following examples.

Example 4.23 The authors of the article "Assessing the Importance of Surgeon Hand Anthropometry on the Design of Medical Devices" (*J. Med. Devices* 2017) investigate whether standard surgical instruments, such as some surgical staplers, might be too large for some physicians' hands. According to their research, the proximal grip distance (a measure of one's index finger) for male surgeons follows a normal distribution with mean 7.20 cm and standard deviation 0.51 cm. To use one particular stapler, the surgeon's proximal grip distance must be at least 6.83 cm. What is the probability a male surgeon's hand is large enough to use this stapler? If we let X denote the proximal grip distance of a randomly selected male surgeon, then standardizing gives $X \ge 6.83$ if and only if

$$\frac{X - 7.20}{0.51} \ge \frac{6.83 - 7.20}{0.51}$$

Thus

$$P(X \ge 6.83) = P\left(\frac{X - 7.20}{0.51} \ge \frac{6.83 - 7.20}{0.51}\right) = P(Z \ge -0.73)$$
$$= 1 - P(Z < -0.73) = 1 - \Phi(-0.73) = 1 - .2327 = .7673$$

This is illustrated in Figure 4.22. In other words, nearly a quarter of male surgeons would not be able to use this particular surgical stapler, because their hands are too small (or the stapler is too large, depending on your perspective).
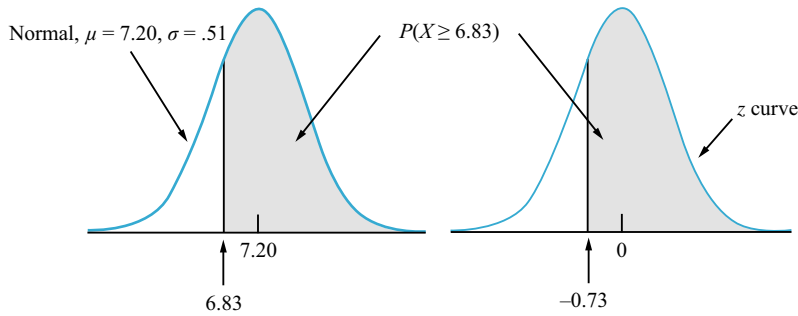
Figure 4.22 Normal curves for Example 4.23

As you might imagine, the situation is worse for female surgeons, whose proximal grip distance distribution can be modeled as $N(6.58, 0.50)$. Denoting the appropriate rv by Y, the probability a female surgeon *cannot* use this stapler is

$$P(Y<6.83) = P\left(\frac{Y-6.58}{0.50} < \frac{6.83-6.58}{0.50}\right) = P(Z<0.5) = \Phi(0.5) = .6915$$

Fortunately, as noted by the authors of the article, another brand of surgical stapler exists for which the required proximal grip distance is only 5.13 cm, meaning that practically all surgeons of either sex can comfortably use this other brand of stapler. ∎

Example 4.24 The amount of distilled water dispensed by a machine is normally distributed with mean value 64 oz and standard deviation .78 oz. What container size c will ensure that overflow occurs only .5% of the time? If X denotes the amount dispensed, the desired condition is that $P(X > c) = .005$, or, equivalently, that $P(X \leq c) = .995$. Thus c is the 99.5th percentile of the normal distribution with $\mu = 64$ and $\sigma = .78$. The 99.5th percentile of the standard normal distribution is $\Phi^{-1}(.995) \approx 2.58$, so

$$c = \eta_{.995} = 64 + (2.58)(.78) = 64 + 2.0 = 66.0 \text{ oz}$$

This is illustrated in Figure 4.23.
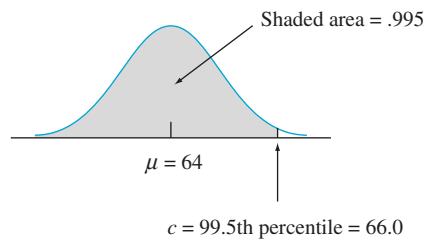
Figure 4.23 Distribution of amount dispensed for Example 4.24

The R command to calculate this percentile is qnorm(.995,64,.78).   ∎

Example 4.25 The return on a diversified investment portfolio is normally distributed. What is the probability that the return is within 1 standard deviation of its mean value? This question can be answered without knowing either μ or σ, as long as the distribution is known to be normal. That is, the answer is the same for *any* normal distribution:

$$
\begin{aligned}
P\left(\begin{array}{c} X \text{ is within one standard} \\ \text{deviation of its mean} \end{array}\right) &= P(\mu - \sigma \le X \le \mu + \sigma) \\
&= P\left(\frac{\mu - \sigma - \mu}{\sigma} \le Z \le \frac{\mu + \sigma - \mu}{\sigma}\right) \\
&= P(-1 \le Z \le 1) \\
&= \Phi(1) - \Phi(-1) = .6826
\end{aligned}
$$

The probability that X is within 2 standard deviations of the mean is $P(-2 \le Z \le 2) = .9544$ and the probability that X is within 3 standard deviations of the mean is $P(-3 \le Z \le 3) = .9973$.   ∎

The results of Example 4.25 are often reported in percentage form and referred to as the *empirical rule* (because empirical evidence has shown that histograms of real data can very frequently be approximated by normal curves).

EMPIRICAL RULE   If the population distribution of a variable is (approximately) normal, then
1. Roughly 68% of the values are within 1 SD of the mean.
2. Roughly 95% of the values are within 2 SDs of the mean.
3. Roughly 99.7% of the values are within 3 SDs of the mean.

It is indeed unusual to observe a value from a normal population that is much farther than 2 standard deviations from μ. These results will be important in the development of hypothesis-testing procedures in later chapters.

The Normal MGF
The moment generating function provides a straightforward way to establish several important results concerning normal distributions.

PROPOSITION   The moment generating function of a normally distributed random variable X is

$$
M_X(t) = e^{\mu t + \sigma^2 t^2 / 2}
$$

Proof Consider first the special case of a standard normal rv Z. Then

$$
M_Z(t) = E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z^2 - 2tz)/2} dz
$$

Completing the square in the exponent, we have

$$M_Z(t) = e^{t^2/2} \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z^2 - 2tz + t^2)/2} dz = e^{t^2/2} \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-t)^2/2} dz$$

The last integral is the area under a normal density with mean t and standard deviation 1, so the value of the integral is 1. Therefore, $M_z(t) = e^{t^2/2}$.

Now let X be any normal rv with mean μ and standard deviation σ. Then, by the proposition earlier in this section, $(X - \mu)/\sigma = Z$, where Z is standard normal. Rewrite this relationship as $X = \mu + \sigma Z$, and use the property $M_{aY+b}(t) = e^{bt} M_Y(at)$:

$$M_X(t) = M_{\mu + \sigma Z}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\sigma^2 t^2/2} = e^{\mu t + \sigma^2 t^2/2} \qquad \blacksquare$$

The normal mgf can be used to establish that μ and σ are indeed the mean and standard deviation of X, as claimed earlier (Exercise 68). Also, by the mgf uniqueness property, any rv X whose moment generating function has the form specified above is necessarily normally distributed. For example, if it is known that the mgf of X is $M_X(t) = e^{8t^2}$, then X must be a normal rv with mean $\mu = 0$ and standard deviation $\sigma = 4$, since the $N(0, 4)$ distribution has e^{8t^2} as its mgf.

It was established earlier in this section that if $X \sim N(\mu, \sigma)$ and $Z = (X - \mu)/\sigma$, then $Z \sim N(0, 1)$, and vice versa. This standardizing transformation is actually a special case of a much more general property.

PROPOSITION   Let $X \sim N(\mu, \sigma)$. Then for any constants a and b with $a \neq 0$, $aX + b$ is also normally distributed. That is, any linear rescaling of a normal rv is normal.

The proof of this proposition uses mgfs and is left as an exercise (Exercise 70). This proposition provides a much easier proof of the earlier relationship between X and Z. The rescaling formulas and this proposition combine to give the following statement: if X is normally distributed and $Y = aX + b$ ($a \neq 0$), then Y is also normal, with mean $\mu_Y = a\mu_X + b$ and standard deviation $\sigma_Y = |a|\sigma_X$.

The Normal Distribution and Discrete Populations
The normal distribution is often used as an approximation to the distribution of values in a discrete population. In such situations, extra care must be taken to ensure that probabilities are computed in an accurate manner.

Example 4.26 IQ (as measured by a standard test) is known to be approximately normally distributed with $\mu = 100$ and $\sigma = 15$. What is the probability that a randomly selected individual has an IQ of at least 125? Letting $X =$ the IQ of a randomly chosen person, we wish $P(X \geq 125)$. The temptation here is to standardize $X \geq 125$ immediately as in previous examples. However, the IQ population is actually discrete, since IQs are integer-valued. So, the normal curve is an approximation to a discrete probability histogram, as pictured in Figure 4.24.
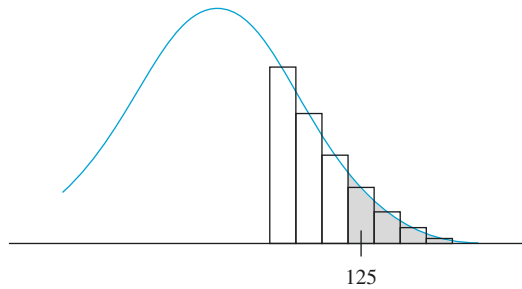
Figure 4.24   A normal approximation to a discrete distribution

The rectangles of the histogram are *centered* at integers, so IQs of at least 125 correspond to rectangles beginning at 124.5, as shaded in Figure 4.24. Thus we really want the area under the approximating normal curve to the right of 124.5. Standardizing this value gives $P(Z \geq 1.63) = .0516$. If we had standardized $X \geq 125$, we would have obtained $P(Z \geq 1.67) = .0475$. The difference is not great, but the answer .0516 is more accurate. Similarly, $P(X = 125)$ would be approximated by the area between 124.5 and 125.5, since the area under the normal curve above the single value 125 is zero. ∎

The correction for discreteness of the underlying distribution in Example 4.26 is often called a **continuity correction**. It is useful in the following application of the normal distribution to the computation of binomial probabilities. The normal distribution was actually created as an approximation to the binomial distribution (by Abraham de Moivre in the 1730s).

Approximating the Binomial Distribution

Recall that the mean value and standard deviation of a binomial random variable X are $\mu = np$ and $\sigma = \sqrt{npq}$, respectively. Figure 4.25a (p. 224) displays a probability histogram for the binomial distribution with $n = 20$, $p = .6$ [so $\mu = 20(.6) = 12$ and $\sigma = \sqrt{20(.6)(.4)} = 2.19$]. A normal curve with mean value and standard deviation equal to the corresponding values for the binomial distribution has been superimposed on the probability histogram. Although the probability histogram is a bit skewed (because $p \neq .5$), the normal curve gives a very good approximation, especially in the middle part of the picture. The area of any rectangle (probability of any particular X value) except those in the extreme tails can be accurately approximated by the corresponding normal curve area. Thus $P(X = 10) = \binom{20}{10}(.6)^{10}(.4)^{10} = .117$, whereas the area under the normal curve between 9.5 and 10.5 is $P(-1.14 \leq Z \leq -.68) = .120$.

On the other hand, a normal distribution is a poor approximation to a discrete distribution that is heavily skewed. For example, Figure 4.25b shows a probability histogram for the Bin(20, .1) distribution and the normal pdf with the same mean and standard deviation ($\mu = 2$ and $\sigma = 1.34$). Clearly, we would not want to use this normal curve to estimate binomial probabilities, even with a continuity correction.
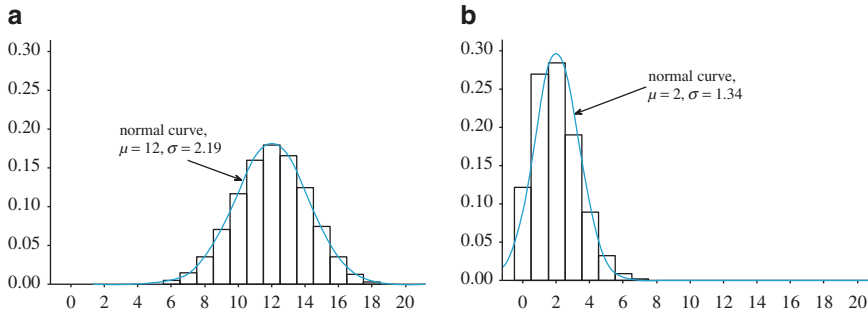
Figure 4.25   Binomial probability histograms with normal approximation curves superimposed: (**a**) $n = 20$ and $p = .6$ (a good fit); (**b**) $n = 20$ and $p = .1$ (a poor fit)

PROPOSITION	Let X be a binomial rv based on n trials with success probability p. Then if the binomial probability histogram is not too skewed, X has approximately a normal distribution with $\mu = np$ and $\sigma = \sqrt{npq}$. In particular, for $x = $ a possible value of X,

$$P(X \leq x) = B(x; n, p) \approx \text{(area under the normal curve to the left of } x + .5)$$
$$= \Phi\left(\frac{x + .5 - np}{\sqrt{npq}}\right)$$

In practice, the approximation is adequate provided that both $np \geq 10$ and $nq \geq 10$.

If either $np < 10$ or $nq < 10$, the binomial distribution may be too skewed for the (symmetric) normal curve to give accurate approximations.

Example 4.27   Suppose that 25% of all licensed drivers in a state do not have insurance. Let X be the number of uninsured drivers in a random sample of size 50 (somewhat perversely, a success is an uninsured driver), so that $p = .25$. Then $\mu = 12.5$ and $\sigma = 3.062$. Since $np = 50(.25) = 12.5 \geq 10$ and $nq = 37.5 \geq 10$, the approximation can safely be applied:

$$P(X \leq 10) = B(10; 50, .25) \approx \Phi\left(\frac{10 + .5 - 12.5}{3.062}\right)$$
$$= \Phi(-.65) = .2578$$

Similarly, the probability that between 5 and 15 (inclusive) of the selected drivers are uninsured is

$$P(5 \leq X \leq 15) = B(15; 50, .25) - B(4; 50, .25)$$
$$\approx \Phi\left(\frac{15.5 - 12.5}{3.062}\right) - \Phi\left(\frac{4.5 - 12.5}{3.062}\right) = .8320$$

The exact probabilities are .2622 and .8348, respectively, so the approximations are quite good. In the last calculation, the probability $P(5 \leq X \leq 15)$ is being approximated by the area under the normal curve between 4.5 and 15.5—the continuity correction is used for both the upper and lower limits. ∎

The wide availability of software for doing binomial probability calculations, even for large values of n, has considerably diminished the importance of the normal approximation. However, it is important for another reason. When the objective of an investigation is to make an inference about a population proportion p, interest will focus on the sample proportion of successes $\hat{P} = X/n$ rather than on X itself. Because this proportion is just X multiplied by the constant $1/n$, the earlier rescaling proposition tells us that $\hat{P}$ will also have approximately a normal distribution (with mean $\mu = p$ and standard deviation $\sigma = \sqrt{pq/n}$) provided that both $np \geq 10$ and $nq \geq 10$. This normal approximation is the basis for several inferential procedures to be discussed in later chapters.

It is quite difficult to give a direct proof of the validity of this normal approximation (the first one goes back almost 300 years to de Moivre). In Chapter 6, we'll see that it is a consequence of an important general result called the Central Limit Theorem.

Normal Distribution Calculations with Software

Many software packages, including R, have built-in functions to determine both probabilities under a normal curve and quantiles (aka percentiles) of any given normal distribution. Table 4.2 summarizes the relevant R code.

Table 4.2 Normal probability and quantile calculations in R

Function	cdf	Quantile; i.e., the $(100p)$th percentile
Notation	$\Phi\left(\frac{x-\mu}{\sigma}\right)$	$\eta_p = \mu + \Phi^{-1}(p) \cdot \sigma$
R	$\texttt{pnorm}(x,\mu,\sigma)$	$\texttt{qnorm}(p,\mu,\sigma)$

In the special case of a standard normal distribution, R will allow the user to drop the last two arguments, μ and σ. That is, the R commands $\texttt{pnorm(x)}$ and $\texttt{pnorm(x,0,1)}$ yield the same result for any number $\texttt{x}$, and a similar comment applies to $\texttt{qnorm}$. R also has a built-in function for the normal pdf: $\texttt{dnorm}(x,\mu,\sigma)$. However, this function is generally only used when one desires to *graph* a normal density curve, x vs. $f(x; \mu, \sigma)$, since the pdf evaluated at particular x does not represent a probability (as discussed in Section 4.1).

Exercises: Section 4.3 (39–70)

39. Let Z be a standard normal random variable and calculate the following probabilities, drawing pictures wherever appropriate.

   a. $P(0 \leq Z \leq 2.17)$
   b. $P(0 \leq Z \leq 1)$
   c. $P(-2.50 \leq Z \leq 0)$
   d. $P(-2.50 \leq Z \leq 2.50)$
   e. $P(Z \leq 1.37)$
   f. $P(-1.75 \leq Z)$
   g. $P(-1.50 \leq Z \leq 2.00)$
   h. $P(1.37 \leq Z \leq 2.50)$
   i. $P(1.50 \leq Z)$
   j. $P(|Z| \leq 2.50)$

40. In each case, determine the value of the constant c that makes the probability statement correct.

   a. $\Phi(c) = .9838$
   b. $P(0 \leq Z \leq c) = .291$
   c. $P(c \leq Z) = .121$
   d. $P(-c \leq Z \leq c) = .668$
   e. $P(c \leq |Z|) = .016$

41. Find the following percentiles for the standard normal distribution. Interpolate where appropriate.

   a. 91st
   b. 9th

c. 75th

d. 25th

e. 6th

42. Determine z_α for the following:

a. $\alpha = .0055$

b. $\alpha = .09$

c. $\alpha = .663$

43. If X is a normal rv with mean 80 and standard deviation 10, compute the following probabilities by standardizing:

a. $P(X \leq 100)$

b. $P(X \leq 80)$

c. $P(65 \leq X \leq 100)$

d. $P(70 \leq X)$

e. $P(85 \leq X \leq 95)$

f. $P(|X - 80| \leq 10)$

44. The plasma cholesterol level (mg/dL) for patients with no prior evidence of heart disease who experience chest pain is normally distributed with mean 200 and standard deviation 35. Consider randomly selecting an individual of this type. What is the probability that the plasma cholesterol level

a. Is at most 250?

b. Is between 300 and 400?

c. Differs from the mean by at least 1.5 standard deviations?

45. The article "Reliability of Domestic-Waste Biofilm Reactors" (*J. Envir. Engr.* 1995: 785–790) suggests that substrate concentration (mg/cm^3) of influent to a reactor is normally distributed with $\mu = .30$ and $\sigma = .06$.

a. What is the probability that the concentration exceeds .25?

b. What is the probability that the concentration is at most .10?

c. How would you characterize the largest 5% of all concentration values?

46. Suppose the diameter at breast height (in.) of trees of a certain type is normally distributed with $\mu = 8.8$ and $\sigma = 2.8$, as suggested in the article "Simulating a Harvester-Forwarder Softwood Thinning" (*Forest Products J.*, May 1997: 36–41).

a. What is the probability that the diameter of a randomly selected tree will be at least 10 in.? Will exceed 10 in.?

b. What is the probability that the diameter of a randomly selected tree will exceed 20 in.?

c. What is the probability that the diameter of a randomly selected tree will be between 5 and 10 in.?

d. What value c is such that the interval $(8.8 - c, \ 8.8 + c)$ includes 98% of all diameter values?

e. If four trees are independently selected, what is the probability that at least one has a diameter exceeding 10 in.?

47. There are two machines available for cutting corks intended for use in wine bottles. The first produces corks with diameters that are normally distributed with mean 3 cm and standard deviation .1 cm. The second machine produces corks with diameters that have a normal distribution with mean 3.04 cm and standard deviation .02 cm. Acceptable corks have diameters between 2.9 and 3.1 cm. Which machine is more likely to produce an acceptable cork?

48. Human body temperatures for healthy individuals have approximately a normal distribution with mean 98.25 °F and standard deviation .75 °F. (The past accepted value of 98.6 °F was obtained by converting the Celsius value of 37°, which is correct to the nearest integer.)

a. Find the 90th percentile of the distribution.

b. Find the 5th percentile of the distribution.

c. What temperature separates the coolest 25% from the others?

49. The article "Monte Carlo Simulation—Tool for Better Understanding of LRFD" (*J. Struct. Engr.* 1993: 1586–1599) suggests that yield strength (ksi) for A36 grade

steel is normally distributed with $\mu = 43$ and $\sigma = 4.5$.

a. What is the probability that yield strength is at most 40? Greater than 60?

b. What yield strength value separates the strongest 75% from the others?

50. The automatic opening device of a military cargo parachute has been designed to open when the parachute is 200 m above the ground. Suppose opening altitude actually has a normal distribution with mean value 200 m and standard deviation 30 m. Equipment damage will occur if the parachute opens at an altitude of less than 100 m. What is the probability that there is equipment damage to the payload of at least 1 of 5 independently dropped parachutes?

51. The temperature reading from a thermocouple placed in a constant temperature medium is normally distributed with mean μ, the actual temperature of the medium, and standard deviation σ. What would the value of σ have to be to ensure that 95% of all readings are within $.1°$ of μ?

52. The distribution of resistance for resistors of a certain type is known to be normal, with 10% of all resistors having a resistance exceeding $10.256\ \Omega$ and 5% having a resistance smaller than 9.671 ohms. What are the mean value and standard deviation of the resistance distribution?

53. If adult female heights are normally distributed, what is the probability that the height of a randomly selected woman is

a. Within 1.5 SDs of its mean value?

b. Farther than 2.5 SDs from its mean value?

c. Between 1 and 2 SDs from its mean value?

54. A machine that produces ball bearings has initially been set so that the true average diameter of the bearings it produces is .500 in. A bearing is acceptable if its diameter is within .004 in. of this target value. Suppose, however, that the setting has changed during the course of production, so that the bearings have normally distributed diameters with mean value .499 in. and standard deviation .002 in. What percentage of the bearings produced will not be acceptable?

55. The Rockwell hardness of a metal is determined by impressing a hardened point into the surface of the metal and then measuring the depth of penetration of the point. Suppose the Rockwell hardness of an alloy is normally distributed with mean 70 and standard deviation 3. (Rockwell hardness is measured on a continuous scale.)

a. If a specimen is acceptable only if its hardness is between 67 and 75, what is the probability that a randomly chosen specimen has an acceptable hardness?

b. If the acceptable range of hardness is $(70 - c,\ 70 + c)$, for what value of c would 95% of all specimens have acceptable hardness?

c. If the acceptable range is as in part (a) and the hardness of each of ten randomly selected specimens is independently determined, what is the expected number of acceptable specimens among the ten?

d. What is the probability that at most 8 of 10 independently selected specimens have a hardness of less than 73.84? [*Hint*: $Y =$ the number among the ten specimens with hardness less than 73.84 is a binomial variable; what is p?]

56. The weight distribution of parcels sent in a certain manner is normal with mean value 12 lb and standard deviation 3.5 lb. The parcel service wishes to establish a weight value c beyond which there will be a surcharge. What value of c is such that 99% of all parcels are at least 1 lb under the surcharge weight?

57. Suppose Appendix Table A.3 contained $\Phi(z)$ only for $z \geq 0$. Explain how you could still compute

    a. $P(-1.72 \leq Z \leq -.55)$
    b. $P(-1.72 \leq Z \leq .55)$

    Is it necessary to table $\Phi(z)$ for z negative? What property of the standard normal curve justifies your answer?

58. Let X be the birth weight, in grams, of a randomly selected full-term baby. The article "Fetal Growth Parameters and Birth Weight: Their Relationship to Neonatal Body Composition" (*Ultrasound Obstetrics Gynecol.* 2009: 441–446) suggests that X is normally distributed with mean 3500 and standard deviation 600.

    a. Sketch the relevant density curve, including tick marks on the horizontal scale.
    b. What is $P(3000 < X < 4500)$, and how does this compare to $P(3000 \leq X \leq 4500)$?
    c. What is the probability that the weight of such a newborn is less than 2500 g?
    d. What is the probability that the weight of such a newborn exceeds 6000 g (roughly 13.2 lb)?
    e. How would you characterize the most extreme .1% of all birth weights?
    f. Use the rescaling proposition from this section to determine the distribution of birth weight expressed in pounds (shape, mean, and standard deviation), and then recalculate the probability from part (c). How does this compare to your previous answer?

59. Based on extensive data from an urban freeway near Toronto, Canada, "it is assumed that free speeds can best be represented by a normal distribution" ["Impact of Driver Compliance on the Safety and Operational Impacts of Freeway Variable Speed Limit Systems" (*J. Transp. Engr.* 2011: 260–268)]. The mean and standard deviation reported in the article were 119 km/h and 13.1 km/h, respectively.

    a. What is the probability that the speed of a randomly selected vehicle is between 100 and 120 km/h?
    b. What speed characterizes the fastest 10% of all speeds?
    c. The posted speed limit was 100 km/h. What percentage of vehicles was traveling at speeds exceeding this posted limit?
    d. If five vehicles are randomly and independently selected, what is the probability that at least one is not exceeding the posted speed limit?
    e. What is the probability that the speed of a randomly selected vehicle exceeds 70 *miles* per hour?

60. Chebyshev's inequality, introduced in Chapter 3 Exercise 45, is valid for continuous as well as discrete distributions. It states that for any number $k \geq 1$, $P(|X - \mu| \geq k\sigma) \leq 1/k^2$ (see the aforementioned exercise for an interpretation and Chapter 3 Exercise 163 for a proof). Obtain this probability in the case of a normal distribution for $k = 1$, 2, and 3, and compare to the Chebyshev upper bound.

61. Let X denote the number of flaws along a 100-m reel of magnetic tape (an integer-valued variable). Suppose X has approximately a normal distribution with $\mu = 25$ and $\sigma = 5$. Use the continuity correction to calculate the probability that the number of flaws is

    a. Between 20 and 30, inclusive.
    b. At most 30. Less than 30.

62. Let X have a binomial distribution with parameters $n = 25$ and p. Calculate each of the following probabilities using the normal approximation (with the continuity correction) for the cases $p = .5, .6,$ and $.8$ and compare to the exact probabilities calculated from Appendix Table A.1.

    a. $P(15 \leq X \leq 20)$
    b. $P(X \leq 15)$
    c. $P(20 \leq X)$

63. Suppose that 10% of all steel shafts produced by a process are nonconforming but can be reworked (rather than having to be scrapped). Consider a random sample of 200 shafts, and let X denote the number among these that are nonconforming and can be reworked. What is the (approximate) probability that X is

    a. At most 30?
    b. Less than 30?
    c. Between 15 and 25 (inclusive)?

64. Suppose only 70% of all drivers in a state regularly wear a seat belt. A random sample of 500 drivers is selected. What is the probability that

    a. Between 320 and 370 (inclusive) of the drivers in the sample regularly wear a seat belt?
    b. Fewer than 325 of those in the sample regularly wear a seat belt? Fewer than 315?

65. In response to concerns about nutritional contents of fast foods, McDonald's announced that it would use a new cooking oil for its French fries that would decrease substantially trans-fatty acid levels and increase the amount of more beneficial polyunsaturated fat. The company claimed that 97 out of 100 people cannot detect a difference in taste between the new and old oils. Assuming that this figure is correct (as a long-run proportion), what is the approximate probability that in a random sample of 1000 individuals who have purchased fries at McDonald's,

    a. At least 40 can taste the difference between the two oils?
    b. At most 5% can taste the difference between the two oils?

66. The following proof that the normal pdf integrates to 1 comes courtesy of Professor Robert Young, Oberlin College. Let $f(z)$ denote the standard normal pdf, and consider the function of two variables

$$g(x, y) = f(x) \cdot f(y)$$
$$= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$
$$= \frac{1}{2\pi} e^{-(x^2 + y^2)/2}$$

Let V denote the volume under the graph of $g(x, y)$ above the xy-plane.

    a. Let A denote the area under the standard normal curve. By setting up the double integral for the volume underneath $g(x, y)$, show that $V = A^2$.
    b. Using the rotational symmetry of $g(x, y)$, V can be determined by adding up the volumes of shells from rotation about the y-axis:

$$V = \int\limits_0^\infty 2\pi\, r \cdot \frac{1}{2\pi} e^{-r^2/2} dr$$

Show this integral equals 1, then use (a) to establish that the area under the standard normal curve is 1.

c. Show that $\int_{-\infty}^{\infty} f(x; \mu, \sigma) dx = 1$. [*Hint:* Write out the integral, and then make a substitution to reduce it to the standard normal case. Then invoke (b).]

67. Suppose $X \sim N(\mu, \sigma)$.
    a. Show via integration that $E(X) = \mu$. [*Hint*: Make the substitution $u = (x - \mu)/\sigma$, which will create two integrals. For one, use the symmetry of the pdf; for the other, use the fact that the standard normal pdf integrates to 1.]
    b. Show via integration that $V(X) = \sigma^2$. [*Hint*: Evaluate the integral for $E[(X - \mu)^2]$ rather than using the variance shortcut formula. Use the same substitution as in part (a).]

68. The moment generating function can be used to find the mean and variance of the normal distribution.
    a. Use derivatives of $M_X(t)$ to verify that $E(X) = \mu$ and $V(X) = \sigma^2$.
    b. Repeat (a) using $R_X(t) = \ln[M_X(t)]$, and compare with part (a) in terms of effort.

69. There is no nice formula for the standard normal cdf $\Phi(z)$, but several good approximations have been published in articles. The following is from "Approximations for Hand Calculators Using Small Integer Coefficients" (*Math. Comput.* 1977: 214–222). For $0 < z \leq 5.5$,

$$P(Z \geq z) = 1 - \Phi(z)$$
$$\approx .5 \exp\left\{ -\left[\frac{(83z + 351)z + 562}{(703/z) + 165} \right] \right\}$$

The relative error of this approximation is less than .042%. Use this to calculate approximations to the following probabilities, and compare whenever possible to the probabilities obtained from Appendix Table A.3.

    a. $P(Z \geq 1)$
    b. $P(Z < -3)$
    c. $P(-4 < Z < 4)$
    d. $P(Z > 5)$

70. a. Use mgfs to show that if X has a normal distribution with parameters μ_X and σ_X, then $Y = aX + b$ (a linear function of X) also has a normal distribution. What are the parameters of the distribution of Y [i.e., μ_Y and σ_Y]?
    b. If when measured in °C, temperature is normally distributed with mean 115 and standard deviation 2, what can be said about the distribution of temperature measured in °F?

4.4   The Gamma Distribution and Its Relatives

The graph of any normal pdf is bell-shaped and thus symmetric. But in many situations, the variable of interest to the experimenter might have a skewed distribution. A family of pdfs that yields a wide variety of skewed distributional shapes is the *gamma* family. To define the family of gamma distributions, we first need to introduce a function that plays an important role in many branches of mathematics.

DEFINITION   For $\alpha > 0$, the **gamma function** $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

The most important properties of the gamma function are the following:
1. For any $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$ (via integration by parts)
2. For any positive integer, n, $\Gamma(n) = (n - 1)!$
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

The following proposition will prove useful for several computations that follow.

PROPOSITION   For any $\alpha, \beta > 0$,

$$\int_0^\infty x^{\alpha-1} e^{-x/\beta} dx = \beta^\alpha \Gamma(\alpha) \tag{4.5}$$

Proof Make the substitution $u = x/\beta$, so that $x = \beta u$ and $dx = \beta\ du$:

$$\int_0^\infty x^{\alpha-1} e^{-x/\beta} dx = \int_0^\infty (\beta u)^{\alpha-1} e^{-u} \beta du = \beta^\alpha \int_0^\infty u^{\alpha-1} e^{-u} du = \beta^\alpha \Gamma(\alpha)$$

The last equality comes from the definition of the gamma function.   ■

The Family of Gamma Distributions
With the preceding proposition in mind, we make the following definition.

DEFINITION   A continuous random variable X is said to have a **gamma distribution**
if the pdf of X is

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \quad x > 0 \tag{4.6}$$

where the parameters α and β satisfy $\alpha > 0$, $\beta > 0$. When $\beta = 1$, X is said to
have a **standard gamma distribution**, and its pdf may be denoted $f(x; \alpha)$.

It's clear that $f(x; \alpha, \beta) \geq 0$ for all x; the previous proposition guarantees that this function integrates to 1, as required. Figure 4.26a illustrates the graphs of the gamma pdf for several (α, β) pairs, whereas Figure 4.26b presents graphs of the standard gamma pdf. For the standard pdf, when $\alpha \leq 1, f(x; \alpha)$ is strictly decreasing as x increases; when $\alpha > 1, f(x; \alpha)$ rises to a maximum and then decreases. The parameter β in (4.6) is called a *scale parameter* because values other than 1 either stretch or compress the pdf in the x direction.
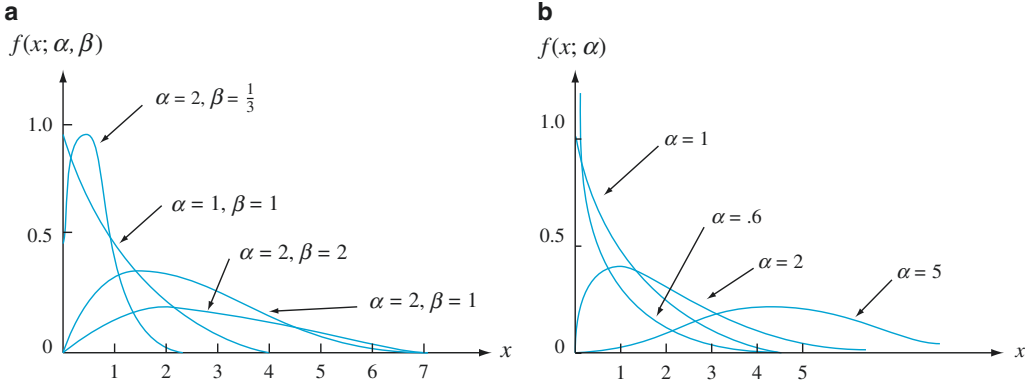
Figure 4.26 (**a**) Gamma density curves; (**b**) standard gamma density curves

---

PROPOSITION   The moment generating function of a gamma random variable is

$$M_X(t) = \frac{1}{(1 - \beta t)^\alpha}$$

Proof By definition, the mgf is

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-x/\beta} dx = \int_0^\infty \frac{x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} e^{-x(-t+1/\beta)} dx$$

Now use Expression (4.5): provided $-t + 1/\beta > 0$, i.e., $t < 1/\beta$,

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-(-t+1/\beta)x} dx = \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot \Gamma(\alpha) \left(\frac{1}{-t+1/\beta}\right)^\alpha = \frac{1}{(1 - \beta t)^\alpha} \qquad ∎$$

The mean and variance can be obtained from the moment generating function (Exercise 82), but they can also be obtained directly through integration (Exercise 83).

PROPOSITION   The mean and variance of a random variable X having the gamma distribution $f(x; \alpha, \beta)$ are

$$E(X) = \mu = \alpha\beta \quad V(X) = \sigma^2 = \alpha\beta^2$$

When X is a standard gamma rv, the cdf of X, which is

$$G(x; \alpha) = \int\limits_0^x \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}\,dy \quad x > 0 \tag{4.7}$$

is called the **incomplete gamma function**. (In mathematics literature, the incomplete gamma function sometimes refers to (4.7) without the denominator $\Gamma(\alpha)$ in the integrand.) In Appendix Table A.4, we present a small tabulation of $G(x; \alpha)$ for $\alpha = 1, 2, \ldots, 10$ and $x = 1, 2, \ldots, 15$. Table 4.3 (p. 236) provides the R commands related to the gamma cdf, which are illustrated in the following examples.

Example 4.28 Suppose the reaction time X (in seconds) of a randomly selected individual to a certain stimulus has a standard gamma distribution with $\alpha = 2$. Since X is continuous,

$$P(3 \le X \le 5) = P(X \le 5) - P(X \le 3) = G(5; 2) - G(3; 2) = .960 - .801 = .159$$

This probability can be obtained in R with `pgamma(5,2) - pgamma(3,2)`.

The probability that the reaction time is more than 4 s is

$$P(X > 4) = 1 - P(X \le 4) = 1 - G(4; 2) = 1 - .908 = .092 \qquad \blacksquare$$

The incomplete gamma function can also be used to compute probabilities involving gamma distributions for any $\beta > 0$.

PROPOSITION   Let X have a gamma distribution with parameters α and β. Then for any $x > 0$, the cdf of X is given by

$$P(X \le x) = G\left(\frac{x}{\beta}; \alpha\right),$$

the incomplete gamma function evaluated at x/β.

The proof is similar to that of Expression (4.5).

Example 4.29 Web servers typically have security algorithms that detect and flag "abnormal" connections from suspicious IP addresses, which can indicate possible hackers. Data from the article "Exact Inferences for a Gamma Distribution" (*J. Quality Technol.* 2014: 140–149) suggests that, for one particular server receiving abnormal connections from one specific IP address, the time X in hours between attempted connections can be modeled using a gamma distribution with $\alpha = 2$ and $\beta = 2.5$. (In fact, the article provides a range of estimates for the parameters; we'll encounter such interval estimates in Chapter 8.) The average time between connections from this suspicious IP address is $E(X) = (2)(2.5) = 5$ h, whereas $V(X) = (2)(2.5)^2 = 12.5$ and $\sigma_X = \sqrt{12.5} \approx 3.5$ h. The probability that a connection from this suspicious IP address will arrive between 5 and 10 h after the previous attempt is

$$P(5 \leq X \leq 10) = P(X \leq 10) - P(X \leq 5)$$
$$= G(10/2.5; 2) - G(5/2.5; 2)$$
$$= G(4; 2) - G(2; 2) = .908 - .594 = .314$$

The probability that two connection attempts from this IP address are separated by more than 15 h is

$$P(X > 15) = 1 - P(X \leq 15)$$
$$= 1 - G(15/2.5; 2) = 1 - G(6; 2) = 1 - .983 = .017$$

Software can also perform these calculations. For instance, the R commands

`pgamma(10,2,1/2.5)-pgamma(5,2,1/2.5)` and `1-pgamma(15,2,1/2.5)`

compute the two probabilities above and return .3144 and .0174, respectively.   ∎

The Exponential Distribution

The family of exponential distributions provides probability models that are widely used in engineering and science disciplines.

DEFINITION   *X* is said to have an **exponential distribution** with parameter λ $(\lambda > 0)$ if the pdf of *X* is

$$f(x; \lambda) = \lambda e^{-\lambda x} \quad x > 0 \tag{4.8}$$

The exponential pdf is a special case of the general gamma pdf (4.6) in which $\alpha = 1$ and $\beta = 1/\lambda$; some sources write the exponential pdf in the form $(1/\beta)e^{-x/\beta}$. The mean and variance of *X* are then

$$\mu = \alpha\beta = \frac{1}{\lambda} \quad \sigma^2 = \alpha\beta^2 = \frac{1}{\lambda^2}$$

Both the mean and standard deviation of the exponential distribution equal $1/\lambda$. Graphs of several exponential pdfs appear in Figure 4.27.
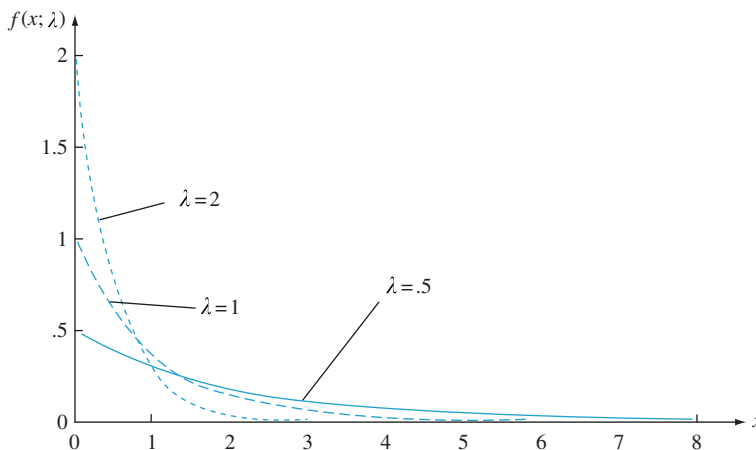
Figure 4.27 Exponential density curves

Unlike the general gamma pdf, the exponential pdf can be easily integrated. In particular, the cdf of X is

$$F(x; \lambda) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

Example 4.30 The response time X at an online computer terminal (the elapsed time between the end of a user's inquiry and the beginning of the system's response to that inquiry) has an exponential distribution with expected response time equal to 5 s. Then $E(X) = 1/\lambda = 5$, so $\lambda = .2$. The probability that the response time is at most 10 s is

$$P(X \leq 10) = F(10; .2) = 1 - e^{-(.2)(10)} = 1 - e^{-2} = 1 - .135 = .865$$

The probability that response time is between 5 and 10 s is

$$P(5 \leq X \leq 10) = F(10; .2) - F(5; .2) = \left(1 - e^{-2}\right) - \left(1 - e^{-1}\right) = .233 \qquad \blacksquare$$

The exponential distribution is frequently used as a model for the distribution of times between the occurrence of successive events, such as customers arriving at a service facility or calls coming into a switchboard. The reason for this is that the exponential distribution is closely related to the Poisson process discussed in Chapter 3.

THEOREM   Suppose that the number of events occurring in any time interval of length t has a Poisson distribution with parameter $\mu = \lambda t$ (where λ, the rate of the event process, is the expected number of events occurring in 1 unit of time) and that numbers of occurrences in nonoverlapping intervals are independent of one another. Then the distribution of elapsed time between the occurrence of two successive events is exponential with parameter λ.

Although a complete proof is beyond the scope of the text, the result is easily verified for the time X_1 until the first event occurs:

$$P(X_1 \leq t) = 1 - P(X_1 > t) = 1 - P(\text{no events in } (0, t])$$
$$= 1 - \frac{e^{-\lambda t} \cdot (\lambda t)^0}{0!} = 1 - e^{-\lambda t}$$

which is exactly the cdf of the exponential distribution.

Example 4.31 Video-on-demand services must carefully model customers' or clients' requests for videos to optimize the use of the available bandwidth. The article "Distributed Client-Assisted Patching for Multicast Video-on-Demand Service in an Enterprise Network" (*J. Comput.* 2017:

511–520) describes a series of experiments in this area, where client requests are modeled by a Poisson process. In one such experiment, the "request rate" was $\lambda = 0.8$ requests per second. Then the time X between successive requests has an exponential distribution with parameter value 0.8. The probability that more than 2 s elapse between requests is

$$P(X > 2) = 1 - P(X \leq 2) = 1 - F(2; 0.8) = e^{-(0.8)(2)} = .202$$

The average time between requests under this setting is $E(X) = 1/\lambda = 1/0.8 = 1.25$ s (you could also deduce this directly from the rate without using the exponential model). ■

Another important application of the exponential distribution is to model the distribution of component lifetime. A partial reason for the popularity of such applications is the "**memoryless**" **property** of the exponential distribution. Suppose component lifetime is exponentially distributed with parameter λ. After putting the component into service, we leave for a period of t_0 h and then return to find the component still working; what now is the probability that it lasts at least an additional t hours? In symbols, we wish $P(X \geq t + t_0 \mid X \geq t_0)$. By the definition of conditional probability,

$$P(X \geq t + t_0 \mid X \geq t_0) = \frac{P[(X \geq t + t_0) \cap (X \geq t_0)]}{P(X \geq t_0)}$$

But the event $X \geq t_0$ in the numerator is redundant, since both events can occur if and only if $X \geq t + t_0$. Therefore,

$$P(X \geq t + t_0 \mid X \geq t_0) = \frac{P(X \geq t + t_0)}{P(X \geq t_0)} = \frac{1 - F(t + t_0; \lambda)}{1 - F(t_0; \lambda)} = \frac{e^{-\lambda(t + t_0)}}{e^{-\lambda t_0}} = e^{-\lambda t}$$

This conditional probability is identical to the original probability $P(X \geq t)$ that the component lasted t hours. Thus *the distribution of additional lifetime is exactly the same as the original distribution of lifetime*, so at each point in time the component shows no effect of wear. In other words, the distribution of remaining lifetime is independent of current age.

Although the memoryless property can be justified at least approximately in many applied problems, in other situations components deteriorate with age or occasionally improve with age (at least up to a certain point). More general lifetime models are then furnished by the gamma, Weibull, and lognormal distributions (the latter two are discussed in the next section).

Gamma and Related Calculations with Software
Table 4.3 summarizes the syntax for the gamma and exponential cdfs in R, which follows the pattern of the other distributions. In a sense, the exponential commands are redundant, since they are just a special case ($\alpha = 1$) of the gamma distribution.

Table 4.3 R code for gamma and exponential cdfs

	Gamma cdf	Exponential cdf
Notation	$G(x/\beta; \alpha)$	$F(x; \lambda) = 1 - e^{-\lambda x}$
R	pgamma $(x, \alpha, 1/\beta)$	pexp (x, λ)

Notice how R parameterizes the distributions: for both the gamma and exponential cdfs, the R functions take as their last input the "rate" parameter $\lambda = 1/\beta$. So, for the gamma rv with parameters $\alpha = 2$ and $\beta = 2.5$ from Example 4.29, $P(X \leq 15)$ would be evaluated as `pgamma(15,2,1/2.5)`. This can be remedied by using a name assignment in the last argument in R; specifically, `pgamma(15,2,scale = 2.5)` will instruct R to use $\beta = 2.5$ in its gamma probability calculation and produce the same answer as the previous expressions. Interestingly, as of this writing the same option does not exist in the `pexp` function.

To graph gamma or exponential density curves, one can request their pdfs in R by replacing the leading letter `p` with `d`. To find quantiles of either of these distributions, the appropriate replacement is `q`. For example, the 75th percentile of the gamma distribution from Example 4.29 can be determined with `qgamma(.75,2,scale = 2.5)`.

Exercises: Section 4.4 (71–83)

71. Evaluate the following:

    a. $\Gamma(6)$
    b. $\Gamma(5/2)$
    c. $G(4; 5)$ (the incomplete gamma function)
    d. $G(5; 4)$
    e. $G(0; 4)$

72. Let X have a standard gamma distribution with $\alpha = 7$. Evaluate the following:

    a. $P(X \leq 5)$
    b. $P(X < 5)$
    c. $P(X > 8)$
    d. $P(3 \leq X \leq 8)$
    e. $P(3 < X < 8)$
    f. $P(X < 4 \text{ or } X > 6)$

73. Suppose the time spent by a randomly selected student at a campus computer laboratory has a gamma distribution with mean 20 min and variance 80 min^2.

    a. What are the values of α and β?
    b. What is the probability that a student uses the laboratory for at most 24 min?
    c. What is the probability that a student spends between 20 and 40 min at the laboratory?

74. Suppose that when a type of transistor is subjected to an accelerated life test, the lifetime X (in weeks) has a gamma distribution with mean 24 weeks and standard deviation 12 weeks.

    a. What is the probability that a transistor will last between 12 and 24 weeks?
    b. What is the probability that a transistor will last at most 24 weeks? Is the median of the lifetime distribution less than 24? Why or why not?
    c. What is the 99th percentile of the lifetime distribution?
    d. Suppose the test will actually be terminated after t weeks. What value of t is such that only .5% of all transistors would still be operating at termination?

75. Let $X =$ the time between two successive arrivals at the drive-up window of a local bank. If X has an exponential distribution with $\lambda = 1$ (which is identical to a standard gamma distribution with $\alpha = 1$), compute the following:

    a. The expected time between two successive arrivals
    b. The standard deviation of the time between successive arrivals
    c. $P(X \leq 4)$
    d. $P(2 \leq X \leq 5)$

76. Let X denote the distance (m) that an animal moves from its birth site to the first

territorial vacancy it encounters. Suppose that for banner-tailed kangaroo rats, X has an exponential distribution with parameter $\lambda = .01386$ (as suggested in the article "Competition and Dispersal from Multiple Nests," *Ecology* 1997: 873–883).

a. What is the probability that the distance is at most 100 m? At most 200 m? Between 100 and 200 m?
b. What is the probability that distance exceeds the mean distance by more than 2 standard deviations?
c. What is the value of the median distance?

77. In studies of anticancer drugs it was found that if mice are injected with cancer cells, the survival time can be modeled with the exponential distribution. Without treatment the expected survival time was 10 h. What is the probability that

a. A randomly selected mouse will survive at least 8 h? At most 12 h? Between 8 and 12 h?
b. The survival time of a mouse exceeds the mean value by more than 2 standard deviations? More than 3 standard deviations?

78. The special case of the gamma distribution in which α is a positive integer n is called an *Erlang distribution*. If we replace β by $1/\lambda$ in Expression (4.6), the Erlang pdf is

$$f(x; \lambda, n) = \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x} \quad x > 0$$

It can be shown that if the times between successive events are independent, each with an exponential distribution with parameter λ, then the total time X that elapses before all of the next n events occur has pdf $f(x; \lambda, n)$.

a. What is the expected value of X? If the time (in minutes) between arrivals of successive customers is exponentially distributed with $\lambda = .5$, how much time can be expected to elapse before the tenth customer arrives?
b. If customer interarrival time is exponentially distributed with $\lambda = .5$, what is the probability that the tenth customer (after the one who has just arrived) will arrive within the next 30 min?
c. The event $\{X \leq t\}$ occurs if and only if at least n events occur in the next t units of time. Use the fact that the number of events occurring in an interval of length t has a Poisson distribution with mean λt to write an expression (involving Poisson probabilities) for the Erlang cumulative distribution function $F(t; \lambda, n) = P(X \leq t)$.

79. A system consists of five identical components connected in series as shown:

As soon as one component fails, the entire system will fail. Suppose each component has a lifetime that is exponentially distributed with $\lambda = .01$ and that components fail independently of one another. Define events $A_i = \{i$th component lasts at least t hours$\}$, $i = 1, \ldots, 5$, so that the A_i's are independent events. Let $X =$ the time at which the system fails—that is, the shortest (minimum) lifetime among the five components.

a. The event $\{X \geq t\}$ is equivalent to what event involving $A_1, \ldots, A_5$?
b. Using the independence of the five A_i's, compute $P(X \geq t)$. Then obtain $F(t) = P(X \leq t)$ and the pdf of X. What type of distribution does X have?
c. Suppose there are n components, each having exponential lifetime with parameter λ. What type of distribution does X have?

80. If X has an exponential distribution with parameter λ, derive a general expression for the $(100p)$th percentile of the distribution. Then specialize to obtain the median.

81. The article "Numerical Prediction of Surface Wear and Roughness Parameters During Running-In for Line Contacts Under Mixed Lubrication" (*J. Tribol.*, Nov. 2018) proposes probability models for several variables that arise in studying the wear of mechanical components like gears and piston rings. These variables include X = wear particle thickness (microns) and W = wear loss (cubic microns).

    a. The article's authors make mathematical arguments that (1) X should follow an exponential distribution and (2) the pdfs of W and X should be related by
    $$f_W(w) \propto w^2 \cdot f_X(w)$$

    What distribution are the authors specifying for W? Identify the name and parameter values of the distribution.

    b. If the variance of W is 3.0 (one of several values considered in the article), what are the numerical values of the parameters of W's distribution, and what is the value of the λ parameter for X's exponential distribution?

82. Determine the mean and variance of the gamma distribution by differentiating the moment generating function $M_X(t)$.

83. Determine the mean and variance of the gamma distribution by first using integration to obtain $E(X)$ and $E(X^2)$. [*Hint*: Express the integrand in terms of a gamma density, and use Expression (4.5).]

4.5 Other Continuous Distributions

The normal, gamma (including exponential), and uniform families of distributions provide a wide variety of probability models for continuous variables, but there are many practical situations in which no member of these families fits a set of observed data very well. Statisticians and other investigators have developed other families of distributions that are often appropriate in practice.

The Weibull Distribution
The family of Weibull distributions was introduced by the Swedish physicist Waloddi Weibull in 1939; his 1951 article "A Statistical Distribution Function of Wide Applicability" (*J. Appl. Mech.* 18: 293–297) discusses a number of applications.

DEFINITION   A random variable X is said to have a **Weibull distribution** with parameters α and β ($\alpha > 0$, $\beta > 0$) if the pdf of X is

$$f(x; \alpha, \beta) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha} \quad x > 0 \tag{4.9}$$

In some situations there are theoretical justifications for the appropriateness of the Weibull distribution, but in many applications $f(x; \alpha, \beta)$ simply provides a good fit to observed data for particular values of α and β. When $\alpha = 1$, the pdf reduces to the exponential distribution (with $\lambda = 1/\beta$), so the exponential distribution is a special case of both the gamma and Weibull distributions. However, there are gamma distributions that are not Weibull distributions and vice versa, so one family is not a

subset of the other. Both α and β can be varied to obtain a number of different distributional shapes, as illustrated in Figure 4.28. Note that β is a scale parameter, so different values stretch or compress the graph in the x-direction.
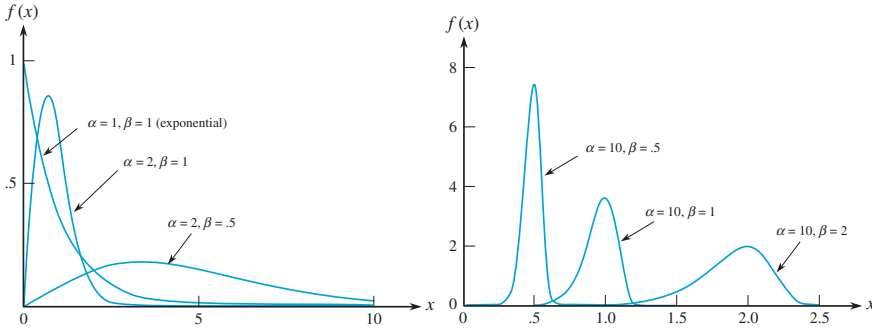
Figure 4.28   Weibull density curves

Integrating to obtain $E(X)$ and $E(X^2)$ yields the mean and variance of X:

$$\mu = \beta\Gamma\left(1 + \frac{1}{\alpha}\right) \quad \sigma^2 = \beta^2\left\{\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2\right\}$$

The computation of μ and σ^2 thus necessitates using the gamma function from the previous section. (The mgf of the Weibull distribution is very complicated, and so we do not include it here.) On the other hand, the integration $\int_0^x f(y; \alpha, \beta)dy$ is easily carried out to obtain the cdf of X:

$$F(x; \alpha, \beta) = \begin{cases} 0 & x < 0 \\ 1 - e^{-(x/\beta)^\alpha} & x \geq 0 \end{cases} \tag{4.10}$$

Example 4.32 One of the most common applications of the Weibull distribution is to model the *time to repair* for some item under industrial use. The article "Supply Chain Inventories of Engineered Shipping Containers" (*Intl. J. Manuf. Engr.* 2016) discusses modeling the time to repair for highly engineered reusable shipping containers, which are quite expensive and need to be monitored carefully. For one specific application, the article suggests using a Weibull distribution with $\alpha = 10$ and $\beta = 3.5$ (the time to repair, X, is measured in months).

The expected time to repair, variance, and standard deviation are

$$\mu = 3.5 \cdot \Gamma\left(1 + \frac{1}{10}\right) = 3.33 \text{ months}$$

$$\sigma^2 = (3.5)^2 \cdot \left\{\Gamma\left(1 + \frac{2}{10}\right) - \left[\Gamma\left(1 + \frac{1}{10}\right)\right]^2\right\} = 0.16 \Rightarrow \sigma = 0.4 \text{ months}$$

The probability that a shipping container requires repair within the first 3 months is

$$P(X \leq 3) = F(3; 10, 3.5) = 1 - e^{-(3/3.5)^{10}} = .193$$

Similarly, $P(2 \leq X \leq 4) = .974$, indicating that the distribution is almost entirely concentrated between 2 and 4 months.

The 95th percentile of this distribution—i.e., the value c which separates the longest-lasting 5% of shipping containers from the rest—is determined from

$$.95 = 1 - e^{-(c/3.5)^{10}}$$

Solving this equation gives $c \approx 3.906$ months. ■

Frequently, a Weibull model may be reasonable except that the smallest possible X value may be some value γ not assumed to be zero (this would also apply to a gamma model). The quantity γ can then be regarded as a third parameter of the distribution, which is what Weibull did in his original work. For, say, $\gamma = 3$, all curves in Figure 4.28 would be shifted 3 units to the right. This is equivalent to saying that $X - \gamma$ has the pdf (4.9), so that the cdf of X is obtained by replacing x in (4.10) by $x - \gamma$.

Example 4.33 An understanding of the volumetric properties of asphalt is important in designing mixtures that will result in high-durability pavement. The article "Is a Normal Distribution the Most Appropriate Statistical Distribution for Volumetric Properties in Asphalt Mixtures" (*J. Testing Eval.*, Sept. 2009: 1–11) used the analysis of some sample data to recommend that for a particular mixture, X = air void volume (%) be modeled with a three-parameter Weibull distribution. Suppose the values of the parameters are $\gamma = 4$, $\alpha = 1.3$, and $\beta = .8$ (quite close to estimates given in the article).

For $x \geq 4$, the cumulative distribution function is

$$F(x; \alpha, \beta, \gamma) = F(x; 1.3, .8, 4) = 1 - e^{-[(x-4)/.8]^{1.3}}$$

The probability that the air void volume of a specimen is between 5% and 6% is

$$P(5 \leq X \leq 6) = F(6; 1.3, .8, 4) - F(5; 1.3, .8, 4) = e^{-[(5-4)/.8]^{1.3}} - e^{-[(6-4)/.8]^{1.3}}$$
$$= .263 - .037 = .226$$

Figure 4.29 shows a graph of the corresponding Weibull density function, in which the shaded area corresponds to the probability just calculated.
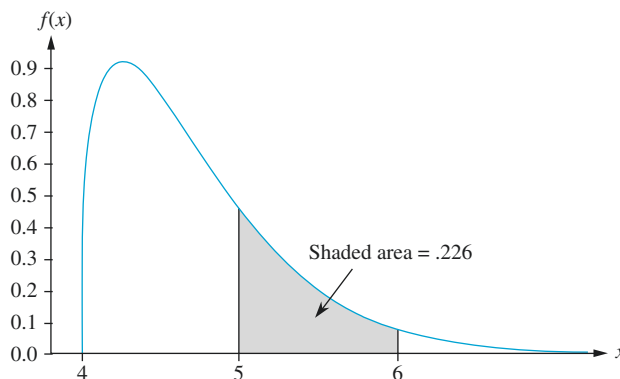
Figure 4.29 Weibull density curve with threshold = 4, shape = 1.3, scale = .8 ■

The Lognormal Distribution

Lognormal distributions have been used extensively in engineering, medicine, and more recently, finance.

DEFINITION A nonnegative rv X is said to have a **lognormal distribution** if the rv $Y = \ln(X)$ has a normal distribution. The resulting pdf of a lognormal rv when $\ln(X)$ is normally distributed with parameters μ and σ is

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-[\ln(x) - \mu]^2 / (2\sigma^2)} \quad x > 0$$

Be careful here: the parameters μ and σ are not the mean and standard deviation of X but of $\ln(X)$. The mean and variance of X can be shown to be

$$E(X) = e^{\mu + \sigma^2/2} \quad V(X) = e^{2\mu + \sigma^2} \cdot (e^{\sigma^2} - 1)$$

In Chapter 6, we will present a theoretical justification for this distribution in connection with the Central Limit Theorem, but as with other distributions, the lognormal can be used as a model even in the absence of such justification. Figure 4.30 illustrates graphs of the lognormal pdf; although a normal curve is symmetric, a lognormal curve has a positive skew.
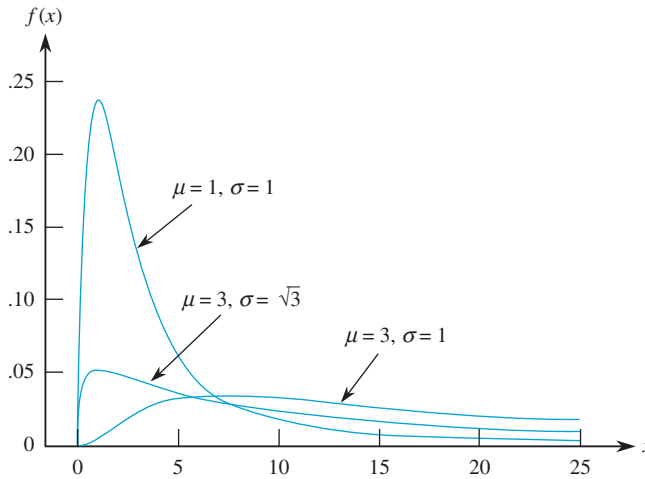
Figure 4.30 Lognormal density curves

Because $\ln(X)$ has a normal distribution, the cdf of X can be expressed in terms of the cdf $\Phi(z)$ of a standard normal rv Z. For $x > 0$,

$$
\begin{aligned}
F(x; \mu, \sigma) = P(X \le x) = P[\ln(X) \le \ln(x)] &= P\left[\frac{\ln(X) - \mu}{\sigma} \le \frac{\ln(x) - \mu}{\sigma}\right] \\
&= P\left[Z \le \frac{\ln(x) - \mu}{\sigma}\right] = \Phi\left[\frac{\ln(x) - \mu}{\sigma}\right]
\end{aligned}
\tag{4.11}
$$

Differentiating $F(x; \mu, \sigma)$ with respect to x gives the lognormal pdf $f(x; \mu, \sigma)$ above.

Example 4.34 According to the article "Predictive Model for Pitting Corrosion in Buried Oil and Gas Pipelines" (*Corrosion* 2009: 332–342), the lognormal distribution has been reported as the best option for describing the distribution of maximum pit depth data from cast iron pipes in soil. The authors suggest that a lognormal distribution with $\mu = .353$ and $\sigma = .754$ is appropriate for maximum pit depth (mm) of buried pipelines. For this distribution, the mean value and variance of pit depth are

$$E(X) = e^{.353 + (.754)^2/2} = e^{.6383} = 1.893$$

$$V(X) = e^{2(.353) + (.754)^2} \cdot (e^{(.754)^2} - 1) = (3.57697)(.765645) = 2.7387$$

The probability that maximum pit depth is between 1 and 2 mm is

$$\begin{aligned}
P(1 \leq X \leq 2) &= P(\ln(1) \leq \ln(X) \leq \ln(2)) \\
&= P(0 \leq \ln(X) \leq .693) \\
&= P\left(\frac{0 - .353}{.754} \leq Z \leq \frac{.693 - .353}{.754}\right) \\
&= \Phi(.45) - \Phi(-.47) = .354
\end{aligned}$$

Figure 4.31 illustrates this probability.
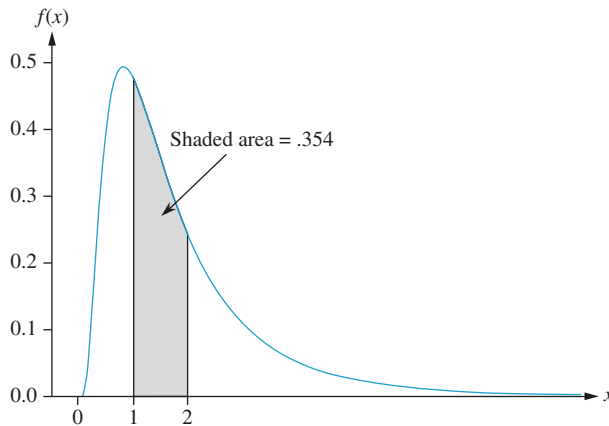
Figure 4.31 Lognormal density curve with $\mu = .353$ and $\sigma = .754$

What value c is such that only 1% of all specimens have a maximum pit depth exceeding c? The desired value satisfies

$$.99 = P(X \leq c) = P\left(Z \leq \frac{\ln(c) - .353}{.754}\right)$$

The z critical value 2.33 captures an upper-tail area of .01 ($z_{.01}$ = 2.33) and thus a cumulative area of .99. This implies that

$$\frac{\ln(c) - .353}{.754} = 2.33$$

from which $\ln(c) = 2.1098$ and $c = 8.247$. Thus 8.247 mm is the 99th percentile of the maximum pit depth distribution.   ∎

As with the Weibull distribution, a third parameter γ can be introduced so that the domain of the distribution is $x > \gamma$ rather than $x > 0$.

The Beta Distribution

All families of continuous distributions discussed so far except for the uniform distribution have positive density over an infinite interval (although typically the density function decreases rapidly to zero beyond a few standard deviations from the mean). The beta distribution provides positive density only for X in an interval of finite length.

DEFINITION   A random variable X is said to have a **beta distribution** with parameters α, β (both positive), A, and B if the pdf of X is

$$f(x; \alpha, \beta, A, B) = \frac{1}{B - A} \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \left(\frac{x - A}{B - A}\right)^{\alpha - 1} \left(\frac{B - x}{B - A}\right)^{\beta - 1} \quad A \le x \le B$$

The case $A = 0$, $B = 1$ gives the **standard beta distribution**.

Figure 4.32 illustrates several standard beta pdfs. Graphs of the general pdf are similar, except they are shifted and then stretched or compressed to fit over [A, B]. Unless α and β are integers, integration of the pdf to calculate probabilities is difficult, so either a table of the incomplete beta function or software is generally used.
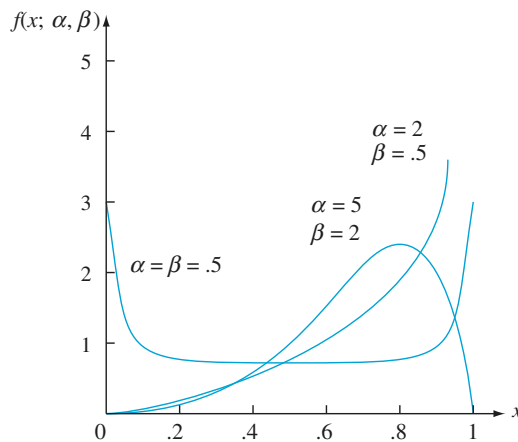
Figure 4.32 Standard beta density curves

The standard beta distribution is commonly used to model variation in the proportion or percentage of a quantity occurring in different samples, such as the proportion of a 24-h day that an individual is asleep or the proportion of a certain element in a chemical compound.

The mean and variance of X are

$$\mu = A + (B - A) \cdot \frac{\alpha}{\alpha + \beta} \qquad \sigma^2 = \frac{(B - A)^2 \alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

Example 4.35 Project managers often use a method labeled PERT—for program evaluation and review technique—to coordinate the various activities making up a large project. (One successful application was in the construction of the *Apollo* spacecraft.) A standard assumption in PERT analysis is that the time necessary to complete any particular activity once it has been started has a beta distribution with A = the optimistic time (if everything goes well) and B = the pessimistic time (if everything goes badly). Suppose that in constructing a single-family house, the time X (in days) necessary for laying the foundation has a beta distribution with $A = 2$, $B = 5$, $\alpha = 2$, and $\beta = 3$. Then $\alpha/(\alpha + \beta) = .4$, so $E(X) = 2 + (3)(.4) = 3.2$. For these values of α and β, the pdf of X is a simple polynomial function. The probability that it takes at most 3 days to lay the foundation is

$$P(X \le 3) = \int_2^3 \frac{1}{3} \cdot \frac{4!}{1! \cdot 2!} \left(\frac{x - 2}{3}\right) \left(\frac{5 - x}{3}\right)^2 dx$$

$$= \frac{4}{27} \int_2^3 (x - 2)(5 - x)^2 dx = \frac{4}{27} \cdot \frac{11}{4} = \frac{11}{27} = .407$$

∎

Many software packages can be used to perform probability calculations for the Weibull, lognormal, and beta distributions. Interested readers should consult the help menus in those packages.

Exercises: Section 4.5 (84–98)

84. The lifetime X (in hundreds of hours) of a type of vacuum tube has a Weibull distribution with parameters $\alpha = 2$ and $\beta = 3$. Compute the following:

    a. $E(X)$ and $V(X)$
    b. $P(X \le 6)$
    c. $P(1.5 \le X \le 6)$

    (This Weibull distribution is suggested as a model for time in service in "On the Assessment of Equipment Reliability: Trading Data Collection Costs for Precision," *J. Engr. Manuf.* 1991: 105–109).

85. Many U.S. railroad tracks were built using A7 steel, and there is renewed interest in the properties of this metal. The article "Stress-State, Temperature, and Strain Rate Dependence of Vintage ASTM A7 Steel" (*J. Engr. Mater. Tech.* 2019) describes, among other things, the distribution of manganese within A7 steel specimens. The authors found that the nearest-neighbor distance (NND, in microns) of manganese particles along longitudinal planes in A7 steel follow a Weibull distribution with (approximate) parameter values $\alpha = 1.18$ and $\beta = 21.61$.

    a. What is the probability of observing a NND between 20 and 40 μm? Less than 20 μm? More than 40 μm?
    b. What are the mean and standard deviation of this distribution?
    c. What is the median of this distribution?

86. In recent years the Weibull distribution has been used to model engine emissions of various pollutants. Let X denote the amount of NO_x emission (g/gal) from a randomly selected four-stroke engine of a certain type, and suppose that X has a Weibull distribution with $\alpha = 2$ and $\beta = 10$ (suggested by information in the article "Quantification of Variability and Uncertainty in Lawn and Garden Equipment NO_x and Total Hydrocarbon Emission Factors," *J. Air Waste Manag. Assoc.* 2002: 435–448).

   a. What is the cdf of X?

   b. Compute $P(X \leq 10)$ and $P(X \geq 10)$.

   c. Determine the mean and standard deviation of X.

   d. Determine the 75th percentile of this distribution.

87. Let X have a Weibull distribution with the pdf from Expression (4.10). Verify that $\mu = \beta\Gamma(1 + 1/\alpha)$. [*Hint*: In the integral for $E(X)$, make the change of variable $y = (x/\beta)^\alpha$, so that $x = \beta y^{1/\alpha}$.]

88. a. In Exercise 84, what is the median lifetime of such tubes? [*Hint*: Use Expression (4.10).]

   b. If X has a Weibull distribution with the cdf from Expression (4.10), obtain a general expression for the $(100p)$th percentile of the distribution.

   c. In Exercise 86, engines whose NO_x emissions exceed a threshold of t g/gal must be replaced to meet new environmental regulations. For what value of t would 10% of these engines require replacement?

89. Let X denote the ultimate tensile strength (ksi) at $-200°$ of a randomly selected steel specimen of a certain type that exhibits "cold brittleness" at low temperatures. Suppose that X has a Weibull distribution with $\alpha = 20$ and $\beta = 100$.

   a. What is the probability that X is at most 105 ksi?

   b. If specimen after specimen is selected, what is the long-run proportion having strength values between 100 and 105 ksi?

   c. What is the median of the strength distribution?

90. The authors of the article "Study on the Life Distribution of Microdrills" (*J. Engr. Manuf.* 2002: 301–305) suggested that a reasonable probability model for drill lifetime was a lognormal distribution with $\mu = 4.5$ and $\sigma = .8$.

   a. What are the mean value and standard deviation of lifetime?

   b. What is the probability that lifetime is at most 100?

   c. What is the probability that lifetime is at least 200? Greater than 200?

91. The article referenced in Exercise 85 also considered the distribution of areas (square microns) of single manganese particles in through thickness planes of A7 steel. The authors determined that a lognormal distribution with parameters $\mu = 1.513$ and $\sigma = 1.006$ to be an appropriate model for these manganese particle areas.

   a. Determine the mean and standard deviation of this distribution.

   b. What is the probability of observing a particle area less than 10 square microns? Between 10 and 20 μm^2?

   c. Determine the probability of observing a manganese particle area less than the mean value. Why does this probability not equal .5?

92. a. Use Equation (4.11) to write a formula for the median $\tilde{\mu}$ of the lognormal distribution. What is the median for the area distribution of the previous exercise?

   b. Recalling that z_α is our notation for the $100(1 - \alpha)$ percentile of the standard normal distribution, write an expression for the $100(1 - \alpha)$ percentile of the lognormal distribution. In the previous exercise, what value will particle area exceed only 5% of the time?

93. A theoretical justification based on a material failure mechanism underlies the assumption that ductile strength X of a material has a lognormal distribution. Suppose the parameters are $\mu = 5$ and $\sigma = .1$.

    a. Compute $E(X)$ and $V(X)$.
    b. Compute $P(X > 125)$.
    c. Compute $P(110 \leq X \leq 125)$.
    d. What is the value of median ductile strength?
    e. If ten different samples of an alloy steel of this type were subjected to a strength test, how many would you expect to have strength of at least 125?
    f. If the smallest 5% of strength values were unacceptable, what would the minimum acceptable strength be?

94. The article "The Statistics of Phytotoxic Air Pollutants" (*J. Roy. Statist Soc*. 1989: 183–198) suggests the lognormal distribution as a model for SO_2 concentration above a forest. Suppose the parameter values are $\mu = 1.9$ and $\sigma = .9$.

    a. What are the mean value and standard deviation of concentration?
    b. What is the probability that concentration is at most 10? Between 5 and 10?

95. What condition on α and β is necessary for the standard beta pdf to be symmetric?

96. Suppose the proportion X of surface area in a randomly selected quadrat that is covered by a certain plant has a standard beta distribution with $\alpha = 5$ and $\beta = 2$.

    a. Compute $E(X)$ and $V(X)$.
    b. Compute $P(X \leq .2)$.
    c. Compute $P(.2 \leq X \leq .4)$.
    d. What is the expected proportion of the sampling region not covered by the plant?

97. Let X have a standard beta density with parameters α and β.

    a. Verify the formula for $E(X)$ given in the section.
    b. Compute $E[(1 - X)^m]$. If X represents the proportion of a substance consisting of a particular ingredient, what is the expected proportion that does not consist of this ingredient?

98. Stress is applied to a 20-in. steel bar that is clamped in a fixed position at each end. Let Y = the distance from the left end at which the bar snaps. Suppose $Y/20$ has a standard beta distribution with $E(Y) = 10$ and $V(Y) = 100/7$.

    a. What are the parameters of the relevant standard beta distribution?
    b. Compute $P(8 \leq Y \leq 12)$.
    c. Compute the probability that the bar snaps more than 2 in. from where you expect it to snap.

4.6   Probability Plots

An investigator will often have obtained a numerical sample consisting of n observations and wish to know whether it is plausible that this sample came from a population distribution of some particular type (e.g., from a normal distribution). For one thing, many formal procedures from statistical inference are based on the assumption that the population distribution is of a specified type. The use of such a procedure is inappropriate if the actual underlying probability distribution differs greatly from the assumed type. Additionally, understanding the underlying distribution can sometimes give insight into the physical mechanisms involved in generating the data. An effective way to check a distributional assumption is to construct what is called a **probability plot**. The basis for our construction is a comparison between percentiles of the sample data and the corresponding percentiles of the assumed underlying distribution.

Sample Percentiles

The details involved in constructing probability plots differ a bit from source to source. Roughly speaking, sample percentiles are defined in the same way that percentiles of a population distribution are defined. The sample 50th percentile (i.e., the sample median) should separate the smallest 50% of the sample from the largest 50%, the sample 90th percentile should be such that 90% of the sample lies below that value and 10% lies above, and so on. Unfortunately, we run into problems when we actually try to compute the sample percentiles for a particular sample of n observations. If, for example, $n = 10$, then we can split off 20% or 30% of the data, but there is no value that will split off exactly 23% of these ten observations. To proceed further, we need an operational definition of sample percentiles (this is one place where different people and different software packages do slightly different things).

Statistical convention states that when n is odd, the sample median is the middle value in the ordered list of sample observations, for example, the sixth-largest value when $n = 11$. This amounts to regarding the middle observation as being half in the lower half of the data and half in the upper half. Similarly, suppose $n = 10$. Then if we call the third-smallest value the 25th percentile, we are regarding that value as being half in the lower group (consisting of the two smallest observations) and half in the upper group (the seven largest observations). This leads to the following general definition of sample percentiles.

DEFINITION Order the n sample observations from smallest to largest. Then the ith-smallest observation in the list is taken to be the **sample $[100(i - .5)/n]$th percentile**.

For example, if $n = 10$, the percentages corresponding to the ordered sample observations are $100(1 - .5)/10 = 5\%$, $100(2 - .5)/10 = 15\%$, 25%, …, and $100(10 - .5)/10 = 95\%$. That is, the smallest observation is the sample 5th percentile, the next-smallest value is the sample 15th percentile, and so on. All other percentiles could then be determined by interpolation; e.g., the sample 10th percentile would then be halfway between the 5th percentile (smallest sample observation) and the 15th percentile (second-smallest observation) of the $n = 10$ values. For the purposes of a probability plot, such interpolation will not be necessary, because a probability plot will be based only on the percentages $100(i - .5)/n$ corresponding to the n sample observations.

A Probability Plot

We now wish to determine whether our sample data could plausibly have come from some particular population distribution (e.g., a normal distribution with $\mu = 10$ and $\sigma = 3$). If the sample was actually selected from the specified distribution, the sample percentiles (ordered sample observations) should be reasonably close to the corresponding population distribution percentiles. That is, for $i = 1, 2, …, n$ there should be reasonable agreement between the ith-smallest sample observation and the theoretical $[100(i - .5)/n]$th percentile for the specified distribution. Consider the (sample percentile, population percentile) pairs—that is, the pairs

$$\begin{pmatrix} i\text{th smallest sample} & [100(i - .5)/n]\text{th percentile} \\ \text{observation} & , & \text{of the population distribution} \end{pmatrix}$$

for $i = 1, …, n$. Each such pair can be plotted as a point on a two-dimensional coordinate system. If the sample percentiles are close to the corresponding population distribution percentiles, the first number in each pair will be roughly equal to the second number, and the plotted points will then fall

close to a 45° line passing through (0, 0). Substantial deviations of the plotted points from this 45° line suggest that the assumed distribution might be wrong.

Example 4.36 The value of a physical constant is known to an experimenter. The experimenter makes $n = 10$ independent measurements of this value using a measurement device and records the resulting measurement errors (error = observed value − true value). These observations appear in the accompanying table.

Percentage	5	15	25	35	45
Sample observation	−1.91	−1.25	−.75	−.53	.20
z percentile	−1.645	−1.037	−.675	−.385	−.126

Percentage	55	65	75	85	95
Sample observation	.35	.72	.87	1.40	1.56
z percentile	.126	.385	.675	1.037	1.645

Is it plausible that the random variable *measurement error* has a standard normal distribution? The needed standard normal (z) percentiles are also displayed in the table and were determined as follows: the 5th percentile of the distribution under consideration, $N(0,1)$, is such that $\Phi(z) = .05$. From software or Appendix Table A.3, the solution is roughly $z = −1.645$. The other nine population (z) percentiles were found in a similar fashion.

Thus the points in the probability plot are (−1.91, −1.645), (−1.25, −1.037), …, and (1.56, 1.645). Figure 4.33 shows the resulting plot. Although the points deviate a bit from the 45° line, the predominant impression is that this line fits the points reasonably well. The plot suggests that the standard normal distribution is a realistic probability model for measurement error.
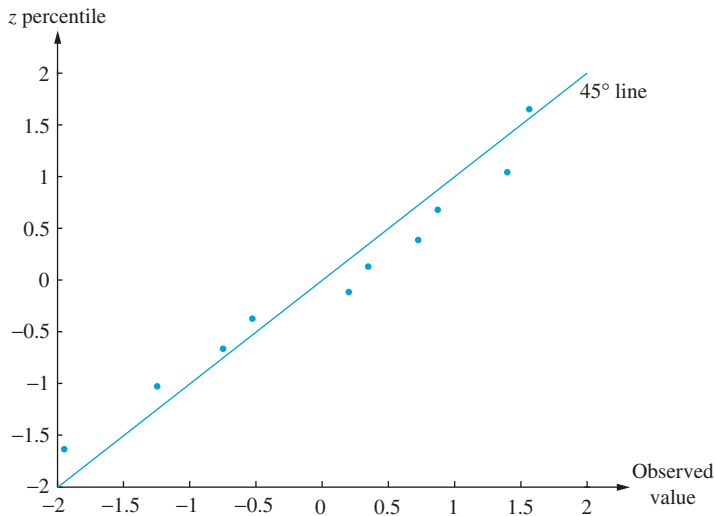
Figure 4.33 Plots of pairs (observed value, z percentile) for the data of Example 4.36 ∎

An investigator is typically not interested in knowing whether a particular probability distribution, such as the normal distribution with $\mu = 0$ and $\sigma = 1$ or the exponential distribution with $\lambda = .1$, is a plausible model for the population distribution from which the sample was selected. Instead, the investigator will want to know whether *some* member of a family of probability distributions provides a plausible model—the family of normal distributions, the family of exponential distributions, the family of Weibull distributions, and so on. The values of any parameters are usually not specified at the outset. If the family of Weibull distributions is under consideration as a model for lifetime data, the issue is whether there are *any* values of the parameters α and β for which the corresponding Weibull distribution gives a good fit to the data. Fortunately, it is almost always the case that just one probability plot will suffice for assessing the plausibility of an entire family. If the plot deviates substantially from a straight line, but not necessarily the 45° line, no member of the family is plausible.

To see why, let's focus on a plot for checking normality. As mentioned earlier, such a plot can be very useful in applied work because many formal statistical procedures are appropriate (i.e., give accurate inferences) only when the population distribution is at least approximately normal. These procedures should generally not be used if a normal probability plot shows a very pronounced departure from linearity. The key to constructing an omnibus normal probability plot is the relationship between standard normal (z) percentiles and those for any other normal distribution, which was presented in Section 4.3:

$$\begin{array}{c} \text{percentile for a} \\ N(\mu, \sigma)\text{distribution} \end{array} = \mu + \sigma \cdot (\text{corresponding } z \text{ percentile})$$

If each sample observation was exactly equal to the corresponding $N(\mu, \sigma)$ percentile, then the pairs (observation, $\mu + \sigma \cdot [z$ percentile]) would fall on the 45° line, $y = x$. But since $\mu + \sigma z$ is itself a linear function, the pairs (observation, z percentile) would also fall on a straight line, just not the line with slope 1 and y-intercept 0. (The latter pairs would pass through the line $z = x/\sigma - \mu/\sigma$, but the equation itself isn't important.)

DEFINITION   A plot of the n pairs

$$(i\text{th-smallest observation}, [100(i - .5)/n]\text{th } z \text{ percentile})$$

on a two-dimensional coordinate system is called a **normal probability plot**. If the sample observations are in fact drawn from a normal distribution, then the points should fall close to a straight line (although not necessarily a 45° line). Thus a plot for which the points fall close to *some* straight line suggests that the assumption of a normal population distribution is plausible.

Example 4.37 The accompanying sample consisting of $n = 20$ observations on dielectric breakdown voltage of a piece of epoxy resin appeared in the article "Maximum Likelihood Estimation in the 3-Parameter Weibull Distribution" (*IEEE Trans. Dielectrics Electr. Insul.* 1996: 43–55). Values of $(i - .5)/n$ for which z percentiles are needed are $(1 - .5)/20 = .025$, $(2 - 5)/20 = .075$, …, and .975.

Observation	24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
z percentile	−1.96	−1.44	−1.15	−.93	−.76	−.60	−.45	−.32	−.19	−.06

Observation	27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88
z percentile	.06	.19	.32	.45	.60	.76	.93	1.15	1.44	1.96

Figure 4.34 shows the resulting normal probability plot. The pattern in the plot is quite straight, indicating it is plausible that the population distribution of dielectric breakdown voltage is normal.
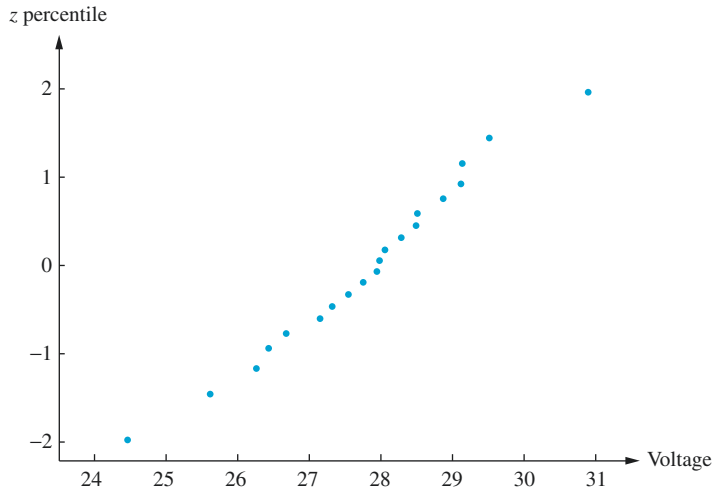
Figure 4.34   Normal probability plot for the dielectric breakdown voltage sample   ■

There is an alternative version of a normal probability plot in which the z percentile axis is replaced by a nonlinear probability axis. The scaling on this axis is constructed so that plotted points should again fall close to a line when the sampled distribution is normal. Figure 4.35 shows such a plot from Minitab for the breakdown voltage data of Example 4.37. Here the z values are replaced by the corresponding normal percentiles. The plot remains the same, and it is just the labeling of the axis that changes. Minitab and various other software packages use the refinement $(i - .375)/(n + .25)$ of the expression $(i - .5)/n$ in order to get a better approximation to what is expected for the ordered values from the standard normal distribution.
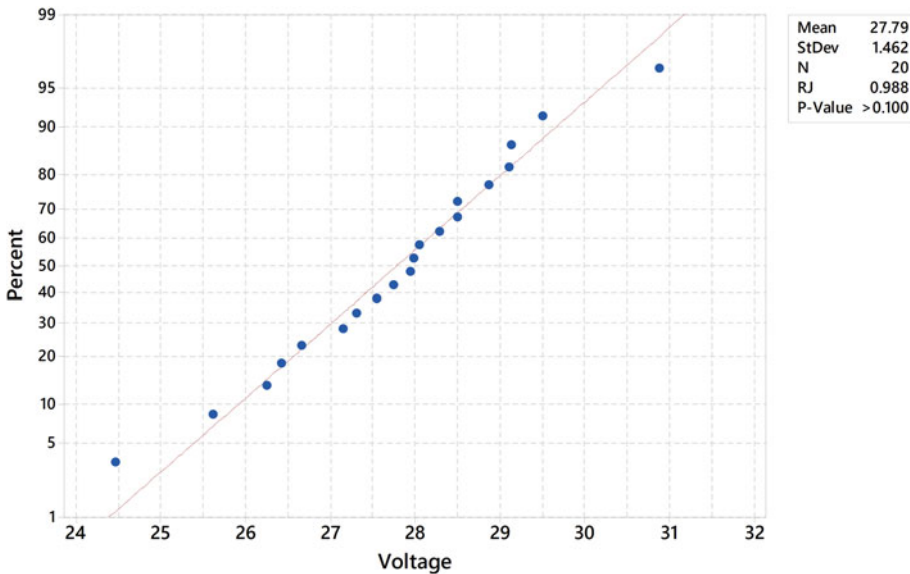
Figure 4.35   Normal probability plot of the breakdown voltage data from Minitab

Departures from Normality

A nonnormal population distribution can often be placed in one of the following three categories:

1. It is symmetric and has "lighter tails" than does a normal distribution; that is, the density curve declines more rapidly out in the tails than does a normal curve.
2. It is symmetric and heavy-tailed compared to a normal distribution.
3. It is skewed.

A uniform distribution is light-tailed, since its density function drops to zero outside a finite interval. The density function $f(x) = 1/[\pi(1 + x^2)]$, for $-\infty < x < \infty$, is one example of a heavy-tailed distribution, since $1/(1 + x^2)$ declines much less rapidly than does $e^{-x^2/2}$. Lognormal and Weibull distributions are among those that are skewed. When the points in a normal probability plot do not adhere to a straight line, the pattern will frequently suggest that the population distribution is in a particular one of these three categories.

Figure 4.36 illustrates typical normal probability plots corresponding to the three situations above. If the sample was selected from a light-tailed distribution, the largest and smallest observations are usually not as extreme as would be expected from a normal random sample. Visualize a straight
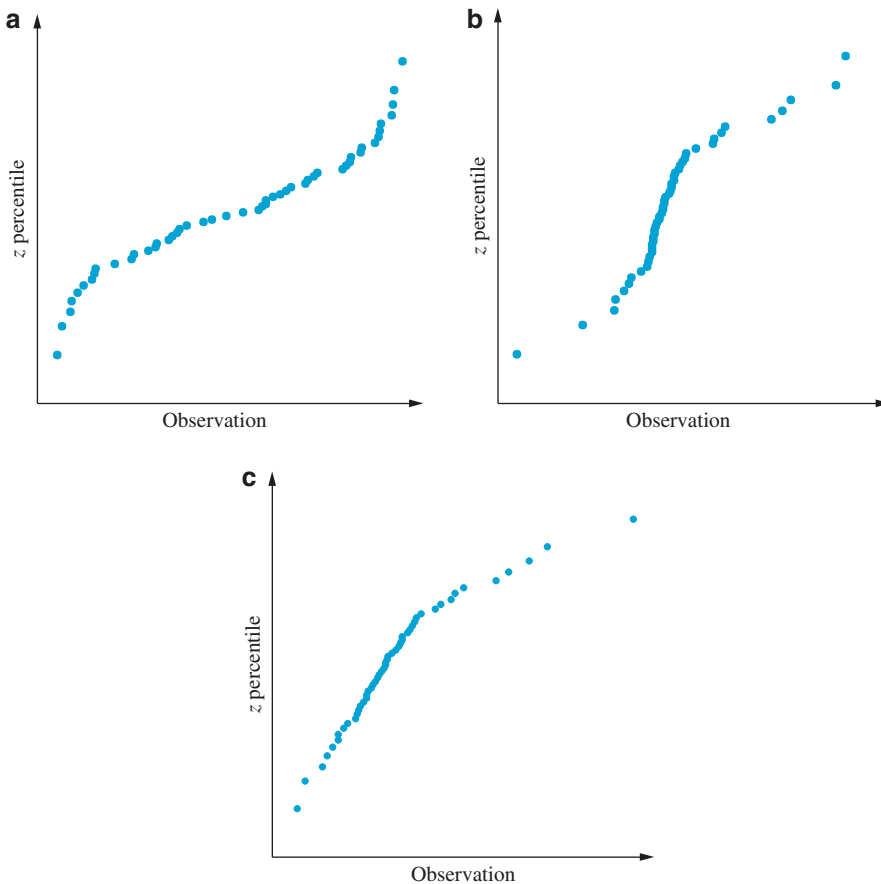
Figure 4.36 Probability plots that suggest a nonnormal distribution:
(**a**) a plot consistent with a light-tailed distribution; (**b**) a plot consistent with
a heavy-tailed distribution; (**b**) a plot consistent with a (positively) skewed distribution

line drawn through the middle part of the plot; points on the far right tend to be above the line (z percentile > observed value), whereas points on the left end of the plot tend to fall below the straight line (z percentile < observed value). The result is an S-shaped pattern of the type pictured in Figure 4.36a. For sample observations from a heavy-tailed distribution, the opposite effect will occur, and a normal probability plot will have an S shape with the opposite orientation, as in Figure 4.36b. If the underlying distribution is positively skewed (a short left tail and a long right tail), the smallest sample observations will be larger than expected from a normal sample and so will the largest observations. In this case, points on both ends of the plot will fall below a straight line through the middle part, yielding a curved pattern, as illustrated in Figure 4.36c. For example, a sample from a lognormal distribution will usually produce such a pattern; a plot of (ln(observation), z percentile) pairs should then resemble a straight line.

Even when the population distribution is normal, the sample percentiles will not coincide exactly with the theoretical percentiles because of sampling variability. How much can the points in the probability plot deviate from a straight line pattern before the assumption of population normality is no longer plausible? This is not an easy question to answer. Generally speaking, a small sample from a normal distribution is more likely to yield a plot with a nonlinear pattern than is a large sample. The book *Fitting Equations to Data* (see the bibliography) presents the results of a simulation study in which numerous samples of different sizes were selected from normal distributions. The authors concluded that there is typically greater variation in the appearance of the probability plot for sample sizes smaller than 30, and only for much larger sample sizes does a linear pattern generally predominate. When a plot is based on a small sample size, only a very substantial departure from linearity should be taken as conclusive evidence of nonnormality. A similar comment applies to probability plots for checking the plausibility of other types of distributions.

Beyond Normality

Consider a family of probability distributions involving two parameters, θ_1 and θ_2, and let $F(x; \theta_1, \theta_2)$ denote the corresponding cdf. The family of normal distributions is one such family, with $\theta_1 = \mu$, $\theta_2 = \sigma$, and $F(x; \mu, \sigma) = \Phi[(x - \mu)/\sigma]$. Another example is the Weibull family, with $\theta_1 = \alpha$, $\theta_2 = \beta$, and

$$F(x; \alpha, \beta) = 1 - e^{-(x/\beta)^{\alpha}}$$

Still another family of this type is the gamma family, for which the cdf is an integral involving the incomplete gamma function that cannot be expressed in any simpler form.

The parameters θ_1 and θ_2 are said to be **location** and **scale parameters**, respectively, if $F(x; \theta_1, \theta_2)$ is a function of $(x - \theta_1)/\theta_2$. The parameters μ and σ of the normal family are location and scale parameters, respectively. Changing μ shifts the location of the bell-shaped density curve to the right or left, and changing σ amounts to stretching or compressing the measurement scale (the scale on the horizontal axis when the density function is graphed). Another example is given by the cdf

$$F(x; \theta_1, \theta_2) = 1 - e^{-e^{(x-\theta_1)/\theta_2}} \qquad -\infty < x < \infty$$

A random variable with this cdf is said to have an *extreme value distribution*. It is used in applications involving component lifetime and material strength.

Although the form of the extreme value cdf might at first glance suggest that θ_1 is the point of symmetry for the density function, and therefore the mean and median, this is not the case. Instead, $P(X \leq \theta_1) = F(\theta_1; \theta_1, \theta_2) = 1 - e^{-1} = .632$, and the density function $f(x; \theta_1, \theta_2) = F'(x; \theta_1, \theta_2)$ is negatively skewed (a long lower tail). Similarly, the scale parameter θ_2 is not the standard deviation ($\mu = \theta_1 - 5772\theta_2$ and $\sigma = 1.283\theta_2$). However, changing the value of θ_1 does change the location of the density curve, whereas a change in θ_2 rescales the measurement axis.

The parameter β of the Weibull distribution is a scale parameter. However, α is not a location parameter but instead is called a **shape parameter**. The same is true for the parameters α and β of the gamma distribution. In the usual form, the density function for any member of either the gamma or Weibull distribution is positive for $x > 0$ and zero otherwise. A location (or shift) parameter can be introduced as a third parameter γ (we noted this for the Weibull distribution in Section 4.5) to shift the density function so that it is positive if $x > \gamma$ and zero otherwise.

When the family under consideration has only location and scale parameters, the issue of whether any family member is a plausible population distribution can be addressed by a single probability plot. This is exactly what we did to obtain an omnibus normal probability plot. One first obtains the percentiles of the *standard distribution*, the one with $\theta_1 = 0$ and $\theta_2 = 1$, for percentages $100(i - .5)/n$ ($i = 1, \ldots, n$). The n (observation, standardized percentile) pairs give the points in the plot.

Somewhat surprisingly, this methodology can be applied to yield an omnibus Weibull probability plot. The key result is that if X has a Weibull distribution with shape parameter α and scale parameter β, then the transformed variable $\ln(X)$ has an extreme value distribution with location parameter $\theta_1 = \ln(\beta)$ and scale parameter $\theta_2 = 1/\alpha$ (see Exercise 154). Thus a plot of the (ln(observation), extreme value standardized percentile) pairs that shows a strong linear pattern provides support for choosing the Weibull distribution as a population model.

Example 4.38 As climate change continues, more areas experience extreme wind events, which both safety engineers and FEMA must accurately model because they affect home damage. Engineers frequently use the Weibull distribution to model maximum wind speed in a given region. The article "Estimation of Extreme Wind Speeds by Using Mixed Distributions" (*Engr. Invest. Technol.* 2013: 153–162) provides measurements of $X =$ maximum wind speed (m/s) for 45 stations in the Netherlands. A Weibull probability plot can be constructed by plotting the logarithms of those observations against the ($100p$)th percentiles of the extreme value distribution for $p = (1 - .5)/45$, $(2 - .5)/45, \ldots, (45 - .5)/45$. The ($100p$)th percentile $\eta(p)$ satisfies

$$p = F(\eta(p)) = 1 - e^{-e^{\eta(p)}}$$

from which $\eta(p) = \ln[-\ln(1 - p)]$.

Percentile	x	$\ln(x)$	Percentile	x	$\ln(x)$
–4.49	17.7	2.87	–0.30	25.8	3.25
–3.38	18.9	2.94	–0.24	25.8	3.25
–2.86	20.9	3.04	–0.18	25.9	3.25
–2.51	21.4	3.06	–0.12	25.9	3.25
–2.25	21.7	3.08	–0.06	26.0	3.26
–2.04	22.3	3.10	0.00	26.2	3.27
–1.86	22.6	3.12	0.06	26.2	3.27
–1.70	22.8	3.13	0.12	26.4	3.27
–1.56	23.0	3.14	0.19	26.6	3.28
–1.44	23.1	3.14	0.25	26.7	3.28

(continued)

Percentile	x	ln(x)	Percentile	x	ln(x)
−1.33	23.2	3.14	0.31	26.8	3.29
−1.22	23.3	3.15	0.38	26.9	3.29
−1.12	23.7	3.17	0.44	26.9	3.29
−1.03	24.0	3.18	0.51	27.0	3.30
−0.94	24.1	3.18	0.58	27.3	3.31
−0.86	24.1	3.18	0.66	28.0	3.33
−0.78	24.2	3.19	0.74	28.1	3.34
−0.71	24.4	3.19	0.83	28.8	3.36
−0.64	25.2	3.23	0.94	29.2	3.37
−0.57	25.6	3.24	1.06	29.4	3.38
−0.50	25.6	3.24	1.22	30.0	3.40
−0.43	25.7	3.25	1.50	31.1	3.44
−0.37	25.7	3.25			

The pairs (2.87, −4.49), (2.94, −3.38), …, (3.44, 1.50) are plotted as points in Figure 4.37. The straightness of the plot argues strongly that $\ln(X)$ is compatible with an extreme value distribution, and so X itself can be well-modeled by a Weibull distribution.
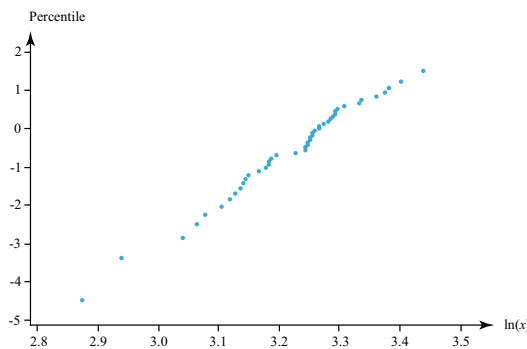
Figure 4.37   A Weibull probability plot of the maximum wind speed data

It should be noted that many statistical software packages have built-in Weibull probability plot functionality that does not require the user to transform the data or calculate the extreme value percentiles. ∎

The gamma distribution is an example of a family involving a shape parameter for which there is no transformation into a distribution that depends only on location and scale parameters. Construction of a probability plot necessitates first estimating the shape parameter from sample data (some methods for doing this are described in Chapter 7).

Sometimes an investigator wishes to know whether the transformed variable X^{θ} has a normal distribution for some value of θ (by convention, $\theta = 0$ is identified with the logarithmic transformation, in which case X has a lognormal distribution). The book *Graphical Methods for Data Analysis* (see the bibliography) discusses this type of problem as well as other refinements of probability plotting.

Formal Tests of a Distributional Fit

Given the limitations of probability plots, there is need for an alternative. Statisticians have developed several formal procedures for assessing whether sample data could plausibly have come from a normally distributed population. The *Ryan-Joiner test* quantifies on a zero-to-one scale how closely the pattern of points in a normal probability plot adheres to a straight line, with higher values corresponding to a more linear pattern. If this quantified value is too low, the test casts doubt on population normality. (In the formal language of Chapter 9, the test "rejects" the claim of a normal population if the probability plot is sufficiently nonlinear.) The Ryan–Joiner measure appears in the top-right corner of Figure 4.35 (RJ = 0.988); its very high value on a [0, 1] scale implies that population normality is plausible. The *Shapiro–Wilk test* proceeds similarly, although it quantifies linearity somewhat differently, and is more ubiquitous among statistical software packages: R, SAS, Stata, SPSS, and JMP all include the Shapiro–Wilk test among their options.

The Ryan–Joiner and Shapiro–Wilk tests are specialized to assessing normality; i.e., they are not designed to detect conformance with other distributions (gamma, Weibull, etc.). The *Anderson–Darling (AD) test* and the *Kolmogorov–Smirnov (KS) test* can both be applied to a wider collection of distributions. Each of these latter tests is based on comparing the cdf $F(x)$ of the theorized distribution (e.g., the Weibull cdf) to the "empirical" cdf $F_n(x)$ of the sample data, defined for any real number x by

$$F_n(x) = \text{the proportion of the sample values} \{x_1, \ldots, x_n\} \text{ that are} \leq x$$

If $F(x)$ and $F_n(x)$ are "too far apart" in some sense, this indicates that the sample data is incompatible with the theorized population distribution (and so that theory should be "rejected"). The AD and KS tests differ in how they quantify the disparity between $F(x)$ and $F_n(x)$. (Specific to assessing normality, a 2011 article in the *Journal of Statistical Modeling and Analysis* found that the Shapiro–Wilk test has greater capability of detecting normality violations than either the AD or KS tests.)

Exercises: Section 4.6 (99–109)

99. The accompanying normal probability plot was constructed from a sample of 30 readings on tension for mesh screens behind the surface of video display tubes. Does it appear plausible that the tension distribution is normal? Explain.
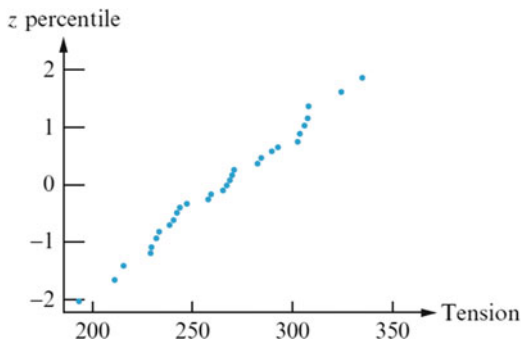
100. A sample of 15 female collegiate golfers was selected and the clubhead velocity (km/h) while swinging a driver was determined for each one, resulting in the following data ("Hip Rotational Velocities during the Full Golf Swing," *J. Sports Sci. Med.* 2009: 296–299):

69.0	69.7	72.7	80.3	81.0
85.0	86.0	86.3	86.7	87.7
89.3	90.7	91.0	92.5	93.0

The corresponding z percentiles are

−1.83	−1.28	−0.97	−0.73	−0.52
−0.34	−0.17	0.0	0.17	0.34
0.52	0.73	0.97	1.28	1.83

Construct a normal probability plot and a dotplot. Is it plausible that the population distribution is normal?

101. Construct a normal probability plot for the following sample of observations on coating thickness for low-viscosity paint ("Achieving a Target Value for a Manufacturing Process: A Case Study," *J. Qual. Tech*. 1992: 22–26). Would you feel comfortable estimating population mean thickness using a method that assumed a normal population distribution? Explain.

.83	.88	.88	1.04	1.09	1.12
1.29	1.31	1.48	1.49	1.59	1.62
1.65	1.71	1.76	1.83		

102. The article "A Probabilistic Model of Fracture in Concrete and Size Effects on Fracture Toughness" (*Mag. Concrete Res*. 1996: 311–320) gives arguments for why fracture toughness in concrete specimens should have a Weibull distribution and presents several histograms of data that appear well fit by superimposed Weibull curves. Consider the following sample of $n = 18$ observations on toughness for high-strength concrete (consistent with one of the histograms); values of $p_i = (i - 5)/18$ are also given.

Observation	.47	.58	.65	.69	.72	.74
p_i	.0278	.0833	.1389	.1944	.2500	.3056
Observation	.77	.79	.80	.81	.82	.84
p_i	.3611	.4167	.4722	.5278	.5833	.6389
Observation	.86	.89	.91	.95	1.01	1.04
p_i	.6944	.7500	.8056	.8611	.9167	.9722

Construct a Weibull probability plot and comment.

103. Construct a normal probability plot for the escape time data given in Exercise 46 of Chapter 1. Does it appear plausible that escape time has a normal distribution? Explain.

104. The article "Reducing Uncertainty of Design Floods of Two-Component Mixture Distributions by Utilizing Flood Timescale to Classify Flood Types in Seasonally Snow Covered Region" (*J. Hydrol*. 2019:

588–608) reports the accompanying data on annual precipitation (mm/yr) at 34 watersheds in Norway.

527.9	598.2	668.5	1136.6	1160.1
1177.0	1512.7	1542.5	1642.6	2383.8
2628.5	2671.5	697.7	859.0	884.3
1182.3	1195.6	1212.8	1872.1	1976.3
2082.9	2872.3	3221.6	3430.2	894.3
1030.7	1035.5	1294.2	1441.7	1475.4
2266.3	2337.0	2365.0	4029.7	

a. Construct a normal probability plot. Is normality plausible?
b. Construct a Weibull probability plot. Is the Weibull distribution family plausible?

105. Construct a probability plot that will allow you to assess the plausibility of the lognormal distribution as a model for the nitrogen data of Example 1.17.

106. The accompanying observations are precipitation values during March over a 30-year period in Minneapolis–St. Paul.

0.77	1.20	3.00	1.62	2.81	2.48
1.74	0.47	3.09	1.31	1.87	0.96
0.81	1.43	1.51	0.32	1.18	1.89
1.20	3.37	2.10	0.59	1.35	0.90
1.95	2.20	0.52	0.81	4.75	2.05

a. Construct and interpret a normal probability plot for this data set.
b. Calculate the square root of each value and then construct a normal probability plot based on this transformed data. Does it seem plausible that the square root of precipitation is normally distributed?
c. Repeat part (b) after transforming by cube roots.

107. The accompanying data set consists of observations on shower-flow rate (L/min) for a sample of $n = 129$ houses in Perth, Australia ("An Application of Bayes Methodology to the Analysis of Diary Records in a Water Use Study," *J. Amer. Statist. Assoc*. 1987: 705–711):

4.6	12.3	7.1	7.0	4.0	9.2	6.7	6.9	11.5	5.1
11.2	10.5	14.3	8.0	8.8	6.4	5.1	5.6	9.6	7.5
7.5	6.2	5.8	2.3	3.4	10.4	9.8	6.6	3.7	6.4
8.3	6.5	7.6	9.3	9.2	7.3	5.0	6.3	13.8	6.2
5.4	4.8	7.5	6.0	6.9	10.8	7.5	6.6	5.0	3.3
7.6	3.9	11.9	2.2	15.0	7.2	6.1	15.3	18.9	7.2
5.4	5.5	4.3	9.0	12.7	11.3	7.4	5.0	3.5	8.2
8.4	7.3	10.3	11.9	6.0	5.6	9.5	9.3	10.4	9.7
5.1	6.7	10.2	6.2	8.4	7.0	4.8	5.6	10.5	14.6
10.8	15.5	7.5	6.4	3.4	5.5	6.6	5.9	15.0	9.6
7.8	7.0	6.9	4.1	3.6	11.9	3.7	5.7	6.8	11.3
9.3	9.6	10.4	9.3	6.9	9.8	9.1	10.6	4.5	6.2
8.3	3.2	4.9	5.0	6.0	8.2	6.3	3.8	6.0	

Construct a normal probability plot of this data and comment.

108. Let the *ordered* sample observations be denoted by $y_1, y_2, \ldots, y_n$ (y_1 being the smallest and y_n the largest). Our suggested check for normality is to plot the $(\Phi^{-1}[(i - .5)/n], y_i)$ pairs. Suppose we believe that the observations come from a distribution with mean 0, and let $w_1, \ldots, w_n$ be the *ordered absolute* values of the x_i's. A **half-normal plot** is a probability plot of the w_i's. That is, since $P(|Z| \leq w) = P(-w \leq Z \leq w) = 2\Phi(w) - 1$, a half-normal plot is a plot of the $(\Phi^{-1}[(p_i + 1)/2], w_i)$ pairs, where $p_i = (i - 5)/n$. The virtue of this plot is that small or large outliers in the original sample will now appear only at the upper end of the plot rather than at both ends. Construct a half-normal plot for the following sample of measurement errors, and comment: $-3.78, -1.27, 1.44, -.39, 12.38, -43.40, 1.15, -3.96, -2.34, 30.84$.

109. The following failure time observations (1000s of hours) resulted from accelerated life testing of 16 integrated circuit chips of a certain type:

82.8	11.6	359.5	502.5	307.8	179.7
242.0	26.5	244.8	304.3	379.1	212.6
229.9	558.9	366.7	204.6		

Use the corresponding percentiles of the exponential distribution with $\lambda = 1$ to construct a probability plot. Then explain why the plot assesses the plausibility of the sample having been generated from *any* exponential distribution.

4.7   Transformations of a Random Variable

Often we need to deal with a transformation $Y = g(X)$ of the random variable X. For example, $g(X)$ could be a simple change of time scale: if X is the time to complete a task in minutes, then $Y = 60X$ is the completion time expressed in seconds. How can we get the pdf of Y from the pdf of X? Consider first a simple example.

Example 4.39 The interval X in minutes between calls to a 911 center is exponentially distributed with mean 2 min, so its pdf $f_X(x) = .5e^{-.5x}$ for $x > 0$. In order to get the pdf of $Y = 60X$, we first obtain its cdf:

$$F_Y(y) = P(Y \leq y) = P(60X \leq y) = P(X \leq y/60) = F_X(y/60)$$

$$= \int_0^{y/60} .5e^{-.5x}dx = 1 - e^{-y/120}$$

Differentiating this with respect to y gives $f_Y(y) = (1/120)e^{-y/120}$ for $y > 0$. We see that the distribution of Y is exponential with mean 120 s (2 min).

There is nothing special here about the mean 2 and the multiplier 60. It should be clear that if we multiply an exponential random variable with mean μ by a positive constant c we get another exponential random variable with mean $c\mu$. ∎

Sometimes it isn't possible to evaluate the cdf in closed form. Could the pdf of Y be obtained without evaluating the integral? Yes, thanks to the following theorem.

TRANSFORMATION THEOREM

Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is monotonic (either strictly increasing or strictly decreasing) on the set of all possible values of X, so it has an inverse function $X = g^{-1}(Y) = h(Y)$. Assume that h has a derivative $h'(y)$. Then

$$f_Y(y) = f_X(h(y)) \cdot |h'(y)| \tag{4.12}$$

Proof Here is the proof assuming that g is monotonically increasing. The proof for g monotonically decreasing is similar. First find the cdf of Y:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq h(y)) = F_X(h(y))$$

The third equality above, wherein $g(X) \leq y$ is true iff $X \leq g^{-1}(y) = h(y)$, relies on g being a monotonically increasing function. Now differentiate the cdf with respect to y, using the Chain Rule:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(h(y)) = F_X'(h(y)) \cdot h'(y) = f_X(h(y)) \cdot h'(y)$$

The absolute value on the derivative in (4.12) is needed only in the other case where g is decreasing. The set of possible values for Y is obtained by applying g to the set of possible values for X. ∎

Example 4.40 Let's apply the Transformation Theorem to the situation introduced in Example 4.39. There $Y = g(X) = 60X$ and $X = h(Y) = Y/60$.

$$f_Y(y) = f_X[h(y)]|h'(y)| = .5e^{-.5x}\left|\frac{1}{60}\right| = \frac{1}{120}e^{-y/120} \quad y > 0$$

This matches the pdf of Y derived through the cdf in Example 4.39. ∎

Example 4.41 Let $X \sim \text{Unif}[0, 1]$, so $f_X(x) = 1$ for $0 \leq x \leq 1$, and define a new variable $Y = 2\sqrt{X}$. The function $g(x) = 2\sqrt{x}$ is monotone on $[0, 1]$, with inverse $x = h(y) = y^2/4$. Apply the Transformation Theorem:

$$f_Y(y) = f_X(h(y))|h'(y)| = (1)\left|\frac{2y}{4}\right| = \frac{y}{2} \quad 0 \leq y \leq 2$$

The range $0 \leq y \leq 2$ comes from the fact that $y = 2\sqrt{x}$ maps $[0, 1]$ to $[0, 2]$. A graphical representation may help in understanding why the transform $Y = 2\sqrt{X}$ yields $f_Y(y) = y/2$ if $X \sim \text{Unif}[0, 1]$. Figure 4.38a shows the uniform distribution with $[0, 1]$ partitioned into ten subintervals. In Figure 4.38b the endpoints of these intervals are shown after transforming according to $y = 2\sqrt{x}$. The heights of the rectangles are arranged so each rectangle still has area .1, and therefore the probability in each interval is preserved. Notice the close fit of the dashed line, which has the equation $f_Y(y) = y/2$.
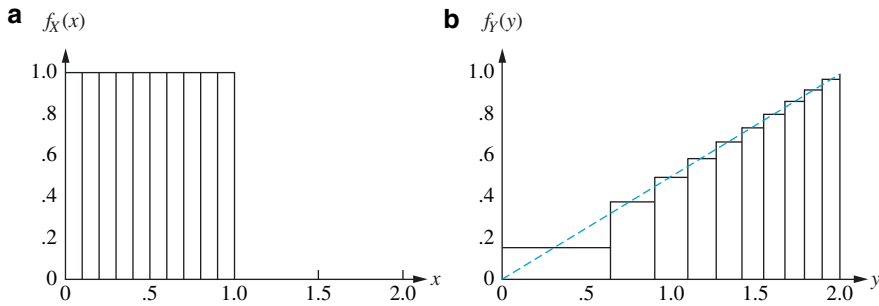
Figure 4.38 The effect on the pdf if X is uniform on [0, 1] and $Y = 2\sqrt{X}$ ∎

Example 4.42 The variation in a certain electrical current source X (in milliamps) can be modeled by the pdf

$$f_X(x) = 1.25 - .25x \quad 2 \le x \le 4$$

If this current passes through a 220-Ω resistor, the resulting power Y (in microwatts) is given by the expression $Y = 220X^2$. The function $y = g(x) = 220x^2$ is monotonically increasing on the range of X, the interval [2, 4], and has inverse function $x = h(y) = g^{-1}(y) = \sqrt{y/220}$. (Notice that $g(x)$ is a parabola and thus not monotone on the entire real number line, but for the purposes of the Transformation Theorem $g(x)$ only needs to be monotone on the range of the rv X.) Apply (4.12):

$$
\begin{aligned}
f_Y(y) &= f_X(h(y)) \cdot |h'(y)| \\
&= f_X(\sqrt{y/220}) \cdot \left| \frac{d}{dy} \sqrt{y/220} \right| \\
&= (1.25 - .25\sqrt{y/220}) \cdot \frac{1}{2\sqrt{220y}} = \frac{5}{8\sqrt{220y}} - \frac{1}{1760}
\end{aligned}
$$

The set of possible Y values is determined by substituting $x = 2$ and $x = 4$ into $g(x) = 220x^2$; the resulting range for Y is [880, 3520]. Therefore, the pdf of $Y = 220X^2$ is

$$
f_Y(y) = \begin{cases} \dfrac{5}{8\sqrt{220y}} - \dfrac{1}{1760} & 880 \le y \le 3520 \\[2mm] 0 & \text{otherwise} \end{cases}
$$

The pdfs of X and Y appear in Figure 4.39.
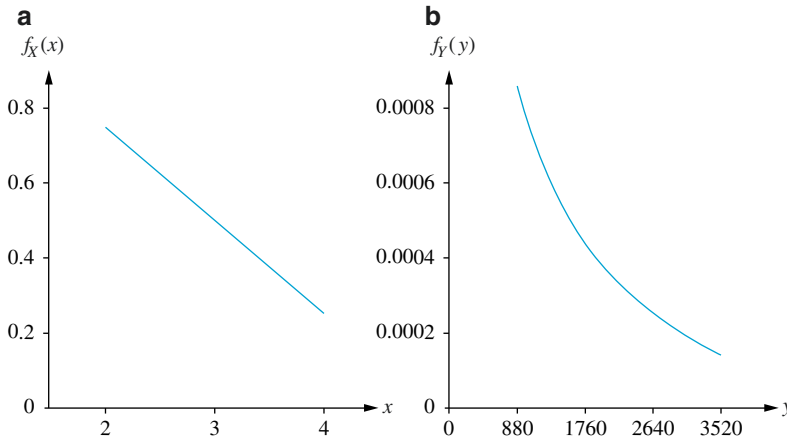
Figure 4.39   pdfs from Example 4.42: (**a**) pdf of X; (**b**) pdf of Y ∎

The Transformation Theorem requires a monotonic transformation, but there are important applications in which the transformation is not monotone. Nevertheless, it may be possible to use the theorem anyway with a little trickery.

Example 4.43   In this example, we start with a standard normal random variable Z, and we transform to $Y = Z^2$. (The squares of normal random variables are important because the sample variance is built from squares, and we will subsequently need the distribution of the sample variance.) This is *not* monotonic over the interval for Z, $(-\infty, \infty)$. However, consider the transformation $U = |Z|$. Because Z has a symmetric distribution, the pdf of U is $f_U(u) = f_Z(u) + f_Z(-u) = 2 f_Z(u)$. Don't despair if this is not intuitively clear, because we'll verify it shortly. For the time being, assume it to be true. Then $Y = Z^2 = |Z|^2 = U^2$, and the transformation in terms of U is monotonic because its set of possible values is $(0, \infty)$. Thus we can use the Transformation Theorem with $h(y) = y^{1/2}$:

$$f_Y(y) = f_U[h(y)]|h'(y)| = 2f_Z[h(y)]|h'(y)|$$

$$= \frac{2}{\sqrt{2\pi}} e^{-.5(y^{1/2})^2} \left| \frac{1}{2} y^{-1/2} \right| = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \quad y > 0$$

You were asked to believe intuitively that $f_U(u) = 2 f_X(u)$. Here is a little derivation that works as long as the distribution of Z is symmetric about 0. If $u > 0$,

$$F_U(u) = P(U \leq u) = P(|Z| \leq u) = P(-u \leq Z \leq u) = 2P(0 \leq Z \leq u)$$
$$= 2[F_Z(u) - F_Z(0)].$$

Differentiating this with respect to u gives $f_U(u) = 2 f_X(u)$.   ∎

Example 4.44   Sometimes the Transformation Theorem cannot be used at all, and you need to use the cdf. Let $f_X(x) = (x + 1)/8$, $-1 \leq x \leq 3$, and $Y = X^2$. The transformation is not monotonic on $[-1, 3]$; and, since $f_X(x)$ is not an even function, we can't employ the symmetry trick of the previous example. Possible values of Y are $\{y: 0 \leq y \leq 9\}$. Considering first $0 \leq y \leq 1$,

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = \int\limits_{-\sqrt{y}}^{\sqrt{y}} \frac{u+1}{8} du = \frac{\sqrt{y}}{4}$$

Then, on the other subinterval, $1 < y \le 9$,

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = P(-1 \le X \le \sqrt{y})$$

$$= \int\limits_{-1}^{\sqrt{y}} \frac{u+1}{8} du = (1 + y + 2\sqrt{y})/16$$

Differentiating, we get

$$f_Y(y) = \begin{cases} \dfrac{1}{8\sqrt{y}} & 0 < y \le 1 \\[2mm] \dfrac{y + \sqrt{y}}{16y} & 1 < y \le 9 \end{cases}$$

Figure 4.40 shows the pdfs of both X and Y.
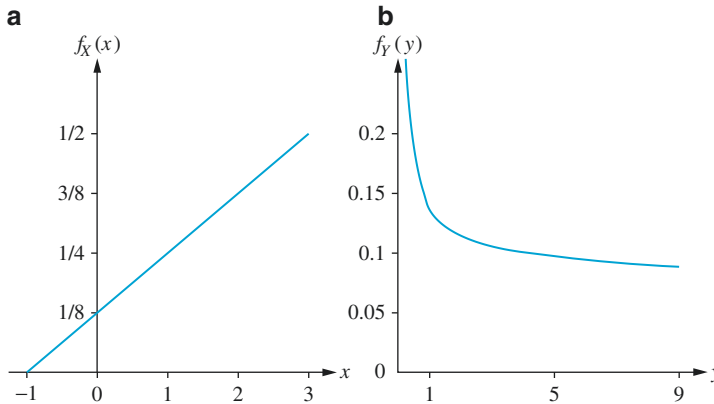
Figure 4.40 Pdfs from Example 4.44: (**a**) pdf of X; (**b**) pdf of Y ■

Exercises: Section 4.7 (110–124)

110. Relative to the winning time, the time X of another runner in a ten kilometer race has pdf $f_X(x) = 2/x^3$, $x > 1$. The reciprocal $Y = 1/X$ represents the ratio of the time for the winner divided by the time of the other runner. Find the pdf of Y. Explain why Y also represents the speed of the other runner relative to the winner.

111. Let X be the fuel efficiency in miles per gallon of an extremely inefficient vehicle (a military tank, perhaps?), and suppose X has

the pdf $f_X(x) = 2x$, $0 < x < 1$. Determine the pdf of $Y = 1/X$, which is fuel efficiency in gallons per mile. [*Note:* The distribution of Y is a special case of the *Pareto distribution* (see Exercise 10).]

112. Let X have the pdf $f_X(x) = 2/x^3$, $x > 1$. Find the pdf of $Y = \sqrt{X}$.

113. Let X have an exponential distribution with mean 2, so $f_X(x) = \frac{1}{2}e^{-x/2}$, $x > 0$. Find the pdf of $Y = \sqrt{X}$. [*Note:* Suppose you choose

a point in two dimensions randomly, with the horizontal and vertical coordinates chosen independently from the standard normal distribution. Then X has the distribution of the squared distance from the origin and Y has the distribution of the distance from the origin. Y has a *Rayleigh distribution* (see Exercise 4).]

114. If X is distributed as $N(\mu, \sigma)$, find the pdf of $Y = e^X$. Verify that the distribution of Y matches the lognormal pdf provided in Section 3.5.

115. If the length of a side of a square X is random with the pdf $f_X(x) = x/8$, $0 < x < 4$, and Y is the area of the square, find the pdf of Y.

116. Let $X \sim$ Unif(0, 1). Determine the pdf of $Y = -\ln(X)$.

117. Let $X \sim$ Unif(0, 1). Determine the pdf of $Y = \tan[\pi(X - .5)]$. [*Note:* The random variable Y has the *Cauchy distribution*, named after the famous mathematician.]

118. If $X \sim$ Unif[0, 1], find a linear transformation $Y = cX + d$ such that Y is uniformly distributed on $[A, B]$, where A and B are any two numbers such that $A < B$. Is there any other solution? Explain.

119. If X has the pdf $f_X(x) = x/8$, $0 < x < 4$, find a transformation $Y = g(X)$ such that $Y \sim$ Unif[0, 1]. [*Hint:* The target is to achieve $f_Y(y) = 1$ for $0 \le y \le 1$. The Transformation Theorem will allow you to find $h(y)$, from which $g(x)$ can be obtained.]

120. a. If a measurement error X is uniformly distributed on [–1, 1], find the pdf of $Y = |X|$, which is the magnitude of the measurement error.
    b. If $X \sim$ Unif[–1, 1], find the pdf of $Y = X^2$.
    c. If $X \sim$ Unif[–1, 3], find the pdf of $Y = X^2$.

121. If a measurement error X is distributed as $N(0, 1)$, find the pdf of $|X|$, which is the magnitude of the measurement error.

122. AAnn is expected at 7:00 pm after an all-day drive. She may be as much as one hour early or as much as three hours late. Assuming that her arrival time X is uniformly distributed over that interval, find the pdf of $|X - 7|$, the absolute difference between her actual and predicted arrival times.

123. A circular target has radius 1 foot. Assume that you hit the target (we shall ignore misses) and that the probability of hitting any region of the target is proportional to the region's area. If you hit the target at a distance Y from the center, then let $X = \pi Y^2$ be the corresponding area. Show that

    a. X is uniformly distributed on $[0, \pi]$. [*Hint:* Show that $F_X(x) = P(X \le x) = x/\pi$.]
    b. Y has pdf $f_Y(y) = 2y$, $0 < y < 1$.

124. In the previous exercise, suppose instead that Y is uniformly distributed on [0, 1]. Find the pdf of $X = \pi Y^2$. Geometrically speaking, why should X have a pdf that is unbounded near 0?

4.8 Simulation of Continuous Random Variables

In Sections 2.6 and 3.8, we discussed the need for simulation of random events and discrete random variables in situations where an "analytic" solution is very difficult or simply not possible. This section presents methods for simulating continuous random variables, including some of the built-in simulation tools of R.

The Inverse CDF Method

Section 3.8 introduced the inverse cdf method for simulating discrete random variables. The basic idea was this: generate a Unif[0, 1] random number and align it with the cdf of the random variable X we want to simulate. Then, determine which X value corresponds to that cdf value. We now extend

this methodology to the simulation of values from a continuous distribution; the heart of the algorithm relies on the following theorem, often called the **probability integral transform**.

THEOREM   Consider a continuous distribution with pdf f and cdf F. Let $U \sim \text{Unif}[0, 1)$, and define a random variable X by

$$X = F^{-1}(U) \tag{4.13}$$

Then the pdf of X is f.

Before proving this theorem, let's consider its practical usage: Suppose we want to simulate a continuous rv whose pdf is $f(x)$, i.e., obtain successive values of X having pdf $f(x)$. If we can determine the corresponding cdf $F(x)$ and apply its inverse F^{-1} to values $u_1, \ldots, u_n$, obtained from a standard uniform distribution, then $x_1 = F^{-1}(u_1), \ldots, x_n = F^{-1}(u_n)$ will be values from the desired distribution f. A graphical description of the algorithm appears in Figure 4.41.
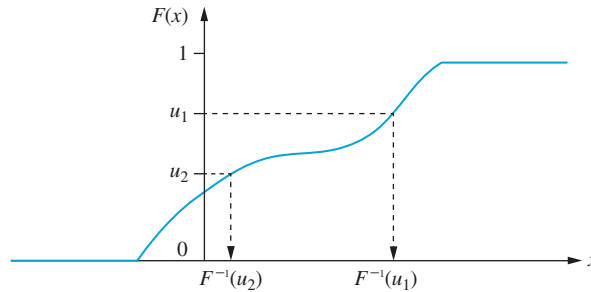
Figure 4.41 The inverse cdf method, illustrated

Proof Apply the Transformation Theorem (Section 4.7) with $f_U(u) = 1$ for $0 \le u < 1$, $X = g(U) = F^{-1}(U)$, and thus $U = h(X) = g^{-1}(X) = F(X)$. The pdf of the transformed variable X is

$$f_X(x) = f_U(h(x)) \cdot |h'(x)| = f_U(F(x)) \cdot |F'(x)| = 1 \cdot |f(x)| = f(x)$$

In the last step, the absolute values may be removed because a pdf is always nonnegative. ∎

The following box describes the implementation of the inverse cdf method justified by the preceding theorem.

INVERSE CDF METHOD	It is desired to simulate n values from a distribution pdf $f(x)$. Let $F(x)$ be the corresponding cdf. Repeat n times: 1. Use a random number generator (RNG) to produce a value, u, from [0, 1). 2. Assign $x = F^{-1}(u)$. The resulting values $x_1, \ldots, x_n$ form a simulation of a random variable with the original pdf, $f(x)$.

Example 4.45 Consider the electrical current distribution model of Example 4.42, where the pdf of X is given by $f(x) = 1.25 - .25x$ for $2 \leq x \leq 4$. Suppose a simulation of X is required as part of some larger system analysis. To implement the above method, the inverse of the cdf of X is required. First, compute the cdf:

$$F(x) = P(X \leq x) = \int_2^x f(y)\,dy$$

$$= \int_2^x (1.25 - .25y)\,dy = -0.125x^2 + 1.25x - 2 \quad 2 \leq x \leq 4$$

To find the probability integral transform (4.13), set $u = F(x)$ and solve for x:

$$u = F(x) = -0.125x^2 + 1.25x - 2 \Rightarrow x = F^{-1}(u) = 5 - \sqrt{9 - 8u}$$

The equation above can be solved using the quadratic formula; care must be taken to select the solution whose values lie in the interval [2, 4] (the other solution, $x = 5 + \sqrt{9 - 8u}$, does not have that feature). Beginning with the usual Unif[0, 1) RNG, the algorithm for simulating X is the following: given a value u from the RNG, assign $x = 5 - \sqrt{9 - 8u}$. Repeating this algorithm n times gives n simulated values of X. An R program that implements this algorithm appears in Figure 4.42; it returns a vector, x, containing $n = 10{,}000$ simulated values of the specified distribution.

```
x <- NULL
for (i in 1:10000){
      u<-runif(1)
      x[i]<-5-sqrt(9-8*u)
}
```

Figure 4.42 R simulation code for Example 3.42

As discussed in Chapter 2, this program can be accelerated by "vectorizing" the operations rather than using a for loop. In fact, a single line of code can produce the desired result:

```
x<-5-sqrt(9-8*runif(10000))
```

The pdf of the rv X and a histogram of simulation results appear in Figure 4.43.
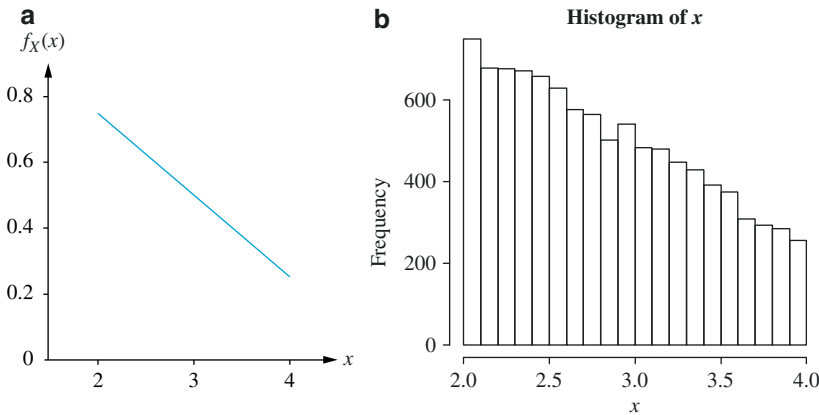
Figure 4.43 (**a**) Theoretical pdf and (**b**) R simulation results for Example 4.45   ∎

Example 4.46 The lifetime of a certain type of drill bit has an exponential distribution with mean 100 h. An analysis of a large manufacturing process that uses these drill bits requires the simulation of this lifetime distribution, which can be achieved through the inverse cdf method. From Section 4.4, the cdf of this exponential distribution is $F(x) = 1 - e^{-.01x}$, and so the inverse cdf is $x = F^{-1}(u) = -100\ln(1 - u)$. Applying this function to Unif[0, 1) random numbers will generate the desired simulation. (Don't let the negative sign at the front worry you: since $0 \leq u < 1$, $1 - u$ lies between 0 and 1, and so its logarithm is negative and the resulting value of x is actually positive.)

As a check, the code x=-100*log(1-runif(10000)) was submitted to R and the resulting sample mean and sd were obtained using mean(x) and sd(x). Exponentially distributed rvs have standard deviation equal to the mean, so the theoretical answers are $\mu = 100$ and $\sigma = 100$. The simulation yielded $\bar{x} = 99.3724$ and $s = 100.8908$, both of which are reasonably close to 100 and validate the inverse cdf formula.

In general, an exponential distribution with mean μ (equivalently, parameter $\lambda = 1/\mu$) can be simulated using the transform $x = -\mu\ln(1 - u)$.   ∎

The preceding two examples illustrated the inverse cdf method for fairly simple density functions: a linear polynomial and an exponential function. In practice, the algebraic complexity of $f(x)$ can often be a barrier to implementing this simulation technique. After all, the algorithm requires that we can (1) obtain the cdf $F(x)$ in closed form and (2) find the inverse function of F in closed form. Consider, for example, attempting to simulate values from the $N(0, 1)$ distribution: its cdf is the function denoted $\Phi(z)$ and given by the integral expression $(1/\sqrt{2\pi}) \int_{-\infty}^{z} e^{-u^2/2} du$. There is no closed-form expression for this integral, let alone a method to solve $u = \Phi(z)$ for z and implement (4.13). (As a reminder, the lack of a closed-form expression for $\Phi(z)$ is the reason that software or tables are always required for calculations involving normal probabilities.) Thankfully, most statistical software packages have built-in tools to simulate normally distributed variates (using a very clever algorithm called the *Box-Muller method*; see Section 5.6). We'll discuss R's built-in simulation tools at the end of this section.

As the next example illustrates, even when $F(x)$ can be determined in closed form we cannot necessarily implement the inverse cdf method, because $F(x)$ cannot always be inverted. This difficulty surfaces in practice when attempting to simulate values from a gamma distribution.

Example 4.47 The measurement error X (in mV) of a particular voltmeter has the following distribution: $f(x) = (4 - x^2)/9$ for $-1 \leq x \leq 2$ (and $f(x) = 0$ otherwise). To use the inverse cdf method to simulate X, begin by calculating its cdf:

$$F(x) = \int_{-1}^{x} \frac{4 - y^2}{9}\, dy = \frac{-x^3 + 12x + 11}{27}$$

To implement step 2 of the inverse cdf method requires solving $F(x) = u$ for x; since $F(x)$ is a cubic polynomial, this is not a simple task. Advanced computer algebra systems can solve this equation, though the general solution is unwieldy (and such a solution doesn't exist at all for 5th-degree and higher polynomials). Readers familiar with numerical analysis methods may recognize that, for any specified numerical value of u, a root-finding algorithm (such as Newton–Raphson) can be implemented to *approximate* the solution x. This latter method, however, is computationally intensive, especially if it's desirable to generate 10,000 or more simulated values of x. ∎

The preceding example suggests that in practice not every continuous distribution can be simulated via the inverse cdf method. When the inverse cdf method of simulation cannot be implemented, the *accept–reject method* provides an alternative. The downside of the accept–reject method is that only some of the random numbers generated by software will be used ("accepted"), while others will be "rejected." As a result, one needs to create more—sometimes, many more—random variates than the desired number of simulated values. For information on the accept–reject method, consult the texts by Ross or Carlton and Devore listed in the bibliography.

Built-in Simulation Packages for R
As was true for the most common discrete distributions, many software packages have built-in tools for simulating values from the continuous models named in this chapter. Table 4.4 summarizes the relevant R functions for the uniform, normal, gamma, and exponential distributions; the variable n refers to the desired number of simulated values of the distribution. R includes similar commands for the Weibull, lognormal, and beta distributions.

Table 4.4 Functions to simulate major continuous distributions in R

Distribution	R code
Unif[A, B]	`runif(n,A,B)`
$N(\mu, \sigma)$	`rnorm(n,μ,σ)`
Gamma(α, β)	`rgamma(n,α,1/β)`
Exponential(λ)	`rexp(n,λ)`

As was the case with the cdf commands discussed in Section 4.4, R parameterizes the gamma and exponential distributions using the "rate" parameter $\lambda = 1/\beta$. In the gamma simulation command, this can be overridden by naming the final argument `scale`, as in `rgamma(n,α,scale=β)`. The command `rnorm(n)` will generate standard normal variates (i.e., with $\mu = 0$ and $\sigma = 1$). Similarly, R will generate standard uniform variates ($A = 0$ and $B = 1$), the basis for many of our simulation methods, with the command `runif(n)`.

Precision of Simulation Results

Section 3.8 discusses in detail the precision of estimates associated with simulating discrete random variables. The same results apply in the continuous case. In particular, the estimated standard error in using a sample proportion $\hat{p}$ to estimate the true probability of an event is still $\sqrt{\hat{p}(1 - \hat{p})/n}$, where n is the simulation size. Also, the estimated standard error in using a sample mean, $\bar{x}$, to estimate the true expected value μ of a (continuous) rv X is $s/\sqrt{n}$, where s is the sample standard deviation of the simulated values of X. Refer back to Section 3.8 for more details.

Exercises: Section 4.8 (125–130)

125. The amount of time (hours) required to complete an unusually short statistics homework assignment is modeled by the pdf $f(x) = x/2$ for $0 < x < 2$ (and $= 0$ otherwise).

    a. Obtain the cdf and then its inverse.
    b. Write a program to simulate 10,000 values from this distribution.
    c. Compare the sample mean and standard deviation of your 10,000 simulated values to the theoretical mean and sd of this distribution (which you can determine by calculating the appropriate integrals).

126. The Weibull distribution was introduced in Section 4.5.

    a. Find the inverse of the Weibull cdf.
    b. Write a program to simulate n values from a Weibull distribution. Your program should have three inputs: the desired number of simulated values n and the two parameters α and β. It should have a single output: an $n \times 1$ vector of simulated values.
    c. Use your program from part (b) to simulate 10,000 values from a Weibull(4, 6) distribution and estimate the mean of this distribution. The correct value of the mean is $6\Gamma(5/4) \approx 5.438$; how close is your sample mean?

127. Consider the pdf for the rv X = magnitude (in newtons) of a dynamic load on a bridge, given in Example 4.7:

$$f(x) = \frac{1}{8} + \frac{3}{8}x \quad 0 \le x \le 2$$

Write a program to simulate values from this distribution using the inverse cdf method.

128. In distributed computing, any given task is split into smaller subtasks which are handled by separate processors (which are then re-combined by a multiplexer). Consider a distributed computing system with 4 processors, and suppose for one particular purpose that pdf of completion time for a particular subtask (microseconds) on any one of the processors is given by $f(x) = 20/(3x^2)$ for $4 \le x \le 10$ and $= 0$ otherwise. That is, the subtask completion times X_1, X_2, X_3, X_4 of the four processors each have the specified pdf.

    a. Write a program to simulate the above pdf using the inverse cdf method.
    b. The overall time to complete any task is the largest of the four subtask completion times: if we call this variable Y, then $Y = \max(X_1, X_2, X_3, X_4)$. (We assume that the multiplexing time is negligible.) Use your program in part (a) to simulate 10,000 values of the rv Y. Create a

histogram of the simulated values of Y, and also use your simulation to estimate both $E(Y)$ and σ_Y.

129. Consider the following pdf:

$$f(x;\theta,\tau) = \frac{\theta}{\tau}(1 - x/\tau)^{\theta-1} \quad 0 \le x < \tau$$

where $\theta > 0$ and $\tau > 0$ are the parameters of the model. [This pdf is suggested for modeling waiting time in the article "A Model of Pedestrians' Waiting Times for Street Crossings at Signalized Intersections" (*Trans. Res.* 2013: 17–28).]

  a. Write a function to simulate values from this distribution, implementing the inverse cdf method. Your function should have three inputs: the desired number of simulated values n and values for the two parameters for θ and τ.
  b. Use your function in part (a) to simulate 10,000 values from this wait time distribution with $\theta = 4$ and $\tau = 80$. Estimate $E(X)$ under these parameter settings. How close is your estimate to the correct value of 16?

130. Explain why the transformation $x = -\mu\ln(u)$ may be used to simulate values from an exponential distribution with mean μ. (This expression is slightly simpler than the one established in this section.)

Supplementary Exercises: (131–159)

131. An insurance company issues a policy covering losses up to 5 (in thousands of dollars). The loss, X, follows a distribution with density function $f(x) = 3/x^4$ for $x \ge 1$ and $= 0$ otherwise. What is the expected value of the amount paid under the policy?

132. A 12-in. bar clamped at both ends is subjected to an increasing amount of stress until it snaps. Let Y = the distance from the left end at which the break occurs. Suppose Y has pdf

$$f(y) = \frac{y}{24}\left(1 - \frac{y}{12}\right) \quad 0 \le y \le 12$$

Compute the following:
  a. The cdf of Y, and graph it.
  b. $P(Y \le 4), P(Y > 6)$, and $P(4 \le Y \le 6)$.
  c. $E(Y), E(Y^2)$, and $V(Y)$.
  d. The probability that the break point occurs more than 2 in. from the expected break point.
  e. The expected length of the shorter segment when the break occurs.

133. Let X denote the time to failure (in years) of a hydraulic component. Suppose the pdf of X is $f(x) = 32/(x + 4)^3$ for $x > 0$.
  a. Verify that $f(x)$ is a legitimate pdf.
  b. Determine the cdf.
  c. Use the result of part (b) to calculate the probability that time to failure is between 2 and 5 years.
  d. What is the expected time to failure?
  e. If the component has a salvage value equal to $100/(4 + x)$ when its time to failure is x, what is the expected salvage value?

134. The completion time X for a task has cdf $F(x)$ given by

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^3}{3} & 0 \le x < 1 \\ 1 - \frac{1}{2}\left(\frac{7}{3} - x\right)\left(\frac{7}{4} - \frac{3}{4}x\right) & 1 \le x < \frac{7}{3} \\ 1 & x \ge \frac{7}{3} \end{cases}$$

  a. Obtain the pdf $f(x)$ and sketch its graph.
  b. Compute $P(.5 \le X \le 2)$.
  c. Compute $E(X)$.

135. The breakdown voltage of a randomly chosen diode of a certain type is known to be normally distributed with mean value 40 V and standard deviation 1.5 V.

  a. What is the probability that the voltage of a single diode is between 39 and 42?

b. What value is such that only 15% of all diodes have voltages exceeding that value?

c. If four diodes are independently selected, what is the probability that at least one has a voltage exceeding 42?

136. The article "Computer Assisted Net Weight Control" (*Qual. Prog.* 1983: 22–25) suggests a normal distribution with mean 137.2 oz and standard deviation 1.6 oz, for the actual contents of jars of a certain type. The stated contents was 135 oz.

a. What is the probability that a single jar contains more than the stated contents?

b. Among ten randomly selected jars, what is the probability that at least eight contain more than the stated contents?

c. Assuming that the mean remains at 137.2, to what value would the standard deviation have to be changed so that 95% of all jars contain more than the stated contents?

137. When circuit boards used in the manufacture of MP3 players are tested, the long-run percentage of defectives is 5%. Suppose that a batch of 250 boards has been received and that the condition of any particular board is independent of that of any other board.

a. What is the approximate probability that at least 10% of the boards in the batch are defective?

b. What is the approximate probability that there are exactly ten defectives in the batch?

138. Let X be a nonnegative continuous random variable with pdf $f(x)$, cdf $F(x)$, and mean $E(X)$.

a. The definition of expected value is $E(X) = \int_0^\infty x f(x) dx$. Replace the first x inside the integral with $\int_0^x 1\, dy$ to create a double integral expression for $E(X)$. [The "order of integration" should be $dy\, dx$.]

b. Re-arrange the order of integration, keeping track of the revised limits of integration, to show that

$$E(X) = \int_0^\infty \int_y^\infty f(x) dx\, dy$$

c. Evaluate the dx integral in (b) to show that $E(X) = \int_0^\infty [1 - F(y)] dy$. (This provides an alternate derivation of the formula established in Exercise 38.)

d. Use the result of (c) to verify that the expected value of an exponentially distributed rv with parameter λ is $1/\lambda$.

139. The reaction time (in seconds) to a stimulus is a continuous random variable with pdf $f(x) = 1.5/x^2$ for $1 \le x \le 3$ and $= 0$ otherwise.

a. Obtain the cdf.

b. Using the cdf, what is the probability that reaction time is at most 2.5 s? Between 1.5 and 2.5 s?

c. Compute the expected reaction time.

d. Compute the standard deviation of reaction time.

e. If an individual takes more than 1.5 s to react, a light comes on and stays on either until one further second has elapsed or until the person reacts (whichever happens first). Determine the expected amount of time that the light remains lit. [*Hint*: Let $h(X)$ = the time that the light is on as a function of reaction time X.]

140. Let X denote the temperature at which a certain chemical reaction takes place. Suppose that X has pdf $f(x) = (4 - x^2)/9$ for $-1 \le x \le 2$ and $= 0$ otherwise.

a. Sketch the graph of $f(x)$.

b. Determine the cdf and sketch it.

c. Is 0 the median temperature at which the reaction takes place? If not, is the median temperature smaller or larger than 0?

d. Suppose this reaction is independently carried out once in each of ten different

laboratories and that the pdf of reaction time in each laboratory is as given. Let Y = the number among the ten laboratories at which the temperature exceeds 1. What kind of distribution does Y have? (Give the name and values of any parameters.)

141. The article "Determination of the MTF of Positive Photoresists Using the Monte Carlo Method" (*Photographic Sci. Engr.* 1983: 254–260) proposes the exponential distribution with parameter $\lambda = .93$ as a model for the distribution of a photon's free path length (μm) under certain circumstances. Suppose this is the correct model.

a. What is the expected path length, and what is the standard deviation of path length?

b. What is the probability that path length exceeds 3.0? What is the probability that path length is between 1.0 and 3.0?

c. What value is exceeded by only 10% of all path lengths?

142. The article "The Prediction of Corrosion by Statistical Analysis of Corrosion Profiles" (*Corrosion Sci.* 1985: 305–315) suggests the following cdf for the depth X of the deepest pit in an experiment involving the exposure of carbon manganese steel to acidified seawater.

$$F(x; \theta_1, \theta_2) = e^{-e^{-(x-\theta_1)/\theta_2}} \quad -\infty < x < \infty$$

(This is called the *Gumbel distribution*.) The investigators proposed the values $\theta_1 = 150$ and $\theta_2 = 90$. Assume this to be the correct model.

a. What is the probability that the depth of the deepest pit is at most 150? At most 300? Between 150 and 300?

b. Below what value will the depth of the maximum pit be observed in 90% of all such experiments?

c. What is the density function of X?

d. The density function can be shown to be unimodal (a single peak). Above what value on the measurement axis does this peak occur? (This value is the *mode*.)

e. It can be shown that $E(X) \approx .5772\theta_2 + \theta_1$. What is the mean for the given values of θ_1 and θ_2, and how does it compare to the median and mode? Sketch the graph of the density function.

143. Let t = the amount of sales tax a retailer owes the government for a certain period. The article "Statistical Sampling in Tax Audits" (*Statistics Law* 2008: 320–343) proposes modeling the uncertainty in t by regarding it as a normally distributed random variable with mean value μ and standard deviation σ (in the article, these two parameters are estimated from the results of a tax audit involving n sampled transactions). If a represents the amount the retailer is assessed, then an underassessment results if $t > a$ and an overassessment if $a > t$. We can express this in terms of a *loss function*, a function that shows zero loss if $t = a$ but increases as the gap between t and a increases. The proposed loss function is $L(a, t) = t - a$ if $t > a$ and $= k(a - t)$ if $t \le a$ ($k > 1$ is suggested to incorporate the idea that overassessment is more serious than underassessment).

a. Show that $a^* = \mu + \sigma \Phi^{-1}(1/(k+1))$ is the value of a that minimizes the expected loss, where Φ^{-1} is the inverse function of the standard normal cdf.

b. If $k = 2$ (suggested in the article), $\mu = \$100,000$, and $\sigma = \$10,000$, what is the optimal value of a, and what is the resulting probability of overassessment?

144. A *mode* of a continuous distribution is a value x^* that maximizes $f(x)$.

a. What is the mode of a normal distribution with parameters μ and σ?

b. Does the uniform distribution with parameters A and B have a single mode? Why or why not?

c. What is the mode of an exponential distribution with parameter λ? (Draw a picture.)

d. If X has a gamma distribution with parameters α and β, and $\alpha > 1$, find the mode. [*Hint*: $\ln[f(x)]$ will be maximized if and only if $f(x)$ is, and it may be simpler to take the derivative of $\ln[f(x)]$.]

145. The article "Error Distribution in Navigation" (*J. Institut. Navigation* 1971: 429–442) suggests that the frequency distribution of positive errors (magnitudes of errors) is well approximated by an exponential distribution. Let $X =$ the lateral position error (nautical miles), which can be either negative or positive. Suppose the pdf of X is

$$f(x) = .1\, e^{-.2|x|} \quad -\infty < x < \infty$$

a. Sketch a graph of $f(x)$ and verify that it is a legitimate pdf (show that it integrates to 1).

b. Obtain the cdf of X and sketch it.

c. Compute $P(X \le 0)$, $P(X \le 2)$, $P(-1 \le X \le 2)$, and the probability that an error of more than 2 miles is made.

146. In some systems, a customer is allocated to one of two service facilities. If the service time for a customer served by facility i has an exponential distribution with parameter λ_i ($i = 1, 2$) and p is the proportion of all customers served by facility 1, then the pdf of $X =$ the service time of a randomly selected customer is

$$f(x; \lambda_1, \lambda_2, p) =$$
$$p\lambda_1 e^{-\lambda_1 x} + (1 - p)\lambda_2 e^{-\lambda_2 x} \quad x > 0$$

This is often called the *hyperexponential* or *mixed exponential distribution*. This

distribution is also proposed in the article "Statistical Behavior Modeling for Driver-Adaptive Precrash Systems" (*IEEE Trans. Intelligent Transp. Syst.* 2013: 1–9) as a model for modeling what the authors call "the criticality level of a situation."

a. Verify that $f(x; \lambda_1, \lambda_2, p)$ is indeed a pdf.

b. If $p = .5$, $\lambda_1 = 40$, $\lambda_2 = 200$ (λ values suggested in the cited article), calculate $P(X > .01)$.

c. If X has $f(x; \lambda_1, \lambda_2, p)$ as its pdf, what is $E(X)$?

d. Using the fact that $E(X^2) = 2/\lambda^2$ when X has an exponential distribution with parameter λ, compute $E(X^2)$ when X has pdf $f(x; \lambda_1, \lambda_2, p)$. Then compute $V(X)$.

e. The *coefficient of variation* of a random variable (or distribution) is $CV = \sigma/\mu$. What is the CV for an exponential rv? What can you say about the value of CV when X has a hyperexponential distribution?

f. What is the CV for an Erlang distribution with parameters λ and n as defined in Exercise 78? [*Note*: In applied work, the sample CV is used to decide which of the three distributions might be appropriate.]

g. For the parameter values given in (b), calculate the probability that X is within one standard deviation of its mean value. Does this probability depend upon the values of the λ's (it does not depend on λ when X has an exponential distribution)?

147. Suppose a state allows individuals filing tax returns to itemize deductions only if the total of all itemized deductions is at least $5000. Let X (in 1000's of dollars) be the total of itemized deductions on a randomly chosen form. Assume that X has the pdf

$$f(x; \alpha) = k/x^{\alpha} \quad x \ge 5$$

a. Find the value of k. What restriction on α is necessary?
b. What is the cdf of X?
c. What is the expected total deduction on a randomly chosen form? What restriction on α is necessary for $E(X)$ to be finite?
d. Show that $\ln(X/5)$ has an exponential distribution with parameter $\alpha - 1$.

148. Let I_i be the input current to a transistor and I_o be the output current. Then the current gain is proportional to $\ln(I_o/I_i)$. Suppose the constant of proportionality is 1 (which amounts to choosing a particular unit of measurement), so that current gain = X = $\ln(I_o/I_i)$. Assume X is normally distributed with $\mu = 1$ and $\sigma = .05$.

a. What type of distribution does the ratio I_o/I_i have?
b. What is the probability that the output current is more than twice the input current?
c. What are the expected value and variance of the ratio of output to input current?

149. The article "Response of SiC_f/Si_3N_4 Composites Under Static and Cyclic Loading—An Experimental and Statistical Analysis" (*J. Engr. Mater. Tech.* 1997: 186–193) suggests that tensile strength (MPa) of composites under specified conditions can be modeled by a Weibull distribution with $\alpha = 9$ and $\beta = 180$.

a. Sketch a graph of the density function.
b. What is the probability that the strength of a randomly selected specimen will exceed 175? Will be between 150 and 175?
c. If two randomly selected specimens are chosen and their strengths are independent of each other, what is the probability that at least one has strength between 150 and 175?
d. What strength value separates the weakest 10% of all specimens from the remaining 90%?

150. Suppose the lifetime X of a component, when measured in hours, has a gamma distribution with parameters α and β.
a. Let Y = lifetime measured in minutes. Derive the pdf of Y.
b. What is the probability distribution of $Y = cX$?

151. Based on data from a dart-throwing experiment, the article "Shooting Darts" (*Chance*, Summer 1997: 16–19) proposed that the horizontal and vertical errors from aiming at a point target should be independent of each other, each with a normal distribution having mean 0 and variance σ^2. It can then be shown that the pdf of the distance V from the target to the landing point is

$$f(v) = \frac{v}{\sigma^2} \cdot e^{-v^2/(2\sigma^2)} \quad v > 0$$

a. This pdf is a member of what family introduced in this chapter?
b. If $\sigma = 20$ mm (close to the value suggested in the paper), what is the probability that a dart will land within 25 mm (roughly 1 in.) of the target?

152. The article "Three Sisters Give Birth on the Same Day" (*Chance*, Spring 2001: 23–25) used the fact that three Utah sisters had all given birth on March 11, 1998, as a basis for posing some interesting questions regarding birth coincidences.

a. Disregarding leap year and assuming that the other 365 days are equally likely, what is the probability that three randomly selected births all occur on March 11? Be sure to indicate what, if any, extra assumptions you are making.
b. With the assumptions used in part (a), what is the probability that three randomly selected births all occur on the same day?
c. The author suggested that, based on extensive data, the length of gestation (time between conception and birth) could be modeled as having a normal

distribution with mean value 280 days and standard deviation 19.88 days. The due dates for the three Utah sisters were March 15, April 1, and April 4, respectively. Assuming that all three due dates are at the mean of the distribution, what is the probability that all births occurred on March 11? [*Hint*: The deviation of birth date from due date is normally distributed with mean 0.]

d. Explain how you would use the information in part (c) to calculate the probability of a common birth date.

153. Let *X* denote the lifetime of a component, with *f*(*x*) and *F*(*x*) the pdf and cdf of *X*. The probability that the component fails in the interval (*x*, *x* + Δ*x*) is approximately *f*(*x*) · Δ*x*. The conditional probability that it fails in (*x*, *x* + Δ*x*) given that it has lasted at least *x* is *f*(*x*) · Δ*x*/[1 − *F*(*x*)]. Dividing this by Δ*x* produces the **failure rate function**:

$$r(x) = \frac{f(x)}{1 - F(x)}$$

An increasing failure rate function indicates that older components are increasingly likely to wear out, whereas a decreasing failure rate is evidence of increasing reliability with age. In practice, a "bathtub-shaped" failure is often assumed.

a. If *X* is exponentially distributed, what is *r*(*x*)?

b. If *X* has a Weibull distribution with parameters α and β, what is *r*(*x*)? For what parameter values will *r*(*x*) be increasing? For what parameter values will *r*(*x*) decrease with *x*?

c. Since $r(x) = -(d/dx)\ln[1 - F(x)]$, $\ln[1 - F(x)] = \int r(x)dx$. Suppose

$$r(x) = \alpha\left(1 - \frac{x}{\beta}\right) \quad 0 \le x \le \beta$$

so that if a component lasts β hours, it will last forever (while seemingly unreasonable, this model can be used to study just "initial wearout"). What are the cdf and pdf of *X*?

154. Let *X* have a Weibull distribution with shape parameter α and scale parameter β. Show that the the transformed variable $Y = \ln(X)$ has an extreme value distribution as defined in Section 4.6, with $\theta_1 = \ln(\beta)$ and $\theta_2 = 1/\alpha$.

155. Let *X* have a Weibull distribution with parameters α = 2 and β. Show that $Y = 2X^2/\beta^2$ has a gamma distribution, and identify its parameters.

156. Let *X* have the pdf $f(x) = 1/[\pi(1 + x^2)]$ for $-\infty < x < \infty$ (a central Cauchy distribution), and show that $Y = 1/X$ has the same distribution. [*Hint*: Consider $P(|Y| \le y)$, the cdf of |*Y*|, then obtain its pdf and show it is identical to the pdf of |*X*|.]

157. A store will order *q* gallons of a liquid product to meet demand during a particular time period. This product can be dispensed to customers in any amount desired, so demand during the period is a continuous random variable *X* with cdf *F*(*x*). There is a fixed cost c_0 for ordering the product plus a cost of c_1 per gallon purchased. The per gallon sale price of the product is *d*. Liquid left unsold at the end of the time period has a salvage value of *e* per gallon. Finally, if demand exceeds *q*, there will be a shortage cost for loss of goodwill and future business; this cost is *f* per gallon of unfulfilled demand. Show that the value of *q* that maximizes expected profit, denoted by *q**, satisfies

$$P(\text{satisfying demand}) = F(q*)$$
$$= \frac{d - c_1 + f}{d - e + f}$$

Then determine the value of *F*(*q**) if *d* = \$35, c_0 = \$25, c_1 = \$15, *e* = \$5, and *f* = \$25. [*Hint*: Let *x* denote a particular value of *X*. Develop an expression for profit when $x \le q$ and another expression for profit when $x > q$. Now write an integral expression for expected profit (as a function of *q*) and differentiate.]

158. A function $g(x)$ is *convex* if the chord connecting any two points on the function's graph lies above the graph. When $g(x)$ is differentiable, an equivalent condition is that for every x, the tangent line at x lies entirely on or below the graph. (See the accompanying figure.) How does $g(\mu) = g[E(X)]$ compare to $E[g(X)]$? [*Hint*: The equation of the tangent line at $x = \mu$ is $y = g(\mu) + g'(\mu) \cdot (x - \mu)$. Use the condition of convexity, substitute X for x, and take expected values.] *Note*: Unless $g(x)$ is linear, the resulting inequality, usually called *Jensen's inequality*, is strict ($<$ rather than $\leq$); it is valid for both continuous and discrete rvs.
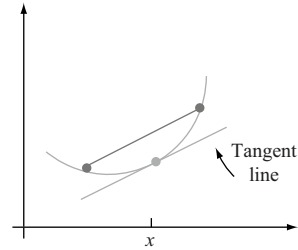
Joint Probability Distributions and Their Applications

<div align="right">**5**</div>

Introduction

In Chapters 3 and 4, we developed probability models for a single random variable. Many problems in probability and statistics lead to models involving several random variables simultaneously. In this chapter, we first discuss probability models for the joint behavior of several random variables, putting special emphasis on the case in which the variables are independent of each other. We then study expected values of functions of several random variables, including covariance and correlation as measures of the degree of association between two variables.

Section 5.3 develops properties of linear combinations of random variables, with particular emphasis on the sum and the average. The next section considers conditional distributions, the distributions of random variables given the values of other random variables. In Section 5.5 we extend the normal distribution of Chapter 4 to two possibly dependent rvs. The next section is about transformations of two or more random variables, generalizing the results of Section 4.7. In the last section of this chapter we discuss the distribution of order statistics: the minimum, maximum, median, and other quantities that can be found by arranging the observations in order.

5.1 Jointly Distributed Random Variables

There are many experimental situations in which more than one random variable (rv) will be of interest to an investigator. For example X might be the number of books checked out from a public library on a particular day and Y the number of videos checked out on the same day. Or X and Y might be the height and weight, respectively, of a randomly selected adult. In general, the two rvs of interest could both be discrete, both be continuous, or one could be discrete and the other continuous. In practice, the two "pure" cases—both of the same type—predominate. We shall first consider joint probability distributions for two discrete rvs, then for two continuous variables, and finally for more than two variables.

The Joint Probability Mass Function for Two Discrete Random Variables

The probability mass function (pmf) of a single discrete rv X specifies how much probability mass is placed on each possible X value. The joint pmf of two discrete rvs X and Y describes how much probability mass is placed on each possible pair of values (x, y).

DEFINITION   Let X and Y be two discrete rvs defined on the sample space $\mathcal{S}$ of an experiment. The
joint probability mass function $p(x, y)$ is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X = x \text{ and } Y = y)$$

A function $p(x, y)$ can be used as a joint pmf provided that $p(x, y) \geq 0$ for all x and y and
$\sum_x \sum_y p(x, y) = 1$. Let A be any set consisting of pairs of (x, y) values, such as $\{(x, y): x + y < 10\}$.
Then the probability that the random pair (X, Y) lies in A is obtained by summing the joint pmf over
pairs in A:

$$P((X, Y) \in A) = \sum_{(x, y) \in A} \sum p(x, y)$$

As in previous chapters, we will display a joint pmf for the values in its support—i.e., the set of all
(x, y) values for which $p(x, y) > 0$—with the understanding that $p(x, y) = 0$ otherwise.

Example 5.1 A large insurance agency services a number of customers who have purchased both a
homeowner's policy and an automobile policy from the agency. For each type of policy, a deductible
amount must be specified. For an automobile policy, the choices are $100 and $250, whereas for a
homeowner's policy, the choices are 0, $100, and $200. Suppose an individual with both types of
policy is selected at random from the agency's files. Let X = the deductible amount on the auto policy
and Y = the deductible amount on the homeowner's policy. Possible (X, Y) pairs are then (100, 0),
(100, 100), (100, 200), (250, 0), (250, 100), and (250, 200); the joint pmf specifies the probability
associated with each one of these pairs, with any other pair having probability zero. Suppose the joint
pmf is given in the accompanying **joint probability table**:

			y	
	$p(x, y)$	0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

Then $p(100, 100) = P(X = 100 \text{ and } Y = 100) = P(\$100 \text{ deductible on both policies}) = .10$. The
probability $P(Y \geq 100)$ is computed by summing probabilities of all (x, y) pairs for which $y \geq 100$:

$$P(Y \geq 100) = p(100, 100) + p(250, 100) + p(100, 200) + p(250, 200) = .75 \qquad \blacksquare$$

Looking at the joint probability table in Example 5.1, we see that $P(X = 100)$, i.e., $p_X(100)$, equals
.20 + .10 + .20 = .50, and similarly $p_X(250) = .05 + .15 + .30 = .50$ as well. That is, the pmf of X at
a specified number is calculated by fixing an x value (say, 100 or 250) and summing across all
possible y values; e.g., $p_X(250) = p(250,0) + p(250,100) + p(250,200)$. The pmf of Y can be obtained
by analogous summation, adding "down" the table instead of "across." In fact, by adding across rows
and down columns, we could imagine writing these probabilities in the margins of the joint proba-
bility table; for this reason, p_X and p_Y are called the *marginal* distributions of X and Y.

DEFINITION The **marginal probability mass functions** of X and of Y, denoted by $p_X(x)$ and $p_Y(y)$, respectively, are given by

$$p_X(x) = \sum_y p(x, y) \qquad p_Y(y) = \sum_x p(x, y)$$

Thus to obtain the marginal pmf of X evaluated at, say, $x = 100$, the probabilities $p(100, y)$ are added over all possible y values. Doing this for each possible X value gives the marginal pmf of X alone (i.e., without reference to Y). From the marginal pmfs, probabilities of events involving only X or only Y can be computed.

Example 5.2 (Example 5.1 continued) The possible X values are $x = 100$ and $x = 250$, so computing row totals in the joint probability table yields

$$p_X(100) = p(100, 0) + p(100, 100) + p(100, 200) = .50$$

And

$$p_X(250) = p(250, 0) + p(250, 100) + p(250, 200) = .50$$

The marginal pmf of X is then

$$p_X(x) = .50 \quad x = 100, 250$$

Similarly, the marginal pmf of Y is obtained from column totals as

$$p_Y(y) = \begin{cases} .25 & y = 0, \ 100 \\ .50 & y = 200 \end{cases}$$

so $P(Y \geq 100) = p_Y(100) + p_Y(200) = .75$ as before. ∎

The Joint Probability Density Function for Two Continuous Random Variables
The probability that the observed value of a continuous rv X lies in a one-dimensional set A (such as an interval) is obtained by integrating the pdf $f(x)$ over the set A. Similarly, the probability that the pair (X, Y) of continuous rvs falls in a two-dimensional set A (such as a rectangle) is obtained by integrating a function called the *joint density function*.

DEFINITION Let X and Y be continuous rvs. Then $f(x, y)$ is the **joint probability density function** for X and Y if for any two-dimensional set A

$$P((X, Y) \in A) = \iint\limits_A f(x, y) \, dx \, dy$$

In particular, if A is the two-dimensional rectangle $\{(x, y) : a \leq x \leq b, c \leq y \leq d\}$, then

$$P((X,Y) \in A) = P(a \le X \le b, c \le Y \le d) = \int\limits_a^b \int\limits_c^d f(x,y)dy\,dx$$

For $f(x, y)$ to be a joint pdf, it must satisfy $f(x, y) \ge 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)dx\,dy = 1$. We can think of $f(x, y)$ as specifying a surface at height $f(x, y)$ above the point (x, y) in a three-dimensional coordinate system. Then $P((X, Y) \in A)$ is the volume underneath this surface and above the region A, analogous to the area under a curve in the one-dimensional case. This is illustrated in Figure 5.1.
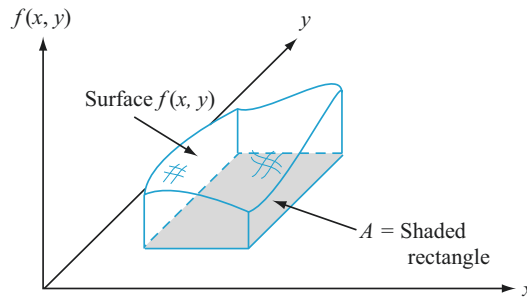
Figure 5.1 $P((X, Y) \in A)$ = volume under density surface above A

Example 5.3 A bank operates both a drive-up facility and a walk-up window. On a randomly selected day, let X = the proportion of time that the drive-up facility is in use (at least one customer is being served or waiting to be served) and Y = the proportion of time that the walk-up window is in use. Then the set of possible values for (X, Y) is the rectangle $D = \{(x, y) : 0 \le x \le 1, 0 \le y \le 1\}$. Suppose the joint pdf of (X, Y) is given by

$$f(x,y) = \frac{6}{5}(x+y^2) \quad 0 \le x \le 1, \ 0 \le y \le 1$$

To verify that this is a legitimate pdf, note that $f(x, y) \ge 0$ and

$$\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x,y)dxdy = \int\limits_0^1 \int\limits_0^1 \frac{6}{5}(x+y^2)dxdy$$

$$= \int\limits_0^1 \int\limits_0^1 \frac{6}{5}xdxdy + \int\limits_0^1 \int\limits_0^1 \frac{6}{5}y^2dxdy$$

$$= \int\limits_0^1 \frac{6}{5}xdx + \int\limits_0^1 \frac{6}{5}y^2dy = \frac{6}{10} + \frac{6}{15} = 1$$

The probability that neither facility is busy more than one-quarter of the time is

$$P\left(0 \le X \le \frac{1}{4}, 0 \le Y \le \frac{1}{4}\right) = \int_0^{1/4} \int_0^{1/4} \frac{6}{5}(x+y^2)dxdy$$

$$= \frac{6}{5}\int_0^{1/4} \int_0^{1/4} xdxdy + \frac{6}{5}\int_0^{1/4} \int_0^{1/4} y^2 dxdy$$

$$= \frac{6}{20} \cdot \frac{x^2}{2}\Big|_{x=0}^{x=1/4} + \frac{6}{20} \cdot \frac{y^3}{3}\Big|_{y=0}^{y=1/4} = \frac{7}{640}$$

$$= .0109 \qquad \blacksquare$$

The marginal pmf of one discrete variable results from summing the joint pmf over all values of the *other* variable. Similarly, the marginal pdf of one continuous variable is obtained by integrating the joint pdf over all values of the other variable.

DEFINITION The **marginal probability density functions** of X and Y, denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad \text{for } -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx \quad \text{for } -\infty < y < \infty$$

Example 5.4 (Example 5.3 continued) The marginal pdf of X, which gives the probability distribution of busy time for the drive-up facility without reference to the walk-up window, is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy = \int_0^1 \frac{6}{5}(x+y^2)\, dy = \frac{6}{5}x + \frac{2}{5}$$

for $0 \le x \le 1$ and 0 otherwise. Similarly, the marginal pdf of Y is

$$f_Y(y) = \frac{6}{5}y^2 + \frac{3}{5} \quad 0 \le y \le 1$$

Then, for example,

$$P\left(\frac{1}{4} \le Y \le \frac{3}{4}\right) = \int_{1/4}^{3/4} \left(\frac{6}{5}y^2 + \frac{3}{5}\right)dy = \frac{37}{80} = .4625. \qquad \blacksquare$$

In Examples 5.3–5.4, the region of positive joint density was a rectangle, which made computation of the marginal pdfs relatively easy. Consider now an example in which the region of positive density is a more complicated figure.

Example 5.5   A nut company markets cans of deluxe mixed nuts containing almonds, cashews, and peanuts. Suppose the net weight of each can is exactly 1 lb, but the weight contribution of each type of nut is random. Because the three weights sum to 1, a joint probability model for any two gives all necessary information about the weight of the third type. Let X = the weight of almonds in a selected can and Y = the weight of cashews. Then the region of positive density is $D = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1, x + y \leq 1\}$, the shaded region pictured in Figure 5.2.
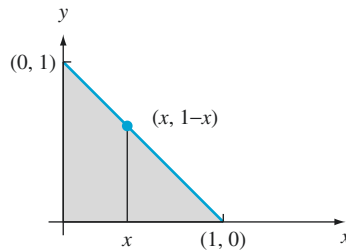
Figure 5.2   Region of positive density for Example 5.5

Now let the joint pdf for (X, Y) be

$$f(x, y) = 24xy \quad 0 \leq x \leq 1, \ \ 0 \leq y \leq 1, \ \ x + y \leq 1$$

For any fixed x, $f(x, y)$ increases with y; for fixed y, $f(x, y)$ increases with x. This is appropriate because the word *deluxe* implies that most of the can should consist of almonds and cashews rather than peanuts, so that the density function should be large near the upper boundary and small near the origin. The surface determined by $f(x, y)$ slopes upward from zero as (x, y) moves away from either axis.

Clearly, $f(x, y) \geq 0$. To verify the second condition on a joint pdf, recall that a double integral is computed as an iterated integral by holding one variable fixed (such as x as in Figure 5.2), integrating over values of the other variable lying along the straight line passing through the value of the fixed variable, and finally integrating over all possible values of the fixed variable. Thus

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dy dx = \iint_{D} f(x, y) dy dx = \int_{0}^{1} \left\{ \int_{0}^{1-x} 24xy \, dy \right\}$$

$$= \int_{0}^{1} 24x \left\{ \left. \frac{y^2}{2} \right|_{y=0}^{y=1-x} \right\} dx = \int_{0}^{1} 12x(1-x)^2 dx = 1$$

To compute the probability that the two types of nuts together make up at most 50% of the can, let $A = \{(x, y) : 0 \le x \le 1, 0 \le y \le 1, \text{and } x + y \le .5\}$, as shown in Figure 5.3. Then
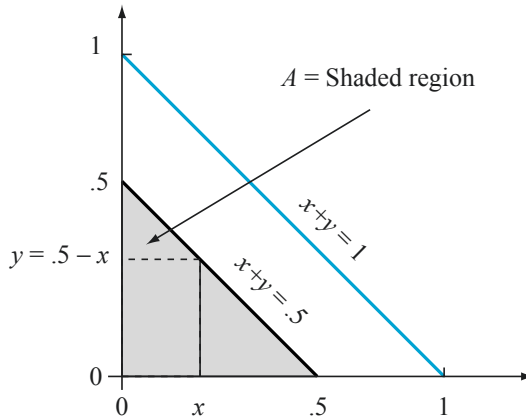
Figure 5.3   Computing $P((X, Y) \in A)$ for Example 5.5

$$P((X, Y) \in A) = \iint_A f(x, y)\, dx\, dy = \int_0^{.5} \int_0^{.5-x} 24xy\, dy\, dx = .0625$$

The marginal pdf for almonds is obtained by holding X fixed at x and integrating $f(x, y)$ along the vertical line through x:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^{1-x} 24xy\, dy = 12x(1 - x)^2 \quad 0 \le x \le 1$$

By symmetry of $f(x, y)$ and the region D, the marginal pdf of Y is obtained by replacing x and X in $f_X(x)$ by y and Y, respectively: $f_Y(y) = 12y(1 - y)^2$ for $0 \le y \le 1$.   ∎

Independent Random Variables

In many situations, information about the observed value of one of the two variables X and Y gives information about the value of the other variable. In Example 5.1, the marginal probability of X at $x = 250$ was .5, as was the probability that $X = 100$. If, however, we are told that the selected individual had $Y = 0$, then $X = 100$ is four times as likely as $X = 250$. Thus there is a dependence between the two variables.

In Chapter 2 we pointed out that one way of defining independence of two events is to say that A and B are independent if $P(A \cap B) = P(A) \cdot P(B)$. Here is an analogous definition for the independence of two rvs.

DEFINITION   Two random variables X and Y are said to be **independent** if for every pair of
                   x and y values,

$$p(x, y) = p_X(x) \cdot p_Y(y) \quad \text{when } X \text{ and } Y \text{ are discrete}$$

                   or   (5.1)

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad \text{when } X \text{ and } Y \text{ are continuous}$$

                   If (5.1) is not satisfied for all (x, y), then X and Y are said to be **dependent**.

The definition says that two variables are independent if their joint pmf or pdf is the product of the
two marginal pmfs or pdfs.

Example 5.6 In the insurance situation of Examples 5.1 and 5.2,

$$p(100, 100) = .10 \neq (.5)(.25) = p_X(100) \cdot p_Y(100)$$

so X and Y are not independent. Independence of X and Y requires that *every* entry in the joint
probability table be the product of the corresponding row and column marginal probabilities.   ∎

Example 5.7 (Example 5.5 continued) Because $f(x, y)$ in the nut scenario has the form of a product,
X and Y might appear to be independent. However, although $f_X\left(\frac{3}{4}\right) = f_Y\left(\frac{3}{4}\right) = \frac{9}{16}$, $f\left(\frac{3}{4}, \frac{3}{4}\right) = 0 \neq \frac{9}{16} \cdot \frac{9}{16}$
so the variables are not in fact independent. To be independent, $f(x, y)$ must have the form $g(x) \cdot h(y)$ *and*
the region of positive density must be a rectangle whose sides are parallel to the coordinate axes.   ∎

   Independence of two random variables is most useful when the description of the experiment
under study tells us that X and Y have no effect on each other. Then once the marginal pmfs or pdfs
have been specified, the joint pmf or pdf is simply the product of the two marginal functions. It
follows that

$$P(\{a \leq X \leq b\} \cap \{c \leq Y \leq d\}) = P(a \leq X \leq b) \cdot P(c \leq Y \leq d)$$

Example 5.8 Suppose that the lifetimes of two components are independent of each other and that
the first lifetime, X_1, has an exponential distribution with parameter λ_1 whereas the second, X_2, has an
exponential distribution with parameter λ_2. Then the joint pdf is

$$f(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) = \lambda_1 e^{-\lambda_1 x_1} \cdot \lambda_2 e^{-\lambda_2 x_2} = \lambda_1 \lambda_2 e^{-\lambda_1 x_1 - \lambda_2 x_2} \quad x_1 > 0, x_2 > 0$$

Let λ_1 = 1/1000 and λ_2 = 1/1200, so that the expected lifetimes are 1000 h and 1200 h, respectively.
The probability that both component lifetimes are at least 1500 h is

$$P(1500 \leq X_1, 1500 \leq X_2) = P(1500 \leq X_1) \cdot P(1500 \leq X_2) = e^{-\lambda_1(1500)} \cdot e^{-\lambda_2(1500)} = (.2231)(.2865) = .0639$$

The probability that the sum of their lifetimes, $X_1 + X_2$, is at most 3000 h requires a double integral of the joint pdf:

$$
\begin{aligned}
P(X_1 + X_2 \leq 3000) = P(X_1 \leq 3000 - X_2) &= \int\limits_0^{3000} \int\limits_0^{3000-x_2} f(x_1, x_2)\, dx_1\, dx_2 \\
&= \int\limits_0^{3000} \int\limits_0^{3000-x_2} \lambda_1 \lambda_2 e^{-\lambda_1 x_1 - \lambda_2 x_2}\, dx_1\, dx_2 \\
&= \int\limits_0^{3000} \lambda_2 e^{-\lambda_2 x_2} \left[-e^{-\lambda_1 x_1}\right]_0^{3000-x_2}\, dx_2 \\
&= \int\limits_0^{3000} \lambda_2 e^{-\lambda_2 x_2} \left[1 - e^{-\lambda_1(3000-x_2)}\right]\, dx_2 \\
&= \lambda_2 \int\limits_0^{3000} \left[e^{-\lambda_2 x_2} - e^{-3000\lambda_1} e^{(\lambda_1 - \lambda_2)x_2}\right]\, dx_2 = .7564 \qquad \blacksquare
\end{aligned}
$$

More than Two Random Variables
To model the joint behavior of more than two random variables, we extend the concept of a joint distribution of two variables.

DEFINITION   If $X_1, X_2, \ldots, X_n$ are all discrete random variables, the **joint pmf** of the variables is the function

$$p(x_1, x_2, \ldots, x_n) = P(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \cdots \cap \{X_n = x_n\})$$

If the variables are continuous, the **joint pdf** of $X_1, X_2, \ldots, X_n$ is the function $f(x_1, x_2, \ldots, x_n)$ such that for any n intervals $[a_1, b_1], \ldots, [a_n, b_n]$,

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int\limits_{a_1}^{b_1} \cdots \int\limits_{a_n}^{b_n} f(x_1, \ldots, x_n)\, dx_n \ldots dx_1$$

and more generally, for any n-dimensional set A, $P((X_1, \ldots, X_n) \in A)$ results from integrating $f(\ )$ over A.

Example 5.9 A binomial experiment consists of n dichotomous (success–failure), homogenous (constant success probability) independent trials. Now consider a *trinomial experiment* in which each of the n trials can result in one of *three* possible outcomes. For example, each successive customer at

a store might pay with cash, a credit card, or a debit card. The trials are assumed independent. Let $p_1 = P$(trial results in a type 1 outcome), and define p_2 and p_3 analogously for type 2 and type 3 outcomes. The random variables of interest here are X_i = the number of trials that result in a type i outcome for i = 1, 2, 3.

In n = 10 trials, the probability that the first five are type 1 outcomes, the next three are type 2, and the last two are type 3—i.e., the probability of the experimental outcome 1111122233—is $p_1^5 \cdot p_2^3 \cdot p_3^2$. This is also the probability of the outcome 1122311123, and in fact the probability of any outcome that has exactly five 1's, three 2's, and two 3's. Now to determine the probability $P(X_1 = 5, X_2 = 3,$ and $X_3 = 2)$, we have to count the number of outcomes that have exactly five 1's, three 2's, and two 3's. First, there are $\binom{10}{5}$ ways to choose five of the trials to be the type 1 outcomes. Now from the remaining five trials, we choose three to be the type 2 outcomes, which can be done in $\binom{5}{3}$ ways. This determines the remaining two trials which consist of type 3 outcomes. So the total number of ways of choosing five 1's, three 2's, and two 3's is

$$\binom{10}{5} \cdot \binom{5}{3} = \frac{10!}{5!5!} \cdot \frac{5!}{3!2!} = \frac{10!}{5!3!2!} = 2520$$

Thus we see that $P(X_1 = 5, X_2 = 3, X_3 = 2) = 2520\, p_1^5 \cdot p_2^3 \cdot p_3^2$. Generalizing this to n trials gives

$$p(x_1, x_2, x_3) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1! x_2! x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}$$

for x_1 = 0, 1, 2, …; x_2 = 0, 1, 2, …; x_3 = 0, 1, 2, … such that $x_1 + x_2 + x_3 = n$. Notice that whereas there are three random variables here, the third variable X_3 is actually redundant, because for example in the case n = 10, having $X_1 = 5$ and $X_2 = 3$ implies that $X_3 = 2$ (just as in a binomial experiment there are actually two rvs—the number of successes and number of failures—but the latter is redundant).

As an example, the genetic allele of a pea section can be either AA, Aa, or aa. A simple genetic model specifies $P(AA) = .25$, $P(Aa) = .50$, and $P(aa) = .25$. If the alleles of ten independently obtained sections are determined, the probability that exactly five of these are Aa and two are AA is

$$p(2, 5, 3) = \frac{10!}{2!5!3!}(.25)^2(.50)^5(.25)^3 = .0769 \qquad \blacksquare$$

The trinomial scenario of Example 5.9 can be generalized by considering a **multinomial experiment** consisting of n independent and identical trials, in which each trial can result in any one of r possible outcomes. Let $p_i = P$(outcome i on any particular trial), and define random variables by X_i = the number of trials resulting in outcome i (i = 1, . . ., r). The joint pmf of X_1, . . ., X_r is called the **multinomial distribution**. An argument analogous to what was done in Example 5.9 gives the joint pmf of X_1, . . ., X_r:

$$p(x_1, \ldots, x_r) = \frac{n!}{x_1! x_2! \cdots x_r!} p_1^{x_1} \cdot \cdots \cdot p_r^{x_r} \quad \text{for } x_i = 0, 1, 2, \ldots \quad \text{with } x_1 + \cdots + x_r = n$$

The case $r = 2$ reduces to the binomial distribution, with X_1 = number of successes and $X_2 = n - X_1$ = number of failures. Both the multinomial and binomial distributions model discrete rvs (counts). Now, let's consider some examples with more than two continuous random variables.

Example 5.10 When a certain method is used to collect a fixed volume of rock samples in a region, there are four resulting rock types. Let X_1, X_2, and X_3 denote the proportion by volume of rock types 1, 2, and 3 in a randomly selected sample (the proportion of rock type 4 is $1 - X_1 - X_2 - X_3$, so a variable X_4 would be redundant). If the joint pdf of X_1, X_2, X_3 is

$$f(x_1, x_2, x_3) = kx_1x_2(1 - x_3) \quad 0 \le x_1 \le 1, 0 \le x_2 \le 1, 0 \le x_3 \le 1, x_1 + x_2 + x_3 \le 1$$

then k is determined by

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, x_3) \ dx_3 dx_2 dx_1$$

$$= \int_0^1 \left\{ \int_0^{1-x_1} \left[\int_0^{1-x_1-x_2} kx_1x_2(1-x_3)dx_3 \right] dx_2 \right\} dx_1$$

This iterated integral has value $k/144$, so $k = 144$. The probability that rocks of types 1 and 2 together account for at most 50% of the sample is

$$P(X_1 + X_2 \le .5) = \iiint_{\left\{ \begin{array}{c} 0 \le x_i \le 1 \text{ for } i = 1, 2, 3 \\ x_1 + x_2 + x_3 \le 1, \ x_1 + x_2 \le .5 \end{array} \right\}} f(x_1, x_2, x_3) dx_3 dx_2 dx_1$$

$$= \int_0^{.5} \left\{ \int_0^{.5-x_1} \left[\int_0^{1-x_1-x_2} 144x_1x_2(1-x_3)dx_3 \right] dx_2 \right\} dx_1 = \frac{5}{32} \qquad \blacksquare$$

The notion of independence of more than two random variables is similar to the notion of independence of more than two events.

DEFINITION   The random variables $X_1, X_2, \ldots, X_n$ are said to be **independent** if for *every* subset $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ of the variables (each pair, each triple, and so on), the joint pmf or pdf of the subset is equal to the product of the marginal pmfs or pdfs.

Thus if the variables are independent with $n = 4$, then the joint pmf or pdf of any two variables is the product of the two marginals, and similarly for any three variables and all four variables together. Most important, once we are told that n variables are independent, then the joint pmf or pdf is the product of the n marginals.

Example 5.11 If X_1, ..., X_n represent the lifetimes of n components, the components operate independently of each other, and each lifetime is exponentially distributed with parameter λ, then

$$f(x_1, x_2, \ldots, x_n) = \left(\lambda e^{-\lambda x_1}\right) \cdot \left(\lambda e^{-\lambda x_2}\right) \cdot \cdots \cdot \left(\lambda e^{-\lambda x_n}\right) = \lambda^n e^{-\lambda \Sigma x_i} \quad x_1 \geq 0, \ \ x_2 \geq 0, \ldots, x_n \geq 0$$

If these n components are connected in series, so that the system will fail as soon as a single component fails, then the probability that the system lasts past time t is

$$P(\{X_1 > t\} \cap \ldots \cap \{X_n > t\}) = \int_t^\infty \ldots \int_t^\infty f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$$

$$= \left(\int_t^\infty \lambda e^{-\lambda x_1}\, dx_1\right) \cdot \cdots \cdot \left(\int_t^\infty \lambda e^{-\lambda x_n}\, dx_n\right)$$

$$= (e^{-\lambda t})^n = e^{-n\lambda t}$$

Therefore,

$$P(\text{system lifetime} \leq t) = 1 - e^{-n\lambda t} \quad \text{for } t \geq 0$$

which shows that *system* lifetime has an exponential distribution with parameter $n\lambda$; the expected value of system lifetime is $1/(n\lambda)$.

A variation on the foregoing scenario appeared in the article "A Method for Correlating Field Life Degradation with Reliability Prediction for Electronic Modules" (*Quality and Reliability Engr. Intl.* 2005: 715–726). The investigators considered a circuit card with n soldered chip resistors. The failure time of a card is the minimum of the individual solder connection failure times (mileages here). It was assumed that the solder connection failure mileages were independent, that failure mileage would exceed t if and only if the shear strength of a connection exceeded a threshold d, and that each shear strength was normally distributed with a mean value and standard deviation that depended on the value of mileage t: $\mu(t) = a_1 - a_2 t$ and $\sigma(t) = a_3 + a_4 t$ (a weld's shear strength typically deteriorates and becomes more variable as mileage increases). Then the probability that the failure mileage of a card exceeds t is

$$P(T > t) = \left(1 - \Phi\left(\frac{d - (a_1 - a_2 t)}{a_3 + a_4 t}\right)\right)^n$$

The cited article suggested values for d and the a_i's based on data. In contrast to the exponential scenario, normality of individual lifetimes does not imply normality of system lifetime. ∎

In many experimental situations to be considered in this book, independence is a reasonable assumption, so that specifying the joint distribution reduces to deciding on appropriate marginal distributions.

Exercises: Section 5.1 (1–22)

1. A service station has both self-service and full-service islands. On each island, there is a single regular unleaded pump with two hoses. Let X denote the number of hoses being used on the self-service island at a particular time, and let Y denote the number of hoses on the full-service island in use at that time. The joint pmf of X and Y appears in the accompanying tabulation.

	$p(x, y)$	0	1	2
			y	
x	0	.10	.04	.02
	1	.08	.20	.06
	2	.06	.14	.30

   a. What is $P(X = 1 \text{ and } Y = 1)$?
   b. Compute $P(X \le 1 \text{ and } Y \le 1)$.
   c. Give a word description of the event $\{X \ne 0 \text{ and } Y \ne 0\}$, and compute the probability of this event.
   d. Compute the marginal pmf of X and of Y. Using $p_X(x)$, what is $P(X \le 1)$?
   e. Are X and Y independent rvs? Explain.

2. A large but sparsely populated county has two small hospitals, one at the south end of the county and the other at the north end. The south hospital's emergency room has 4 beds, whereas the north hospital's emergency room has only 3 beds. Let X denote the number of south beds occupied at a particular time on a given day, and let Y denote the number of north beds occupied at the same time on the same day. Suppose that these two rvs are independent, that the pmf of X puts probability masses .1, .2, .3, .2, and .2 on the x values 0, 1, 2, 3, and 4, respectively, and that the pmf of Y distributes probabilities .1, .3, .4, and .2 on the y values 0, 1, 2, and 3, respectively.

   a. Display the joint pmf of X and Y in a joint probability table.

   b. Compute $P(X \le 1 \text{ and } Y \le 1)$ by adding probabilities from the joint pmf, and verify that this equals the product of $P(X \le 1)$ and $P(Y \le 1)$.
   c. Express the event that the total number of beds occupied at the two hospitals combined is at most 1 in terms of X and Y, and then calculate this probability.
   d. What is the probability that at least one of the two hospitals has no beds occupied?

3. A market has both an express checkout line and a superexpress checkout line. Let X_1 denote the number of customers in line at the express checkout at a particular time of day, and let X_2 denote the number of customers in line at the superexpress checkout at the same time. Suppose the joint pmf of X_1 and X_2 is as given in the accompanying table.

				x_2	
		0	1	2	3
	0	.08	.07	.04	.00
	1	.06	.15	.05	.04
x_1	2	.05	.04	.10	.06
	3	.00	.03	.04	.07
	4	.00	.01	.05	.06

   a. What is $P(X_1 = 1, X_2 = 1)$, that is, the probability that there is exactly one customer in each line?
   b. What is $P(X_1 = X_2)$, that is, the probability that the numbers of customers in the two lines are identical?
   c. Let A denote the event that there are at least two more customers in one line than in the other line. Express A in terms of X_1 and X_2, and calculate the probability of this event.
   d. What is the probability that the total number of customers in the two lines is exactly four? At least four?
   e. Determine the marginal pmf of X_1, and then calculate the expected number of customers in line at the express checkout.
   f. Determine the marginal pmf of X_2.

g. By inspection of $P(X_1 = 4), P(X_2 = 0)$, and $P(X_1 = 4, X_2 = 0)$, are X_1 and X_2 independent random variables? Explain your reasoning.

4. Suppose 51% of the individuals in a certain population have brown eyes, 32% have blue eyes, and the remainder have green eyes. Consider a random sample of 10 people from this population.

   a. What is the probability that 5 of the 10 people have brown eyes, 3 of 10 have blue eyes, and the other 2 have green eyes?
   b. What is the probability that exactly one person in the sample has blue eyes and exactly one has green eyes?
   c. What is the probability that at least 7 of the 10 people have brown eyes? [*Hint:* Think of brown as a success and all other eye colors as failures.]

5. At a certain university, 20% of all students are freshmen, 18% are sophomores, 21% are juniors, and 41% are seniors. As part of a promotion, the university bookstore is running a raffle for which all students are eligible. Ten students will be randomly selected to receive prizes (in the form of textbooks for the term).

   a. What is the probability the winners consist of two freshmen, two sophomores, two juniors, and four seniors?
   b. What is the probability the winners are split equally among under-classmen (freshmen and sophomores) and upper-classmen (juniors and seniors)?
   c. The raffle resulted in no freshmen being selected. The freshman class president complained that something must be amiss for this to occur. Do you agree? Explain.

6. According to the Mars Candy Company, the long-run percentages of various colors of M&M milk chocolate candies are as follows:

Blue:	Orange:	Green:	Yellow:	Red:	Brown:
24%	20%	16%	14%	13%	13%

a. In a random sample of 12 candies, what is the probability that there are exactly two of each color?
b. In a random sample of 6 candies, what is the probability that at least one color is not included?
c. In a random sample of 10 candies, what is the probability that there are exactly 3 blue candies and exactly 2 orange candies?
d. In a random sample of 10 candies, what is the probability that there are at most 3 orange candies? [*Hint:* Think of an orange candy as a success and any other color as a failure.]
e. In a random sample of 10 candies, what is the probability that at least 7 are either blue, orange, or green?

7. The number of customers waiting for gift-wrap service at a department store is an rv X with possible values 0, 1, 2, 3, 4 and corresponding probabilities .1, .2, .3, .25, .15. A randomly selected customer will have 1, 2, or 3 packages for wrapping with probabilities .6, .3, and .1, respectively. Let Y = the total number of packages to be wrapped for the customers waiting in line (assume that the number of packages submitted by one customer is independent of the number submitted by any other customer).

   a. Determine $P(X = 3, \ Y = 3)$, that is, $p(3, 3)$.
   b. Determine $p(4, 11)$.

8. Let X denote the number of Sony 65" 4 K Ultra HD televisions sold during a particular week by a certain store. The pmf of X is

x	0	1	2	3	4
$p_X(x)$	.1	.2	.3	.25	.15

Sixty percent of all customers who purchase these TVs also buy an extended warranty. Let Y denote the number of purchasers during this week who buy an extended warranty.

   a. What is $P(X = 4, \ Y = 2)$? [*Hint:* This probability is $P(Y = 2|X = 4) \cdot P(X = 4)$;

now think of the four purchases as four trials of a binomial experiment, with success on a trial corresponding to buying an extended warranty.]

b. Calculate $P(X = Y)$.

c. Determine the joint pmf of X and Y and then the marginal pmf of Y.

9. The joint probability distribution of the number X of cars and the number Y of buses per signal cycle at a proposed left-turn lane is displayed in the accompanying joint probability table.

		y		
	$p(x, y)$	0	1	2
	0	.025	.015	.010
	1	.050	.030	.020
x	2	.125	.075	.050
	3	.150	.090	.060
	4	.100	.060	.040
	5	.050	.030	.020

a. What is the probability that there is exactly one car and exactly one bus during a cycle?

b. What is the probability that there is at most one car and at most one bus during a cycle?

c. What is the probability that there is exactly one car during a cycle? Exactly one bus?

d. Suppose the left-turn lane is to have a capacity of five cars, and one bus is equivalent to three cars. What is the probability of an overflow during a cycle?

e. Are X and Y independent rvs? Explain.

10. A stockroom currently has 30 components of a certain type, of which 8 were provided by supplier 1, 10 by supplier 2, and 12 by supplier 3. Six of these are to be randomly selected for a particular assembly. Let X = the number of supplier 1's components selected, Y = the number of supplier 2's components selected, and $p(x, y)$ denote the joint pmf of X and Y.

a. What is $p(3, 2)$? [*Hint*: Each sample of size 6 is equally likely to be selected.

Therefore, $p(3, 2)$ = (number of outcomes with $X = 3$ and $Y = 2$)/(total number of outcomes). Now use the product rule for counting to obtain the numerator and denominator.]

b. Using the logic of part (a), obtain $p(x, y)$. (This can be thought of as a multivariate hypergeometric distribution— sampling without replacement from a finite population consisting of more than two categories.)

11. Each front tire of a vehicle is supposed to be filled to a pressure of 26 psi. Suppose the actual air pressure in each tire is a random variable—X for the right tire and Y for the left tire, with joint pdf

$$f(x, y) = k(x^2 + y^2) \quad 20 \le x \le 30, \quad 20 \le y \le 30$$

a. What is the value of k?

b. What is the probability that both tires are under-filled?

c. What is the probability that the difference in air pressure between the two tires is at most 2 psi?

d. Determine the (marginal) distribution of air pressure in the right tire alone.

e. Are X and Y independent rvs?

12. Annie and Alvie have agreed to meet between 5:00 p.m. and 6:00 p.m. for dinner at a local health-food restaurant. Let X = Annie's arrival time and Y = Alvie's arrival time. Suppose X and Y are independent with each uniformly distributed on the interval [5, 6].

a. What is the joint pdf of X and Y?

b. What is the probability that they both arrive between 5:15 and 5:45?

c. If the first one to arrive will wait only 10 min before leaving to eat elsewhere, what is the probability that they have dinner at the health-food restaurant? [*Hint*: The event of interest is $A = \left\{ (x, y) : |x - y| \le \frac{1}{6} \right\}$.]

13. Two different professors have just submitted final exams for duplication. Let X denote the number of typographical errors on the first professor's exam and Y denote the number of such errors on the second exam. Suppose X has a Poisson distribution with parameter μ_1, Y has a Poisson distribution with parameter μ_2, and X and Y are independent.

    a. What is the joint pmf of X and Y?

    b. What is the probability that at most one error is made on both exams combined?

    c. Obtain a general expression for the probability that the total number of errors in the two exams is m (where m is a nonnegative integer). [*Hint*: $A = \{(x, y) : x + y = m\} = \{(m, 0), (m - 1, 1), \ldots, (1, m - 1), (0, m)\}$. Now sum the joint pmf over $(x, y) \in A$ and use the binomial theorem, which says that $\sum_{k=0}^{m} \binom{m}{k} a^k b^{m-k} = (a + b)^m$ for any a, b.]

14. Two components of a computer have the following joint pdf for their useful lifetimes X and Y:

$$f(x, y) = xe^{-x(1+y)} \quad x \geq 0, \ y \geq 0$$

    a. What is the probability that the lifetime X of the first component exceeds 3?

    b. What are the marginal pdfs of X and Y? Are the two lifetimes independent? Explain.

    c. What is the probability that the lifetime of at least one component exceeds 3?

15. You have two lightbulbs for a particular lamp. Let $X =$ the lifetime of the first bulb and $Y =$ the lifetime of the second bulb (both in thousands of hours). Suppose that X and Y are independent and that each has an exponential distribution with parameter $\lambda = 1$.

    a. What is the joint pdf of X and Y?

    b. What is the probability that each bulb lasts at most 1000 h (i.e., $X \leq 1$ and $Y \leq 1$)?

    c. What is the probability that the total lifetime of the two bulbs is at most 2? [*Hint*: Draw a picture of the region $A = \{(x, y): x \geq 0, y \geq 0, x + y \leq 2\}$ before integrating.]

    d. What is the probability that the total lifetime is between 1 and 2?

16. Suppose that you have ten lightbulbs, that the lifetime of each is independent of all the other lifetimes, and that each lifetime has an exponential distribution with parameter λ.

    a. What is the probability that all ten bulbs fail before time t?

    b. What is the probability that exactly k of the ten bulbs fail before time t?

    c. Suppose that nine of the bulbs have lifetimes that are exponentially distributed with parameter λ and that the remaining bulb has a lifetime that is exponentially distributed with parameter θ (it is made by another manufacturer). What is the probability that exactly five of the ten bulbs fail before time t?

17. Consider a system consisting of three components as pictured. The system will continue to function as long as the first component functions and either component 2 or component 3 functions. Let X_1, X_2, and X_3 denote the lifetimes of components 1, 2, and 3, respectively. Suppose the X_i's are independent of each other and each X_i has an exponential distribution with parameter λ.
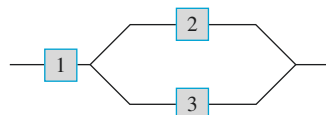
a. Let Y denote the system lifetime. Obtain the cumulative distribution function of Y and differentiate to obtain the pdf. [*Hint*: $F(y) = P(Y \le y)$; express the event $\{Y \le y\}$ in terms of unions and/or intersections of the three events $\{X_1 \le y\}, \{X_2 \le y\}$, and $\{X_3 \le y\}$.]

b. Compute the expected system lifetime.

18. a. For $f(x_1, x_2, x_3)$ as given in Example 5.10, compute the **joint marginal density function** of X_1 and X_3 alone (by integrating over x_2).

b. What is the probability that rocks of types 1 and 3 together make up at most 50% of the sample? [*Hint*: Use the result of part (a).]

c. Compute the marginal pdf of X_1 alone. [*Hint*: Use the result of part (a).]

19. An ecologist selects a point inside a circular sampling region according to a uniform distribution. Let $X =$ the x coordinate of the point selected and $Y =$ the y coordinate of the point selected. If the circle is centered at $(0, 0)$ and has radius r, then the joint pdf of X and Y is

$$f(x, y) = \frac{1}{\pi r^2} \quad x^2 + y^2 \le r^2$$

a. What is the probability that the selected point is within $r/2$ of the center of the circular region? [*Hint*: Draw a picture of the region of positive density D. Because $f(x, y)$ is constant on D, computing a probability reduces to computing an area.]

b. What is the probability that both X and Y differ from 0 by at most $r/2$?

c. Answer part (b) for $r/\sqrt{2}$ replacing $r/2$.

d. What is the marginal pdf of X? Of Y? Are X and Y independent?

20. Each customer making a particular Internet purchase must pay with one of three types of credit cards (think Visa, MasterCard, Amex). Let A_i ($i = 1, 2, 3$) be the event that a type i credit card is used, with $P(A_1) = .5$, $P(A_2) = .3$, $P(A_3) = .2$. Suppose that the number of customers who make a purchase on a given day, N, is a Poisson rv with parameter μ. Define rvs X_1, X_2, X_3 by $X_i =$ the number among the N customers who use a type i card ($i = 1, 2, 3$). Show that these three rvs are independent with Poisson distributions having parameters $.5\mu$, $.3\mu$, and $.2\mu$, respectively. [*Hint*: For nonnegative integers x_1, x_2, x_3, let $n = x_1 + x_2 + x_3$, so $P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3, N = n)$. Now condition on $N = n$, in which case the three X_i's have a trinomial distribution (multinomial with 3 categories) with category probabilities .5, .3, and .2.]

21. Consider randomly selecting two points A and B on the circumference of a circle by selecting their angles of rotation, in degrees, independently from a uniform distribution on the interval [0, 360]. Connect points A and B with a straight line segment. What is the probability that this random chord is longer than the side of an equilateral triangle inscribed inside the circle? [*Hint:* Place one of the vertices of the inscribed triangle at A. You should then be able to intuit the answer visually without having to do any integration.]
(This is called *Bertrand's Chord Problem* in the probability literature. There are other ways of randomly selecting a chord that give different answers from the one appropriate here.)

22. Consider the following technique, called the *accept–reject method*, for simulating values from a continuous distribution f. Identify a distribution g from which values can already be simulated and a constant $c \ge 1$ such that $f(x) \le cg(x)$ for all x. Proceed as follows: (1) Generate $Y \sim g$ and, independently, $U \sim$ Unif[0, 1]. (2) If $u \le f(y)/cg(y)$, then let $x = y$ (i.e., "accept" the y value); otherwise, discard ("reject") y. (3) Repeat steps (1)–(2) until the desired number of x values is obtained.

a. Show that the probability a y value is "accepted" equals $1/c$. [*Hint:* According to the algorithm, this occurs iff $U \leq f(Y)/cg(Y)$. Compute the relevant double integral.]

b. Argue that the average number of y values required to generate a single accepted x value is c.

c. Show that the accept–reject method does result in observations from f by showing that $P(\text{accepted value} \leq x) = F(x)$, where F is the cdf corresponding to f. [*Hint:* Let X denote the accepted value. Then $P(X \leq x) = P(Y \leq x \mid Y$ is accepted$) = P(Y \leq x \cap Y$ is acc.$)/P(Y$ is acc.$).$]

5.2 Expected Values, Covariance, and Correlation

We previously saw that any function $h(X)$ of a single rv X is itself a random variable. However, to compute $E[h(X)]$, it was not necessary to obtain the probability distribution of $h(X)$; instead, $E[h(X)]$ was computed as a weighted average of $h(X)$ values, where the weight function was the pmf $p(x)$ or pdf $f(x)$ of X. A similar result holds for a function $h(X, Y)$ of two jointly distributed random variables.

LAW OF THE UNCONSCIOUS STATISTICIAN	Let X and Y be jointly distributed rvs with pmf $p(x, y)$ or pdf $f(x, y)$ according to whether the variables are discrete or continuous. Then the expected value of a function $h(X, Y)$, denoted by $E[h(X, Y)]$ or $\mu_{h(X,Y)}$, is given by

$$E[h(X, Y)] = \begin{cases} \sum_x \sum_y h(x,y) \cdot p(x,y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) \cdot f(x,y) dx\, dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases} \quad (5.2)$$

The Law of the Unconscious Statistician generalizes to computing the expected value of a function $h(X_1, \ldots, X_n)$ of n random variables. If the X_i's are discrete, $E[h(X_1, \ldots, X_n)]$ is an n-dimensional sum; if the X_i's are continuous, it is an n-dimensional integral.

Example 5.12 Five friends have purchased tickets to a concert. If the tickets are for seats 1–5 in a particular row and the tickets are randomly distributed among the five, what is the expected number of seats separating any particular two of the five? Let X and Y denote the seat numbers of the first and second individuals, respectively. Possible (X, Y) pairs are $\{(1, 2), (1, 3), \ldots, (5, 4)\}$, from which

$$p(x, y) = .05 \quad x = 1, \ldots, 5; \quad y = 1, \ldots, 5; \quad x \neq y$$

The number of seats separating the two individuals is $h(X, Y) = |X - Y| - 1$. The accompanying table gives $h(x, y)$ for each possible (x, y) pair.

	$h(x, y)$	1	2	3	4	5
	1	–	0	1	2	3
	2	0	–	0	1	2
y	3	1	0	–	0	1
	4	2	1	0	–	0
	5	3	2	1	0	–

Thus

$$E[h(X,Y)] = \sum_{(x,y)} \sum h(x,y) \cdot p(x,y) = \sum_{\substack{x=1 \\ x \neq y}}^{5} \sum_{y=1}^{5} (|x-y|-1) \cdot \frac{1}{20} = 1 \qquad \blacksquare$$

Example 5.13 In Example 5.5, the joint pdf of the amount X of almonds and amount Y of cashews in a 1-lb can of nuts was

$$f(x,y) = 24xy \qquad 0 \leq x \leq 1, \ 0 \leq y \leq 1, \ x+y \leq 1$$

If 1 lb of almonds costs the company \$6.00, 1 lb of cashews costs \$10.00, and 1 lb of peanuts costs \$3.50, then the total cost of the contents of a can is

$$h(X,Y) = 6X + 10Y + 3.5(1-X-Y) = 3.5 + 2.5X + 6.5Y$$

(since $1-X-Y$ of the weight consists of peanuts). The expected total cost is

$$\begin{aligned} E[h(X,Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) \cdot f(x,y) \, dx \, dy \\ &= \int_{0}^{1} \int_{0}^{1-x} (3.5 + 2.5x + 6.5y) \cdot 24xy \, dy \, dx = \$7.10 \qquad \blacksquare \end{aligned}$$

Properties of Expected Value

In Chapters 3 and 4, we saw that expected values can be distributed across addition, subtraction, and multiplication by constants. In the language of mathematics, expected value is a *linear operator*. This was a simple consequence of expectation being a sum or an integral, both of which are linear. This obvious but important property, linearity of expectation, extends to more than one variable.

LINEARITY OF EXPECTATION Let X and Y be random variables. Then, for any functions h_1, h_2 and any constants a_1, a_2, b,

$$E[a_1 h_1(X,Y) + a_2 h_2(X,Y) + b] = a_1 E[h_1(X,Y)] + a_2 E[h_2(X,Y)] + b$$

In the previous example, $E(3.5 + 2.5X + 6.5Y)$ can be rewritten as $3.5 + 2.5E(X) + 6.5E(Y)$; the means of X and Y can be computed either by using (5.2) or by first finding the marginal pdfs of X and Y and then performing the appropriate single integrals.

As another illustration, linearity of expectation tells us that for any two rvs X and Y,

$$E(5XY^2 - 4XY + e^X + 12) = 5E(XY^2) - 4E(XY) + E(e^X) + 12 \qquad (5.3)$$

In general, we cannot distribute the expected value operation any further. But when $h(X, Y)$ is a product of a function of X and a function of Y, the expected value simplifies in the case of independence.

PROPOSITION   Let X and Y be *independent* random variables. If $h(X, Y) = g_1(X) \cdot g_2(Y)$, then

$$E[h(X, Y)] = E[g_1(X) \cdot g_2(Y)] = E[g_1(X)] \cdot E[g_2(Y)],$$

assuming $E[g_1(X)]$ and $E[g_2(Y)]$ exist.

Proof Consider two continuous rvs; the discrete case is similar. Apply (5.2):

$$E[h(X, Y)] = E[g_1(X) \cdot g_2(Y)] = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} g_1(x) \cdot g_2(y) \cdot f(x, y) dx\, dy \qquad \text{by (5.2)}$$

$$= \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} g_1(x) \cdot g_2(y) \cdot f_X(x) \cdot f_Y(y) dx\, dy \qquad \text{because } X \text{ and } Y \text{ are independent}$$

$$= \left(\int\limits_{-\infty}^{\infty} g_1(x) \cdot f_X(x) dx \right) \left(\int\limits_{-\infty}^{\infty} g_2(y) \cdot f_Y(y) dy \right) = E[g_1(X)]E[g_2(Y)] \qquad \blacksquare$$

So, if X and Y are independent, Expression (5.3) simplifies further, to $5E(X)E(Y^2) - 4E(X)E(Y) + E(e^X) + 12$. Not surprisingly, both linearity of expectation and the foregoing proposition can be extended to more than two random variables.

Covariance

When two random variables X and Y are not independent, it is frequently of interest to assess how strongly they are related to each other.

DEFINITION The **covariance** between two rvs X and Y is

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$= \begin{cases} \displaystyle\sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\[2mm] \displaystyle\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)f(x, y) dx\, dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

The rationale for the definition is as follows. Suppose X and Y have a strong positive relationship to each other, by which we mean that large values of X tend to occur with large values of Y and small values of X with small values of Y (e.g., X = height and Y = weight). Then most of the probability mass or density will be associated with $(x - \mu_X)$ and $(y - \mu_Y)$ either both positive (both X and Y above

their respective means) or both negative, so the product $(x - \mu_X)(y - \mu_Y)$ will tend to be positive. Thus for a strong positive relationship, Cov(X, Y) should be quite positive. For a strong negative relationship, the signs of $(x - \mu_X)$ and $(y - \mu_Y)$ will tend to be opposite, yielding a negative product. Thus for a strong negative relationship, Cov(X, Y) should be quite negative. If X and Y are not strongly related, positive and negative products will tend to cancel each other, yielding a covariance near 0. Figure 5.4 illustrates the different possibilities. The covariance depends on *both* the set of possible pairs and the probabilities. In Figure 5.4, the probabilities could be changed without altering the set of possible pairs, and this could drastically change the value of Cov(X, Y).
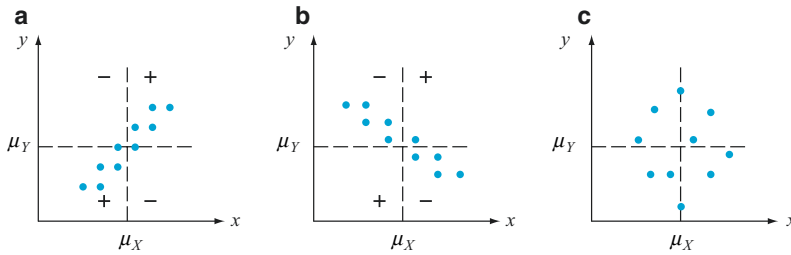
Figure 5.4 $p(x, y) = \frac{1}{10}$ for each of ten pairs corresponding to indicated points; (**a**) positive covariance; (**b**) negative covariance; (**c**) covariance near zero

Example 5.14 The joint and marginal pmfs for X = automobile policy deductible amount and Y = homeowner policy deductible amount in Example 5.1 were

p(x, y)		y			x		100	250		y		0	100	200
		0	100	200										
x	100	.20	.10	.20	$p_X(x)$		.5	.5		$p_Y(y)$		.25	.25	.50
	250	.05	.15	.30										

from which $\mu_X = \sum x \cdot p_X(x) = 175$ and $\mu_Y = 125$. Therefore,

$$\text{Cov}(X, Y) = \sum\sum_{(x, y)} (x - 175)(y - 125)p(x, y)$$

$$= (100 - 175)(0 - 125)(.20) + \cdots + (250 - 175)(200 - 125)(.30)$$

$$= 1875 \qquad\qquad\blacksquare$$

The following proposition summarizes some important properties of covariance.

PROPOSITION   For any two random variables X and Y,

1. Cov(X, Y) = Cov(Y, X)
2. Cov(X, X) = V(X)
3. (Covariance shortcut formula) Cov(X, Y) = $E(XY) - \mu_X \cdot \mu_Y$
4. (Distributive property of covariance) For any rv Z and any constants, a, b, c,
$$\text{Cov}(aX + bY + c, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$$

Proof Property 1 is obvious from the definition of covariance. To establish Property 2, replace Y with X in the definition:

$$\text{Cov}(X, X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = V(X)$$

To prove Property 3, apply linearity of expectation:

$$\begin{aligned}
\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\
&= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y) \\
&= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y \\
&= E(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y = E(XY) - \mu_X \mu_Y
\end{aligned}$$

Property 4 also follows from linearity of expectation (Exercise 39).   ∎

According to Property 3 (the covariance shortcut), no intermediate subtractions are necessary to calculate covariance; only at the end of the computation is $\mu_X \cdot \mu_Y$ subtracted from $E(XY)$.

Example 5.15 (Example 5.5 continued) The joint and marginal pdfs of $X =$ *amount of almonds* and $Y =$ *amount of cashews* were

$$f(x, y) = 24xy \qquad 0 \le x \le 1, \ \ 0 \le y \le 1, \ \ x + y \le 1$$

$$f_X(x) = 12x(1 - x)^2 \qquad 0 \le x \le 1 \quad f_Y(y) = 12y(1 - y)^2 \qquad 0 \le y \le 1$$

It is easily verified that $\mu_X = \mu_Y = \dfrac{2}{5}$, and

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y) dx\, dy = \int_0^1 \int_0^{1-x} xy \cdot 24xy \, dy\, dx$$

$$= 8 \int_0^1 x^2 (1 - x)^3 dx = \frac{2}{15}$$

Thus $\text{Cov}(X, Y) = \frac{2}{15} - \left(\frac{2}{5}\right)\left(\frac{2}{5}\right) = \frac{2}{15} - \frac{4}{25} = -\frac{2}{75}$. A negative covariance is reasonable here because more almonds in the can imply fewer cashews.   ∎

Correlation

It would appear that the relationship in the insurance example is quite strong since $\text{Cov}(X, Y) = 1875$, whereas in the nut example $\text{Cov}(X, Y) = -2/75$ would seem to imply quite a weak relationship. Unfortunately, the covariance has a serious defect that makes it impossible to interpret a computed value of the covariance. In the insurance example, suppose we had expressed the deductible amount in cents rather than in dollars. Then $100X$ would replace X, $100Y$ would replace Y, and the resulting covariance would be $\text{Cov}(100X, 100Y) = (100)(100)\text{Cov}(X, Y) = 18{,}750{,}000$. [To see why, apply properties 1 and 4 of the previous proposition.] If, on the other hand, the deductible

amount had been expressed in hundreds of dollars, the computed covariance would have changed to (.01) (.01) (1875) = .1875. *The defect of covariance is that its computed value depends critically on the units of measurement.* Ideally, the choice of units should have no effect on a measure of strength of relationship. This is achieved by scaling the covariance.

DEFINITION The **correlation coefficient** of X and Y, denoted by Corr(X, Y), or $\rho_{X,Y}$, or just ρ, is defined by

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Example 5.16 It is easily verified that in the insurance scenario of Example 5.14, $E(X^2) = 36{,}250, \sigma_X^2 = 36{,}250 - (175)^2 = 5625, \sigma_X = 75, E(Y^2) = 22{,}500, \sigma_Y^2 = 6875$, and $\sigma_Y = 82.92$. This gives

$$\rho = \frac{1875}{(75)(82.92)} = .301$$ ∎

The following proposition shows that ρ remedies the defect of Cov(X, Y) and also suggests how to recognize the existence of a strong (linear) relationship.

PROPOSITION For any two rvs X and Y,

1. Corr(X, Y) = Corr(Y, X)
2. Corr(X, X) = 1
3. (Scale invariance property) If a, b, c, d are constants and $ac > 0$,
   Corr($aX + b, cY + d$) = Corr(X, Y)
4. $-1 \leq$ Corr(X, Y) ≤ 1.

Proof Property 1 is clear from the definition of correlation and the corresponding property of covariance. To see why Property 2 is true, write Corr(X, X) = Cov(X, X)/$[\sigma_X \cdot \sigma_X] = V(X)/\sigma_X^2 = 1$. The second-to-last step uses Property 2 of covariance. The proofs of Properties 3 and 4 appear as exercises. ∎

Property 3 (scale invariance) says precisely that the correlation coefficient is not affected by a linear change in the units of measurement. If, say, Y = completion time for a chemical reaction in seconds and X = temperature in °C, then $Y/60$ = time in minutes and $1.8X + 32$ = temperature in °F, but Corr(X, Y) will be exactly the same as Corr($1.8X + 32, Y/60$).

According to Properties 2 and 4, the strongest possible positive relationship is evidenced by $\rho = +1$ whereas the strongest possible negative relationship corresponds to $\rho = -1$. Therefore, the correlation coefficient provides information about both the nature and strength of the relationship between X and Y: The sign of ρ indicates whether X and Y are positively or negatively related, and the magnitude of ρ describes the strength of that relationship on an absolute 0 to 1 scale.

If we think of $p(x, y)$ or $f(x, y)$ as prescribing a mathematical model for how the two numerical variables X and Y are distributed in some population (height and weight, verbal SAT score and quantitative SAT score, etc.), then ρ is a population characteristic or parameter that measures how

strongly X and Y are related in the population. In Chapter 12, we will consider taking a sample of pairs $(x_1, y_1), \ldots, (x_n, y_n)$ from the population. The sample correlation coefficient r will then be defined and used to make inferences about ρ.

While superior to covariance, the correlation coefficient ρ is actually not a completely general measure of the strength of a relationship.

PROPOSITION   1. If X and Y are independent, then $\rho = 0$, but $\rho = 0$ does not imply independence.
2. $\rho = 1$ or -1 iff $Y = aX + b$ for some numbers a and b with $a \neq 0$.

Exercise 38 and Example 5.17 relate to Statement 1, and Statement 2 is investigated in Exercises 41 and 42(d).

This proposition says that ρ is a measure of the degree of *linear* relationship between X and Y, and only when the two variables are perfectly related in a linear manner will ρ be as positive or negative as it can be. A ρ less than 1 in absolute value indicates only that the relationship is not completely linear, but there may still be a very strong nonlinear relation. Also, $\rho = 0$ does not imply that X and Y are independent, but only that there is complete absence of a linear relationship. When $\rho = 0$, X and Y are said to be **uncorrelated**. Two variables could be uncorrelated yet highly dependent because of a strong nonlinear relationship, so be careful not to conclude too much from knowing that $\rho = 0$.

Example 5.17 In the manufacture of metal disks, small divots sometimes occur on the surface. If we represent the disk surface by the region $x^2 + y^2 \leq r^2$, one possible joint density function for the location (X, Y) of a divot is

$$f(x, y) = \frac{3}{2\pi r^3} \sqrt{x^2 + y^2} \qquad x^2 + y^2 \leq r^2$$

(This model reflects the fact that it's more likely to see blemishes closer to the disk's edge, since that's where cutting has occurred.)

Since $f(x, y)$ is an even function of x and y, simple symmetry arguments show that $E(X) = 0$, $E(Y) = 0$, and $E(XY) = 0$, from which $\rho_{X,Y} = 0$. So, by definition X and Y are uncorrelated.

However, X and Y are clearly not independent. For instance, if $X = 0$ (so the divot is on the midline), then Y can range from $-r$ to r; however, if $X \approx r$ (divot near the "right" edge), then Y must necessarily be close to 0.

You could also verify that X and Y are not independent by determining their marginal distributions and observing that $f(x, y) \neq f_X(x) \cdot f_Y(y)$, but the marginal pdfs are tedious here. ∎

The next result provides an alternative view of zero correlation.

PROPOSITION   Two rvs X and Y are uncorrelated if, and only if, $E[XY] = \mu_X \cdot \mu_Y$.

Proof By its definition, $\text{Corr}(X, Y) = 0$ iff $\text{Cov}(X, Y) = 0$. Apply the covariance shortcut formula:

$$\rho = 0 \Leftrightarrow \text{Cov}(X, Y) = 0 \Leftrightarrow E[XY] - \mu_X \cdot \mu_Y = 0 \Leftrightarrow E[XY] = \mu_X \cdot \mu_Y \qquad \blacksquare$$

Contrast this with an earlier proposition from this section: If X and Y are *independent*, then $E[g_1(X)g_2(Y)] = E[g_1(X)] \cdot E[g_2(Y)]$ for *all* functions g_1 and g_2. Thus, independence is stronger than zero correlation, the latter just being the special case corresponding to $g_1(X) = X$ and $g_2(Y) = Y$.

Correlation Versus Causation

A value of ρ near 1 does not necessarily imply that increasing the value of X *causes* Y to increase. It implies only that large X values are *associated* with large Y values. For example, in the population of children, vocabulary size and number of cavities are quite positively correlated, but it is certainly not true that cavities cause vocabulary to grow. Instead, the values of both these variables tend to increase as the value of age, a third variable, increases. For children of a fixed age, there is probably a very low correlation between number of cavities and vocabulary size. In summary, association (a high correlation) is not the same as causation.

Exercises: Section 5.2 (23–42)

23. The two most common types of errors made by programmers are syntax errors and logic errors. Let X denote the number of syntax errors and Y the number of logic errors on the first run of a program. Suppose X and Y have the following joint pmf for a particular programming assignment:

		x		
$p(x, y)$	0	1	2	3
0	.71	.03	.02	.01
y 1	.04	.06	.03	.01
2	.03	.03	.02	.01

a. What is the probability a program has more syntax errors than logic errors on the first run?
b. Find the marginal pmfs of X and Y.
c. Are X and Y independent? How can you tell?
d. What is the average number of syntax errors in the first run of a program? What is the average number of logic errors?
e. Suppose an evaluator assigns points to each program with the formula $100 - 4X - 9Y$. What is the expected point score for a randomly selected program?

24. An instructor has given a short quiz consisting of two parts. For a randomly selected student, let $X =$ the number of points earned on the first part and $Y =$ the number of points earned on the second part. Suppose that the joint pmf of X and Y is given in the accompanying table.

			y	
$p(x, y)$	0	5	10	15
0	.02	.06	.02	.10
x 5	.04	.15	.20	.10
10	.01	.15	.14	.01

a. If the score recorded in the grade book is the total number of points earned on the two parts, what is the expected recorded score $E(X + Y)$?
b. If the maximum of the two scores is recorded, what is the expected recorded score?

25. The difference between the number of customers in line at the express checkout and the number in line at the superexpress checkout in Exercise 3 is $X_1 - X_2$. Calculate the expected difference.

26. Six individuals, including A and B, take seats around a circular table in a completely random fashion. Suppose the seats are numbered 1, …, 6. Let $X =$ A's seat number and $Y =$ B's seat number. If A sends a written message around the table to B in the

direction in which they are closest, how many individuals (including A and B) would you expect to handle the message?

27. A surveyor wishes to lay out a square region with each side having length L. However, because of measurement error, he instead lays out a rectangle in which the north–south sides both have length X and the east–west sides both have length Y. Suppose that X and Y are independent and that each one is uniformly distributed on the interval $[L - A, \ L + A]$ (where $0 < A < L$). What is the expected area of the resulting rectangle?

28. Consider a small ferry that can accommodate cars and buses. The toll for cars is $3, and the toll for buses is $10. Let X and Y denote the number of cars and buses, respectively, carried on a single trip. Suppose the joint distribution of X and Y is as given in the table of Exercise 9. Compute the expected revenue from a single trip.

29. Annie and Alvie have agreed to meet for lunch between noon (0:00 p.m.) and 1:00 p.m. Denote Annie's arrival time by X, Alvie's by Y, and suppose X and Y are independent with pdfs

$$f_X(x) = 3x^2 \quad 0 \le x \le 1 \quad f_Y(y) = 2y \quad 0 \le y \le 1$$

What is the expected amount of time that the one who arrives first must wait for the other person? [*Hint*: $h(X, Y) = |X - Y|$.]

30. Suppose that X and Y are independent rvs with moment generating functions $M_X(t)$ and $M_Y(t)$, respectively. If $Z = X + Y$, show that $M_Z(t) = M_X(t) \cdot M_Y(t)$. [*Hint*: Use the proposition on the expected value of a product.]

31. Compute the correlation coefficient ρ for X and Y of Example 5.15 (the covariance has already been computed).

32. a. Compute the covariance for X and Y in Exercise 24.
    b. Compute ρ for X and Y in the same exercise.

33. Compute $\text{Cov}(X, Y)$ and ρ for the variables in Exercise 11.

34. Reconsider the computer component lifetimes X and Y as described in Exercise 14. Determine $E(XY)$. What can be said about $\text{Cov}(X, Y)$ and ρ?

35. Referring back to Exercise 23, calculate both $\text{Cov}(X, Y)$ and ρ.

36. In practice, it is often desired to predict the value of a variable Y from the known value of some other variable, X. For example, a doctor might wish to predict the lifespan Y of someone who smokes X cigarettes a day, or an engineer may require predictions of the tensile strength Y of steel made with concentration X of a certain additive. A *linear predictor* of Y is anything of the form $\hat{Y} = a + bX$; the "hat" ^ on Y indicates prediction.

    A common measure of the quality of a predictor is given by the *mean square prediction error*, $E[(Y - \hat{Y})^2]$.

    a. Show that the choices of a and b that minimize mean square prediction error are

    $$b = \rho \cdot \frac{\sigma_Y}{\sigma_X} \qquad a = \mu_Y - b \cdot \mu_X$$

    where $\rho = \text{Corr}(X, \ Y)$. The resulting expression for $\hat{Y}$ is often called the *best linear predictor* of Y, given X. [*Hint*: Expand the expression for mean square prediction error, apply linearity of expectation, and then use calculus.]

    b. Determine the mean square prediction error for the best linear predictor. How does the value of ρ affect this quantity?

37. Recalling the definition of σ^2 for a single rv X, write a formula that would be appropriate for computing the variance of a function $h(X, Y)$ of two random variables. [*Hint*: Remember that variance is just a special expected value.] Then use this formula to compute the variance of the recorded score $h(X, Y)$ [= max(X, Y)] in part (b) of Exercise 24.

38. Show that when X and Y are independent, $\text{Cov}(X, Y) = \text{Corr}(X, Y) = 0$.

39. Use linearity of expectation to establish the covariance property

$$\text{Cov}(aX + bY + c, Z) = a\text{Cov}(X, Z)$$
$$+ b\text{Cov}(Y, Z)$$

40. a. Use the properties of covariance to show that $\text{Cov}(aX + b, cY + d) = ac\,\text{Cov}(X, Y)$.
    b. Use part (a) along with the rescaling properties standard deviation to show that $\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$ when $ac > 0$ (this is the scale invariance property of correlation).
    c. What happens if a and c have opposite signs, so $ac < 0$?

41. Verify that if $Y = aX + b$ $(a \neq 0)$, then $\text{Corr}(X, Y) = +1$ or -1. Under what conditions will $\rho = +1$?

42. Consider the standardized variables $Z_X = (X - \mu_X)/\sigma_X$ and $Z_Y = (Y - \mu_Y)/\sigma_Y$, and let $\rho = \text{Corr}(X, Y)$.

    a. Use properties of covariance and correlation to verify that $\text{Corr}(X, Y) = \text{Cov}(Z_X, Z_Y) = E(Z_X Z_Y)$.
    b. Use linearity of expectation along with part (a) to show that $E[(Z_Y - \rho Z_X)^2] = 1 - \rho^2$. [*Hint:* If Z is a standardized rv, what are its mean and variance, and how can you use those to determine $E(Z^2)$?]
    c. Use part (b) to show that $-1 \leq \rho \leq 1$.
    d. Use part (b) to show that $\rho = 1$ implies that $Y = aX + b$ where $a > 0$, and $\rho = -1$ implies that $Y = aX + b$ where $a < 0$.

5.3   Linear Combinations

A **linear combination** of random variables refers to anything of the form $a_1X_1 + \cdots + a_nX_n + b$, where the X_i's are random variables and the a_i's and b are numerical constants. (Some sources do not include the constant b in the definition.) For example, suppose your investment portfolio with a particular financial institution includes 100 shares of stock #1, 200 shares of stock #2, and 500 shares of stock #3. Let X_1, X_2, and X_3 denote the share prices of these three stocks at the end of the current fiscal year. Suppose also that the financial institution will levy a management fee of \$150. Then the value of your investments with this institution at the end of the year is $100X_1 + 200X_2 + 500X_3 - 150$, which is a particular linear combination. Important special cases include the total $X_1 + \cdots + X_n$ (take $a_1 = \cdots = a_n = 1$, $b = 0$), the difference of two rvs $X_1 - X_2$ ($n = 2, a_1 = 1, a_2 = -1, b = 0$), and anything of the form $aX + b$ (take $n = 1$ or, equivalently, set $a_2 = \cdots = a_n = 0$). Another very important linear combination is the sample mean $\overline{X} = (X_1 + \cdots + X_n)/n$; just take $a_1 = \cdots = a_n = 1/n$ and $b = 0$.

Notice that we are not requiring the X_i's to be independent or to have the same probability distribution. All the X_i's could have different distributions and therefore different mean values and standard deviations. In this section, we investigate the general properties of linear combinations. Section 6.2 will explore some special properties of the total and sample mean under additional assumptions.

We first consider the expected value and variance of a linear combination.

THEOREM   Let the rvs $X_1, X_2, \ldots, X_n$ have mean values $\mu_1, \ldots, \mu_n$ and standard deviations $\sigma_1, \ldots, \sigma_n$, respectively.

1. Whether or not the X_i's are independent,

$$E(a_1X_1 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b$$
$$= a_1\mu_1 + \cdots + a_n\mu_n + b \tag{5.4}$$

and

$$V(a_1X_1 + \cdots + a_nX_n + b) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\mathrm{Cov}(X_i, X_j)$$
$$= \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum_{i<j}\sum a_ia_j\mathrm{Cov}(X_i, X_j) \tag{5.5}$$

2. If X_1, …, X_n are independent,

$$V(a_1X_1 + \cdots + a_nX_n + b) = a_1^2V(X_1) + \cdots + a_n^2V(X_n)$$
$$= a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2 \tag{5.6}$$

and

$$\sigma_{a_1X_1 + \cdots + a_nX_n + b} = \sqrt{a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2}$$

A paraphrase of (5.4) is that the expected value of a linear combination is the same linear combination of the expected values—for example, $E(2X_1 + 5X_2) = 2\mu_1 + 5\mu_2$. The result (5.6) in Statement 2 is a special case of (5.5) in Statement 1: When the X_i's are independent, $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$ (this simplification actually occurs when the X_i's are uncorrelated, a weaker condition than independence).

Proof ($n = 2$)   To establish (5.4), we could invoke linearity of expectation from Section 5.2, but we present a direct proof here. Suppose that X_1 and X_2 are continuous with joint pdf $f(x_1, x_2)$. Then

$$E(a_1X_1 + a_2X_2 + b) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (a_1x_1 + a_2x_2 + b)f(x_1, x_2)dx_1dx_2$$
$$= a_1\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1f(x_1, x_2)dx_2dx_1 + a_2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_2f(x_1, x_2)dx_1dx_2$$
$$+ b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1, x_2)dx_1dx_2$$
$$= a_1\int_{-\infty}^{\infty} x_1\left[\int_{-\infty}^{\infty} f(x_1, x_2)dx_2\right]dx_1 + a_2\int_{-\infty}^{\infty} x_2\left[\int_{-\infty}^{\infty} f(x_1, x_2)dx_1\right]dx_2 + b(1)$$
$$= a_1\int_{-\infty}^{\infty} x_1f_{X_1}(x_1)dx_1 + a_2\int_{-\infty}^{\infty} x_2f_{X_2}(x_2)dx_2 + b$$
$$= a_1E(X_1) + a_2E(X_2) + b$$

Summation replaces integration in the discrete case. The argument for (5.5) does not require specifying whether either variable is discrete or continuous. Recalling that $V(Y) = E[(Y - \mu_Y)^2]$,

$$
\begin{aligned}
V(a_1X_1 + a_2X_2 + b) &= E[(a_1X_1 + a_2X_2 + b - (a_1\mu_1 + a_2\mu_2 + b))^2] \\
&= E[(a_1X_1 - a_1\mu_1 + a_2X_2 - a_2\mu_2)^2] \\
&= E[a_1^2(X_1 - \mu_1)^2 + a_2^2(X_2 - \mu_2)^2 + 2a_1a_2(X_1 - \mu_1)(X_2 - \mu_2)] \\
&= a_1^2E[(X_1 - \mu_1)^2] + a_2^2E[(X_2 - \mu_2)^2] + 2a_1a_2E[(X_1 - \mu_1)(X_2 - \mu_2)]
\end{aligned}
$$

where the last equality comes from linearity of expectation. We recognize the terms in this last expression as variances and covariance, all together $a_1^2V(X_1) + a_2^2V(X_2) + 2a_1a_2\text{Cov}(X_1, X_2)$, as required. ∎

Example 5.18 A gas station sells three grades of gasoline: regular, plus, and premium. These are priced at \$3.50, \$3.65, and \$3.80 per gallon, respectively. Let X_1, X_2, and X_3 denote the amounts of these grades purchased (gallons) on a particular day. Suppose the X_i's are independent with $\mu_1 = 1000$, $\mu_2 = 500$, $\mu_3 = 300$, $\sigma_1 = 100$, $\sigma_2 = 80$, and $\sigma_3 = 50$. The revenue from sales is $Y = 3.5X_1 + 3.65X_2 + 3.8X_3$, and

$$
\begin{aligned}
E(Y) &= 3.5\mu_1 + 3.65\mu_2 + 3.8\mu_3 = \$6465 \\
V(Y) &= 3.5^2\sigma_1^2 + 3.65^2\sigma_2^2 + 3.8^2\sigma_3^2 = 243{,}864 \\
\sigma_Y &= \sqrt{243{,}864} = \$493.83
\end{aligned}
$$

∎

Example 5.19 Recall that a hypergeometric rv X is the number of successes in a random sample of size n selected without replacement from a population of size N consisting of M successes and $N - M$ failures. It is tricky to obtain the mean value and variance of X directly from the pmf, and the hypergeometric moment generating function is very complicated. We now show how the foregoing proposition on linear combinations can be used to accomplish this task.

To this end, let $X_1 = 1$ if the first individual or object selected is a success and $X_1 = 0$ if it is a failure; define $X_2, X_3, \ldots, X_n$ analogously for the second selection, third selection, and so on. Each X_i is a Bernoulli rv, and each has the same marginal distribution: $p(1) = M/N$ and $p(0) = 1 - M/N$ (this is obvious for X_1, which is based on the very first draw from the population, and can be verified for the other draws as well). Thus $E(X_i) = 0(1 - M/N) + 1(M/N) = M/N$. The total number of success in the sample is $X = X_1 + \cdots + X_n$ (a 1 is added in for each success and a 0 for each failure), so

$$
E(X) = E(X_1) + \cdots + E(X_n) = M/N + M/N + \cdots + M/N = n(M/N) = np
$$

where p denotes the success probability on any particular draw (trial). That is, just as in the case of a binomial rv, the expected value of a hypergeometric rv is the success probability on any trial multiplied by the number of trials. Notice that we were able to apply Equation (5.4), even though the X_i's are not independent.

Since each X_i is Bernoulli, it follows that $V(X_i) = p(1 - p)$ or $M/N(1 - M/N)$. However, the variance of X here is *not* the same as the binomial variance, precisely because the successive draws are not independent. Consider $p(x_1, x_2)$, the joint pmf of X_1 and X_2:

$$p(1,1) = \frac{M}{N}\left(\frac{M-1}{N-1}\right), \quad p(0,0) = \left(\frac{N-M}{N}\right)\left(\frac{N-M-1}{N-1}\right), \quad p(1,0) = p(0,1) = \frac{M}{N}\left(\frac{N-M}{N-1}\right)$$

This is also the joint pmf of any pair X_i, X_j. A slightly tedious calculation then results in

$$\text{Cov}(X_i, X_j) = -\frac{p(1-p)}{N-1} \quad \text{for } i \neq j$$

Applying Equation (5.5) yields

$$V(X) = V(X_1 + \cdots + X_n) = \sum_{i=1}^{n} V(X_i) + 2\sum\sum_{i<j}\text{Cov}(X_i, X_j)$$

$$= nV(X_1) + 2\binom{n}{2}\text{Cov}(X_1, X_2)$$

$$= np(1-p) + n(n-1)\cdot\frac{-p(1-p)}{N-1} = \cdots = np(1-p)\left(\frac{N-n}{N-1}\right)$$

This is quite close to the binomial variance provided that n is much smaller than N so that the last term in parentheses is close to 1. ∎

The following corollary expresses the $n = 2$ case of the main theorem for ease of use, including the important special cases of the sum and the difference of two random variables.

COROLLARY   For any two rvs X_1 and X_2, and any constants a_1, a_2, b,

$$E(a_1X_1 + a_2X_2 + b) = a_1E(X_1) + a_2E(X_2) + b$$

and

$$V(a_1X_1 + a_2X_2 + b) = a_1^2V(X_1) + a_2^2V(X_2) + 2a_1a_2\text{Cov}(X_1, X_2)$$

In particular, $E(X_1 + X_2) = E(X_1) + E(X_2)$ and, if X_1 and X_2 are independent, $V(X_1 + X_2) = V(X_1) + V(X_2)$.[1] Also, $E(X_1 - X_2) = E(X_1) - E(X_2)$ and, if X_1 and X_2 are independent, $V(X_1 - X_2) = V(X_1) + V(X_2)$.

The expected value of a difference is the difference of the two expected values, but the variance of a difference between two independent variables is the *sum*, not the difference, of the two variances. There is just as much variability in $X_1 - X_2$ as in $X_1 + X_2$: Writing $X_1 - X_2 = X_1 + (-1)X_2$, the term $(-1)X_2$ has the same amount of variability as X_2 itself.

Example 5.20 An automobile manufacturer equips a particular model with either a six-cylinder engine or a four-cylinder engine. Let X_1 and X_2 be fuel efficiencies (mpg) for independently and randomly selected six-cylinder and four-cylinder cars, respectively. With $\mu_1 = 22$, $\mu_2 = 26$, $\sigma_1 = 1.2$, and $\sigma_2 = 1.5$,

---

[1]This property of independent rvs can also be written as $\sigma_1^2 + \sigma_2^2 = \sigma_{X_1+X_2}^2$. In part because the formula has the format $a^2 + b^2 = c^2$, statisticians sometimes call this property the *Pythagorean Theorem*.

$$E(X_1 - X_2) = \mu_1 - \mu_2 = 22 - 26 = -4 \text{ mpg}$$
$$V(X_1 - X_2) = \sigma_1^2 + \sigma_2^2 = 1.2^2 + 1.5^2 = 3.69$$
$$\sigma_{X_1 - X_2} = \sqrt{3.69} = 1.92 \text{ mpg}$$

If we re-label so that X_1 refers to the four-cylinder car, then $E(X_1 - X_2) = 26 - 22 = 4$ mpg, but the standard deviation of the difference is still 1.92 mpg. ∎

The PDF of a Sum of Continuous RVs

Generally speaking, knowing the mean and standard deviation of a random variable W is not enough to specify its probability distribution and thus be able to compute probabilities such as $P(W > 10)$ or $P(W \le -2)$. In the case of independent rvs, a general method exists for determining the pdf of the sum $X_1 + \cdots + X_n$ from their marginal pdfs. We present first the result for two random variables.

THEOREM   Suppose X and Y are independent, continuous rvs with marginal pdfs $f_X(x)$ and $f_Y(y)$, respectively. Then the pdf of the rv $W = X + Y$ is given by

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w - x) dx$$

[In mathematics, this integral operation is known as the **convolution** of $f_X(x)$ and $f_Y(y)$ and is sometimes denoted $f_W = f_X \star f_Y$.] The limits of integration are determined by which x values make both $f_X(x) > 0$ and $f_Y(w - x) > 0$.

Proof Since X and Y are independent, their joint pdf is given by $f_X(x) \cdot f_Y(y)$. The cdf of W is then

$$F_W(w) = P(W \le w) = P(X + Y \le w)$$

To calculate $P(X + Y \le w)$, we must integrate over the set of numbers $\{(x, y): x + y \le w\}$, which is the shaded region indicated in Figure 5.5.
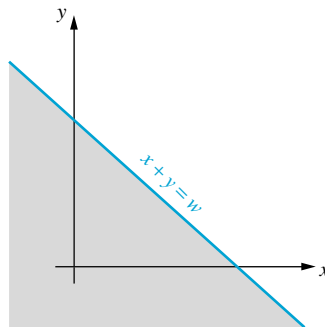
Figure 5.5 Region of integration for $P(X + Y \le w)$

The resulting limits of integration are $-\infty < x < \infty$ and $-\infty < y \le w - x$, and so

$$
\begin{aligned}
F_W(w) &= P(X + Y \le w) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{w-x} f_X(x) f_Y(y) \, dy \, dx = \int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{w-x} f_Y(y) \, dy \, dx \\
&= \int_{-\infty}^{\infty} f_X(x) F_Y(w - x) \, dx
\end{aligned}
$$

The pdf of W is the derivative of this expression with respect to w; taking the derivative underneath the integral sign yields the desired result. ∎

By a similar argument, the pdf of $W = X + Y$ can be determined even when X and Y are not independent. Assuming X and Y have joint pdf $f(x, y)$, $f_W(w) = \int_{-\infty}^{\infty} f(x, w - x) dx$.

Example 5.21 In a *standby system*, a component is used until it wears out and is then immediately replaced by another, not necessarily identical, component. (The second component is said to be "in standby mode," i.e., waiting to be used.) The overall lifetime of a standby system is just the sum of the lifetimes of its individual components. Let X and Y denote the lifetimes of the two components of a standby system, and suppose X and Y are independent exponentially distributed random variables with mean lifetimes 3 weeks and 4 weeks, respectively. Let $W = X + Y$, the system lifetime.

Using Equation (5.4), the expected lifetime of the standby system is $E(W) = E(X) + E(Y) = 3 + 4 = 7$ weeks. Since X and Y are exponential, the variance of each one is the square of its mean (9 and 16, respectively); since they are also independent,

$$
V(W) = V(X) + V(Y) = 3^2 + 4^2 = 25
$$

It follows that $\sigma_W = 5$ weeks. Since $\mu_W \ne \sigma_W$, W cannot itself be exponentially distributed, but we can use the previous theorem to find its pdf.

The marginal pdfs of X and Y are $f_X(x) = (1/3)e^{-x/3}$ for $x > 0$ and $f_Y(y) = (1/4)e^{-y/4}$ for $y > 0$. Substituting $y = w - x$, the inequalities $x > 0$ and $w - x > 0$ imply $0 < x < w$, which specify the limits of integration of the convolution integral:

$$
\begin{aligned}
f_W(w) &= \int_{-\infty}^{\infty} f_X(x) f_Y(w - x) dx = \int_{0}^{w} (1/3)e^{-x/3}(1/4)e^{-(w-x)/4} dx = \frac{1}{12} e^{-w/4} \int_{0}^{w} e^{-x/12} dx \\
&= e^{-w/4}(1 - e^{-w/12}) \quad w > 0
\end{aligned}
$$

A graph of this pdf appears in Figure 5.6. As a check, the mean and variance of W can be verified directly from its pdf.
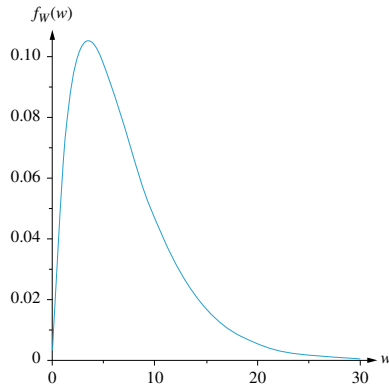
Figure 5.6 The pdf of $W = X + Y$ for Example 5.21

The probability that the system lasts more than its expected lifetime of 7 weeks is given by

$$P(W > 7) = \int_7^\infty f_W(w)dw = \int_7^\infty e^{-w/4}(1 - e^{-w/12})dw = .4042 \qquad \blacksquare$$

As a generalization of the previous proposition, the pdf of the sum $W = X_1 + \cdots + X_n$ of n independent, continuous rvs can be determined by successive convolution: $f_W = f_1 \star \cdots \star f_n$. In most situations, it isn't practical to evaluate such a complicated object. Thankfully, as we'll see next, such tedious computations can sometimes be avoided with the use of moment generating functions.

Moment Generating Functions for Linear Combinations

A proposition in Section 5.2 stated that the expected value of a product of functions of independent random variables is the product of the individual expected values. We now use this to formulate the moment generating function of a linear combination of independent random variables.

PROPOSITION

Let X_1, X_2, ..., X_n be independent rvs with moment generating functions $M_{X_1}(t), M_{X_2}(t), \ldots, M_{X_n}(t)$, respectively. Then the moment generating function of $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b$ is

$$M_Y(t) = e^{bt} \cdot M_{X_1}(a_1t) \cdot M_{X_2}(a_2t) \cdot \cdots \cdot M_{X_n}(a_nt)$$

In the special case that $a_1 = a_2 = \cdots = a_n = 1$ and $b = 0$, so $Y = X_1 + \cdots + X_n$,

$$M_Y(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdot \cdots \cdot M_{X_n}(t)$$

That is, the mgf of a sum of independent rvs is the *product* of the individual mgfs.

Proof First, we write the moment generating function of Y as the expected value of a product.

$$M_Y(t) = E[e^{tY}] = E[e^{t(a_1X_1 + a_2X_2 + \cdots + a_nX_n + b)}]$$
$$= E[e^{ta_1X_1 + ta_2X_2 + \cdots + ta_nX_n + tb}] = e^{bt}E[e^{a_1tX_1} \cdot e^{a_2tX_2} \cdot \cdots \cdot e^{a_ntX_n}]$$

The last expression inside brackets is the product of functions of X_1, X_2, …, X_n. Since the X_i's are independent, the expected value can be distributed across this product:

$$e^{bt}E[e^{ta_1X_1} \cdot e^{ta_2X_2} \cdot \cdots \cdot e^{ta_nX_n}] = e^{bt}E[e^{ta_1X_1}] \cdot E[e^{ta_2X_2}] \cdot \cdots \cdot E[e^{ta_nX_n}]$$
$$= e^{bt}M_{X_1}(a_1t) \cdot M_{X_2}(a_2t) \cdot \cdots \cdot M_{X_n}(a_nt) \qquad \blacksquare$$

Now suppose we wish to determine the pdf of some linear combination of independent rvs. Provided we have their mgfs, the previous proposition makes it easy to determine the mgf of the linear combination. Then, if we can recognize this mgf as belonging to some known distributional family (binomial, exponential, etc.), the uniqueness property of mgfs guarantees our linear combination has that particular distribution. The next several propositions illustrate this technique.

PROPOSITION   If X_1, X_2, …, X_n are independent, normally distributed rvs (with possibly different means and/or sds), then any linear combination of the X_i's also has a normal distribution. In particular, the sum of independent normally distributed rvs itself has a normal distribution, and the difference $X_1 - X_2$ between two independent, normally distributed variables is itself normally distributed.

Proof Let $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b$, where X_i is normally distributed with mean μ_i and standard deviation σ_i, and the X_i are independent. From Section 4.3, $M_{X_i}(t) = e^{\mu_i t + \sigma_i^2 t^2/2}$. Therefore,

$$M_Y(t) = e^{bt}M_{X_1}(a_1t) \cdot M_{X_2}(a_2t) \cdot \cdots \cdot M_{X_n}(a_nt)$$
$$= e^{bt}e^{\mu_1 a_1 t + \sigma_1^2 a_1^2 t^2/2}e^{\mu_2 a_2 t + \sigma_2^2 a_2^2 t^2/2} \cdot \cdots \cdot e^{\mu_n a_n t + \sigma_n^2 a_n^2 t^2/2}$$
$$= e^{(\mu_1 a_1 + \mu_2 a_2 + \cdots + \mu_n a_n + b)t + (\sigma_1^2 a_1^2 + \sigma_2^2 a_2^2 + \cdots + \sigma_n^2 a_n^2)t^2/2}$$
$$= e^{\mu t + \sigma^2 t^2/2},$$

where $\mu = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n + b$ and $\sigma^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2$. We recognize this function as the mgf of a normal random variable, and it follows that Y is normally distributed by the uniqueness property of mgfs. Notice that the mean and variance are in agreement with the first proposition of this section. $\blacksquare$

Example 5.22 (Example 5.18 continued) The total revenue from the sale of the three grades of gasoline on a particular day was $Y = 3.5X_1 + 3.65X_2 + 3.8X_3$, and we calculated $\mu_Y = \$6465$ and (assuming independence) $\sigma_Y = \$493.83$. If the X_i's are (approximately) normally distributed, the probability that revenue exceeds $5000 is

$$P(Y > 5000) \approx P\left(Z > \frac{5000 - 6465}{493.83}\right) = P(Z > -2.967) = 1 - \Phi(-2.967) = .9985 \qquad \blacksquare$$

This same method may be applied to discrete rvs, as the next proposition indicates.

PROPOSITION   Suppose $X_1, \ldots, X_n$ are independent Poisson random variables, where X_i has mean μ_i. Then $Y = X_1 + \cdots + X_n$ also has a Poisson distribution, with mean $\mu_1 + \cdots + \mu_n$.

Proof From Section 3.6, the mgf of a Poisson rv with mean μ is $e^{\mu(e^t - 1)}$. Since Y is the sum of the X_i's, and the X_i's are independent,

$$M_Y(t) = M_{X_1}(t) \cdot \cdots \cdot M_{X_n}(t) = e^{\mu_1(e^t - 1)} \cdot \cdots \cdot e^{\mu_n(e^t - 1)} = e^{(\mu_1 + \cdots + \mu_n)(e^t - 1)}$$

This is the mgf of a Poisson rv with mean $\mu_1 + \cdots + \mu_n$. Therefore, by the uniqueness property of mgfs, Y has a Poisson distribution with mean $\mu_1 + \cdots + \mu_n$. $\qquad \blacksquare$

Example 5.23 During the open enrollment period at a large university, the number of freshmen registering for classes through the online registration system in one hour follows a Poisson distribution with mean 80 students; denote this rv by X_1. Define X_2, X_3, and X_4 similarly for sophomores, juniors, and seniors, and suppose the corresponding means are 125, 118, and 140, respectively. Assume these four counts are independent. The rv $Y = X_1 + X_2 + X_3 + X_4$ represents the total number of undergraduate students registering in one hour; by the preceding proposition, Y is also a Poisson rv, but with mean $80 + 125 + 118 + 140 = 463$ students and standard deviation $\sqrt{463} = 21.5$ students. The probability that more than 500 students enroll during one hour, exceeding the registration system's capacity, is then $P(Y > 500) = 1 - P(Y \leq 500) = .042$ (using software). $\qquad \blacksquare$

Because of the properties stated in the preceding two propositions, both the normal and Poisson models are sometimes called *additive distributions*, meaning that the sum of independent rvs from that family (normal or Poisson) will also belong to that family. The next proposition shows that not all of the major probability distributions are additive; its proof is left as an exercise (Exercise 65).

PROPOSITION   Suppose $X_1, \ldots, X_n$ are independent exponential random variables with common parameter λ. Then $Y = X_1 + \cdots + X_n$ has a gamma distribution, with parameters $\alpha = n$ and $\beta = 1/\lambda$.

Therefore, the exponential distribution is *not* additive, although it can be shown that its "parent," the gamma distribution, is additive under certain conditions (see Exercise 64). Notice that this proposition requires the X_i's to have the same "rate" parameter λ; i.e., the X_i's must be independent *and* identically distributed for their sum to have a gamma distribution. As we saw in Example 5.21, the sum of two independent exponential rvs with different parameter values follows neither an exponential nor a gamma distribution.

Exercises: Section 5.3 (43–67)

43. A shipping company handles containers in three different sizes: (1) 27 ft^3 ($3 \times 3 \times 3$), (2) 125 ft^3, and (3) 512 ft^3. Let $X_i (i = 1, 2, 3)$ denote the number of type i containers shipped during a given week. With $\mu_i = E(X_i)$ and $\sigma_i^2 = V(X_i)$, suppose that the mean values and standard deviations are as follows:

$$\mu_1 = 200 \quad \mu_2 = 250 \quad \mu_3 = 100$$
$$\sigma_1 = 10 \quad \sigma_2 = 12 \quad \sigma_3 = 8$$

   a. Assuming that X_1, X_2, X_3 are independent, calculate the expected value and standard deviation of the total volume shipped. [*Hint*: Volume = $27X_1 + 125X_2 + 512X_3$.]
   b. Would your calculations necessarily be correct if the X_i's were not independent? Explain.
   c. Suppose the X_i's are independent with each one (approximately) normal. What is the (approximate) probability that the total volume shipped is at most $100{,}000 \text{ ft}^3$?

44. Let X_1, X_2, and X_3 represent the times necessary to perform three successive repair tasks at a service facility. Suppose they are independent, normal rvs with expected values μ_1, μ_2, and μ_3 and variances σ_1^2, σ_2^2, and σ_3^2, respectively.

   a. If $\mu_1 = \mu_2 = \mu_3 = 60$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 15$, calculate $P(X_1 + X_2 + X_3 \le 200)$.
   b. Using the μ_i's and σ_i's from (a), what is $P(150 \le X_1 + X_2 + X_3 \le 200)$?
   c. Using the μ_i's and σ_i's given in part (a), calculate $P(55 \le \overline{X})$ and $P(58 \le \overline{X} \le 62)$. [*Hint*: $\overline{X} = (X_1 + X_2 + X_3)/3$.]
   d. Using the values from part (a), calculate $P(-10 \le X_1 - .5X_2 - .5X_3 \le 5)$.
   e. If $\mu_1 = 40$, $\mu_2 = 50$, $\mu_3 = 60$, $\sigma_1^2 = 10$, $\sigma_2^2 = 12$, and $\sigma_3^2 = 14$, calculate $P(X_1 + X_2 + X_3 \le 160)$ and also $P(X_1 + X_2 \ge 2X_3)$.

45. Five automobiles of the same type are to be driven on a 300-mile trip. The first two have six-cylinder engines, and the other three have four-cylinder engines. Let X_1, X_2, X_3, X_4, and X_5 be the observed fuel efficiencies (mpg) for the five cars. Suppose these variables are independent and normally distributed with $\mu_1 = \mu_2 = 30$, $\mu_3 = \mu_4 = \mu_5 = 35$, and $\sigma = 2.5$ for the two larger engines and 3.6 for the three smaller engines. Define a rv Y by

$$Y = \frac{X_1 + X_2}{2} - \frac{X_3 + X_4 + X_5}{3}$$

so that Y is a measure of the difference in efficiency between the six-cylinder and four-cylinder engines. Compute $P(Y \ge 0)$ and $P(-3 \le Y \le 3)$. [*Hint*: Y is a linear combination; what are the a_i's?]

46. Exercise 28 introduced random variables X and Y, and the number of cars and buses, respectively, carried by a ferry on a single trip. These rvs are, in fact, independent.

   a. Compute the expected value, variance, and standard deviation of the total number of vehicles on a single trip.
   b. If each car is charged \$3 and each bus \$10, compute the expected value, variance, and standard deviation of the revenue resulting from a single trip.

47. A concert has three pieces of music to be played before intermission. The time taken to play each piece has a normal distribution. Assume that the three times are independent of each other. The mean times are 15, 30, and 20 min, respectively, and the standard deviations are 1, 2, and 1.5 min, respectively. What is the probability that this part of the concert takes at most one hour? Are there reasons to question the independence assumption? Explain.

48. Refer to Exercise 3.

    a. Calculate the covariance between $X_1 = $ the number of customers in the express checkout and $X_2 = $ the number of customers in the superexpress checkout.
    b. Calculate $V(X_1 + X_2)$. How does this compare to $V(X_1) + V(X_2)$?

49. Suppose your waiting time for a bus in the morning is uniformly distributed on [0, 8], whereas waiting time in the evening is uniformly distributed on [0, 10] independent of morning waiting time.

    a. If you take the bus each morning and evening for a week, what is your total expected waiting time? [*Hint*: Define rvs $X_1, \ldots, X_{10}$ and use a rule of expected value.]
    b. What is the variance of your total waiting time?
    c. What are the expected value and variance of the difference between morning and evening waiting times on a given day?
    d. What are the expected value and variance of the difference between total morning waiting time and total evening waiting time for a particular week?

50. An insurance office buys paper by the ream (500 sheets) for use in the copier, fax, and printer. Each ream lasts an average of 4 days, with standard deviation 1 day. The distribution is normal, independent of previous reams.

    a. Find the probability that the next ream outlasts the present one by more than two days.
    b. How many reams must be purchased if they are to last at least 60 days with probability at least 80%?

51. If two loads are applied to a cantilever beam as shown in the accompanying drawing, the bending moment at 0 due to the loads is $a_1X_1 + a_2X_2$.
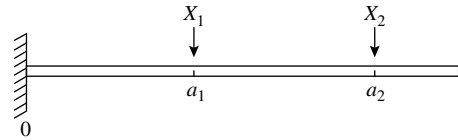
    a. Suppose that X_1 and X_2 are independent rvs with means 2 and 4 kips, respectively, and standard deviations .5 and 1.0 kip, respectively. If $a_1 = 5$ ft and $a_2 = 10$ ft, what is the expected bending moment and what is the standard deviation of the bending moment?
    b. If X_1 and X_2 are normally distributed, what is the probability that the bending moment will exceed 75 kip-ft?
    c. Suppose the positions of the two loads are random variables. Denoting them by A_1 and A_2, assume that these variables have means of 5 and 10 ft, respectively, that each has a standard deviation of .5, and that all A_i's and X_i's are independent of each other. What is the expected moment now?
    d. For the situation of part (c), what is the variance of the bending moment?
    e. If the situation is as described in part (a) except that $\text{Corr}(X_1, X_2) = .5$ (so that the two loads are not independent), what is the variance of the bending moment?

52. One piece of PVC pipe is to be inserted inside another piece. The length of the first piece is normally distributed with mean value 20 in. and standard deviation .5 in. The length of the second piece is a normal rv with mean and standard deviation 15 in. and .4 in., respectively. The amount of overlap is normally distributed with mean value 1 in. and standard deviation .1 in. Assuming that the lengths and amount of overlap are independent of each other, what is the probability that the total length after insertion is between 34.5 in. and 35 in.?

53. Two airplanes are flying in the same direction in adjacent parallel corridors. At time $t = 0$, the first airplane is 10 km ahead of the second one. Suppose the speed of the first plane (km/h) is normally distributed with mean 520 and standard deviation 10 and the second plane's speed, independent of the first, is also normally distributed with mean and standard deviation 500 and 10, respectively.

   a. What is the probability that after 2 h of flying, the second plane has not caught up to the first plane?
   b. Determine the probability that the planes are separated by at most 10 km after 2 h.

54. Three different roads feed into a particular freeway entrance. Suppose that during a fixed time period, the number of cars coming from each road onto the freeway is a random variable, with expected value and standard deviation as given in the table.

	Road 1	Road 2	Road 3
Expected value	800	1000	600
Standard deviation	16	25	18

   a. What is the expected total number of cars entering the freeway at this point during the period? [*Hint*: Let X_i = the number from road i.]
   b. What is the standard deviation of the total number of entering cars? Have you made any assumptions about the relationship between the numbers of cars on the different roads?
   c. With X_i denoting the number of cars entering from road i during the period, suppose $Cov(X_1, X_2) = 80, Cov(X_1, X_3) = 90$, and $Cov(X_2, X_3) = 100$ (so that the three streams of traffic are not independent). Compute the expected total number of entering cars and the standard deviation of the total.

55. Consider independent rvs $X_1, \ldots, X_n$ from a continuous distribution having median 0, so that the probability of any one observation being positive is .5. Now disregard the signs of the observations, rank them from smallest to largest in absolute value, and then let W = the sum of the ranks of the observations having positive signs. For example, if the observations are $-.3, +.7, +2.1$, and -2.5, then the ranks of positive observations are 2 and 3, so $W = 5$. In the statistics literature, W is called *Wilcoxon's signed-rank statistic*. W can be represented as follows:

$$W = 1 \cdot Y_1 + 2 \cdot Y_2 + 3 \cdot Y_3 + \cdots + n \cdot Y_n$$
$$= \sum_{i=1}^{n} i \cdot Y_i$$

where the Y_i's are independent Bernoulli rvs, each with $p = .5$ ($Y_i = 1$ corresponds to the observation with rank i being positive). Compute the following:

   a. $E(Y_i)$ and then $E(W)$ using the equation for W [*Hint*: The first n positive integers sum to $n(n + 1)/2$.]
   b. $V(Y_i)$ and then $V(W)$ [*Hint*: The sum of the squares of the first n positive integers is $n(n + 1)(2n + 1)/6$.]

56. In Exercise 51, the weight of the beam itself contributes to the bending moment. Assume that the beam is of uniform thickness and density so that the resulting load is uniformly distributed on the beam. If the weight of the beam is random, the resulting load from the weight is also random; denote this load by W (kip-ft).

   a. If the beam is 12 ft long, W has mean 1.5 and standard deviation .25, and the fixed loads are as described in part (a) of Exercise 51, what are the expected value and variance of the bending moment? [*Hint:* If the load due to the beam were w kip-ft, the contribution to the bending moment would be $w \int_0^{12} x\,dx$.]

b. If all three variables (X_1, X_2, and W) are normally distributed, what is the probability that the bending moment will be at most 200 kip-ft?

57. A professor has three errands to take care of in the Administration Building. Let X_i = the time that it takes for the ith errand (i = 1, 2, 3), and let X_4 = the total time in minutes that she spends walking to and from the building and between each errand. Suppose the X_i's are independent, normally distributed, with the following means and standard deviations: $\mu_1 = 15$, $\sigma_1 = 4$, $\mu_2 = 5$, $\sigma_2 = 1$, $\mu_3 = 8$, $\sigma_3 = 2$, $\mu_4 = 12$, $\sigma_4 = 3$. She plans to leave her office at precisely 10:00 a.m. and wishes to post a note on her door that reads, "I will return by t a.m." What time t should she write down if she wants the probability of her arriving after t to be .01?

58. In an area having sandy soil, 50 small trees of a certain type were planted, and another 50 trees were planted in an area having clay soil. Let X = the number of trees planted in sandy soil that survive 1 year and Y = the number of trees planted in clay soil that survive 1 year. If the probability that a tree planted in sandy soil will survive 1 year is .7 and the probability of 1-year survival in clay soil is .6, compute an approximation to $P(-5 \le X - Y \le 5)$. [*Hint:* Use a normal approximation from Section 3.3. Do not bother with the continuity correction.]

59. Let X and Y be independent rvs, with $X \sim N(0, 1)$ and $Y \sim N(0, 1)$.

a. Use convolution to show that $X + Y$ is also normal, and identify its mean and standard deviation.

b. Use the additive property of the normal distribution presented in this section to verify your answer to part (a).

60. Karen throws two darts at a board with radius 10 in.; let X and Y denote the distances of the two darts from the center of the board. Under the system Karen uses, the score she obtains depends on $W = X + Y$, the sum of these two distances. Assume X and Y are independent.

a. If X and Y are both uniform distributed on the interval [0, 10], use convolution to determine the pdf of $W = X + Y$. Be very careful with your limits of integration!

b. Based on the pdf in part (a), calculate $P(X + Y \le 5)$.

c. If Karen's darts are equally likely to land anywhere on the board, it can be shown that the pdfs of X and Y are $f_X(x) = x/50$ for $0 \le x \le 10$ and $f_Y(y) = y/50$ for $0 \le y \le 10$. Use convolution to determine the pdf of $W = X + Y$. Then, calculate $P(X + Y \le 5)$.

61. Siblings Matt and Liz both enjoy playing roulette. One day, Matt brought $10 to the local casino and Liz brought $15. They sat at different tables, each made $1 wagers on red on consecutive spins (10 spins for Matt, 15 for Liz). Let X = the number of times Matt won and Y = the number of times Liz won.

a. What is a reasonable probability model for X? [*Hint:* Successive spins of a roulette wheel are independent, and P(land on red) = 18/38.]

b. What is a reasonable probability model for Y?

c. What is a reasonable probability model for $X + Y$, the total number of times Matt and Liz win that day? Explain. [*Hint:* Since the siblings sat at different table, their gambling results are independent.]

d. Use moment generating functions, along with your answers to (a) and (b), to show that your answer to part (c) is correct.

e. Generalize part (d): If X_1, …, X_k are independent binomial rvs, with $X_i \sim \text{Bin}(n_i, p)$, show that their sum is also binomially distributed.

f. Does the result of part (e) hold if the parameter p has a different value for each X_i (e.g., if Matt bets on red but Liz bets on the number 27)?

62. The children attending Milena's birthday party are enjoying taking swings at a piñata. Let X = the number of swings it takes Milena to hit the piñata once (since she's the birthday girl, she goes first), and let Y = the number of swings it takes her brother Lucas to hit the piñata once (he goes second). Assume the results of successive swings are independent (the children don't improve, since they're blindfolded), and that each child has a .2 probability of hitting the piñata on any attempt.

a. What is a reasonable probability model for X?
b. What is a reasonable probability model for Y?
c. What is a reasonable probability model for $X + Y$, the total number of swings taken by Milena and Lucas? Explain. (Assume Milena's and Lucas' results are independent.)
d. Use moment generating functions, along with your answers to (a) and (b), to show that your answer to part (c) is correct.
e. Generalize part (d): If X_1, ..., X_r are independent geometric rvs with common parameter p, show that their sum has a negative binomial distribution.
f. Does the result of part (e) hold if the probability parameter p is different for each X_i (e.g., if Milena has probability .4 on each attempt while Lucas' success probability is only .1)?

63. Let X_1, ..., X_n be independent rvs, with X_i having a negative binomial distribution with parameters r_i and p ($i = 1$, ..., n). Use moment generating functions to show that $X_1 + \cdots + X_n$ has a negative binomial distribution, and identify the parameters of this distribution. Explain why this answer makes sense, based on the negative binomial model. [*Note:* Each X_i may have a different parameter r_i, but all have the same p parameter.]

64. Let X and Y be independent gamma random variables, both with the same scale parameter β. The value of the other parameter is α_1 for X and α_2 for Y. Use moment generating functions to show that $X + Y$ is also gamma distributed, with shape parameter $\alpha_1 + \alpha_2$ and scale parameter β. Is $X + Y$ gamma distributed if the scale parameters are different? Explain.

65. Let X and Y be independent exponential random variables with common parameter λ.

a. Use convolution to show that $X + Y$ has a gamma distribution, and identify the parameters of that gamma distribution.
b. Use the previous exercise to establish the same result.
c. Generalize part (b): If X_1, ..., X_n are independent exponential rvs with common parameter λ, what is the distribution of their sum?

66. For men, pulse rates (in beats per minute) are normally distributed with mean 70 and standard deviation 10. Women's pulse rates are normally distributed with mean 77 and standard deviation 12. Let $\overline{X}$ = the sample average pulse rate for a random sample of 40 men and let $\overline{Y}$ = the sample average pulse rate for a random sample of 36 women.

a. What is the distribution of $\overline{X}$? Of $\overline{Y}$? [*Hint:* $\overline{X} = \frac{1}{40}X_1 + \cdots + \frac{1}{40}X_{40}$, and similarly for $\overline{Y}$.]
b. What is the distribution of $\overline{X} - \overline{Y}$? Justify your answer.
c. Calculate $P(-2 \leq \overline{X} - \overline{Y} \leq 1)$.

d. Calculate $P(\overline{X} - \overline{Y} \leq -15)$. If you actually observed $\overline{X} - \overline{Y} \leq -15$, would you doubt that $\mu_1 - \mu_2 = -7$? Explain.

67. The *Laplace* (or *double exponential*) *distribution* has pdf $f(x) = \frac{1}{2}e^{-|x|}$ for $-\infty < x < \infty$.

   a. The mean of the Laplace distribution is clearly 0, by symmetry. Determine the variance of the Laplace distribution.

   b. Show that the mgf of the Laplace distribution is $M_X(t) = 1/(1 - t^2)$ for $-1 < t < 1$.

   c. Now let $Y_n = X_1 + \cdots + X_n$, where the X_i are iid Laplace rvs. Determine the mean, variance, and mgf of Y_n.

   d. Define a standardized version of Y_n by $Z_n = (Y_n - \mu_{Y_n})/\sigma_{Y_n}$. Determine the mgf of Z_n.

   e. Show that as $n \to \infty$, the limiting mgf of Z_n is $e^{t^2/2}$, the mgf of a standard normal rv.

(This is a preview of the celebrated Central Limit Theorem, which we'll encounter in Chapter 6.)

5.4 Conditional Distributions and Conditional Expectation

The distribution of Y can depend strongly on the value of another variable X. For example, if X is height and Y is weight, the weight distribution for men who are 6 ft tall is very different from the weight distribution for short men. The conditional distribution of Y given $X = x$ describes for each possible x how probability is distributed over the set of possible y values. We define the conditional distribution of Y given X, but the conditional distribution of X given Y can be obtained by just reversing the roles of X and Y. Both definitions are analogous to that of the conditional probability $P(A|B)$ as the ratio $P(A \cap B)/P(B)$.

DEFINITION

Let X and Y be two discrete random variables with joint pmf $p(x, y)$ and marginal X pmf $p_X(x)$. Then for any x value such that $p_X(x) > 0$, the **conditional probability mass function of Y given $X = x$** is

$$p_{Y|X}(y|x) = \frac{p(x, y)}{p_X(x)}$$

An analogous formula holds in the continuous case. Let X and Y be two continuous random variables with joint pdf $f(x,y)$ and marginal X pdf $f_X(x)$. Then for any x value such that $f_X(x) > 0$, the **conditional probability density function of Y given $X = x$** is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}$$

Example 5.24 For a discrete example, reconsider Example 5.1, where X represents the deductible amount on an automobile policy and Y represents the deductible amount on a homeowner's policy. Here is the joint distribution again.

	$p(x, y)$		y	
		0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

The distribution of Y depends on X. In particular, let's find the conditional probability that Y is 200, given that X is 250, using the definition of conditional probability from Section 2.4:

$$P(Y = 200|X = 250) = \frac{P(Y = 200 \cap X = 250)}{P(X = 250)} = \frac{.3}{.05 + .15 + .3} = .6$$

With our new definition we obtain the same result:

$$p_{Y|X}(200|250) = \frac{p(250, 200)}{p_X(250)} = \frac{.3}{.05 + .15 + .3} = .6$$

The conditional probabilities for the two other possible values of Y are

$$p_{Y|X}(0|250) = \frac{p(250, 0)}{p_X(250)} = \frac{.05}{.05 + .15 + .3} = .1$$

$$p_{Y|X}(100|250) = \frac{p(250, 100)}{p_X(250)} = \frac{.15}{.05 + .15 + .3} = .3$$

Thus, $p_{Y|X}(0|250) + p_{Y|X}(100|250) + p_{Y|X}(200|250) = .1 + .3 + .6 = 1$. This is no coincidence; conditional probabilities satisfy the properties of ordinary probabilities. They are nonnegative and they sum to 1. Essentially, the denominator in the definition of conditional probability is designed to make the total be 1.

Reversing the roles of X and Y, we find the conditional probabilities for X, given that $Y = 0$:

$$p_{X|Y}(100|0) = \frac{p(100, 0)}{p_Y(0)} = \frac{.20}{.20 + .05} = .8$$

$$p_{X|Y}(250|0) = \frac{p(250, 0)}{p_Y(0)} = \frac{.05}{.20 + .05} = .2$$

Again, the conditional probabilities add to 1.   ∎

Example 5.25 For a continuous example, recall Example 5.5, where X is the weight of almonds and Y is the weight of cashews in a can of mixed nuts. The sum of $X + Y$ is at most one pound, the total weight of the can of nuts. The joint pdf of X and Y is

$$f(x, y) = 24xy \quad 0 \le x \le 1, \ 0 \le y \le 1, \ x + y \le 1$$

In Example 5.5 it was shown that

$$f_X(x) = 12x(1 - x)^2 \quad 0 \le x \le 1$$

The conditional pdf of Y given that $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{24xy}{12x(1-x)^2} = \frac{2y}{(1-x)^2} \qquad 0 \le y \le 1-x$$

This can be used to calculate conditional probabilities for Y. For example,

$$P(Y \le .25|X = .5) = \int_{-\infty}^{.25} f_{Y|X}(y|.5)dy = \int_{0}^{.25} \frac{2y}{(1-.5)^2}dy = \left[4y^2\right]_0^{.25} = .25$$

Given that the weight of almonds (X) is .5 lb, the probability is .25 for the weight of cashews (Y) to be less than .25 lb.

Just as in the discrete case, the conditional distribution assigns a total probability of 1 to the set of all possible Y values. That is, integrating the conditional density over its set of possible values should yield 1:

$$\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = \int_{0}^{1-x} \frac{2y}{(1-x)^2}dy = \left[\frac{y^2}{(1-x)^2}\right]_0^{1-x} = 1$$

Whenever you calculate a conditional density, we recommend doing this integration as a validity check.   ∎

Conditional Distributions and Independence

Recall that in Section 5.1 two random variables were defined to be independent if their joint pmf or pdf factors into the product of the marginal pmfs or pdfs. We can understand this definition better with the help of conditional distributions. For example, suppose there is independence in the discrete case. Then

$$p_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)} = \frac{p_X(x)p_Y(y)}{p_X(x)} = p_Y(y)$$

That is, independence implies that the conditional distribution of Y is the same as the unconditional (i.e., marginal) distribution, and that this is true no matter the value of X. The implication works in the other direction, too. If $p_{Y|X}(y|x) = p_Y(y)$, then

$$\frac{p(x,y)}{p_X(x)} = p_Y(y)$$

so $p(x, y) = p_X(x)\, p_Y(y)$, and therefore X and Y are independent.

In Example 5.7 we said that independence necessitates the region of positive density being a rectangle (possibly infinite in extent). In terms of conditional distributions, this region tells us the domain of Y for each possible x value. For independence we need to have the domain of Y (the interval of positive density) be the same for each x, implying a rectangular region.

Conditional Expectation and Variance

Because the conditional distribution is a valid probability distribution, it makes sense to define the conditional mean and variance.

DEFINITION   Let X and Y be two discrete random variables with conditional probability mass function $p_{Y|X}(y|x)$. Then the **conditional expectation** (or **conditional mean**) of Y given $X = x$ is

$$\mu_{Y|X=x} = E(Y|X = x) = \sum_y y \cdot p_{Y|X}(y|x)$$

Analogously, for two continuous rvs X and Y with conditional probability density function $f_{Y|X}(y|x)$,

$$\mu_{Y|X=x} = E(Y|X = x) = \int\limits_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x)dy$$

More generally, the conditional mean of any function $h(Y)$ is given by

$$E(h(Y)|X = x) = \begin{cases} \sum\limits_y [h(y) \cdot p_{Y|X}(y|x)] & \text{(discrete case)} \\ \int\limits_{-\infty}^{\infty} h(y) \cdot f_{Y|X}(y|x)dy & \text{(continuous case)} \end{cases}$$

In particular, the **conditional variance of Y given $X = x$** is

$$\sigma^2_{Y|X=x} = V(Y|X = x) = E[(Y - \mu_{Y|X=x})^2|X = x] = E(Y^2|X = x) - \mu^2_{Y|X=x}$$

Example 5.26 Having previously found the conditional distribution of Y given $X = 250$ in Example 5.24, let's compute the conditional mean and variance.

$$\begin{aligned} \mu_{Y|X=250} = E(Y|X = 250) &= 0p_{Y|X}(0|250) + 100p_{Y|X}(100|250) \\ &+ 200p_{Y|X}(200|250) = 0(.1) + 100(.3) + 200(.6) = 150 \end{aligned}$$

The average homeowner's policy deductible, among customers with a \$250 auto deductible, is \$150. Given that the possibilities for Y are 0, 100, and 200 and most of the probability is on the latter two values, it is reasonable that the conditional mean should be between 100 and 200.

Using the alternative (shortcut) formula for the conditional variance requires first obtaining the conditional expectation of Y^2:

$$\begin{aligned} E(Y^2|X = 250) &= 0^2 p_{Y|X}(0|250) + 100^2 p_{Y|X}(100|250) + 200^2 p_{Y|X}(200|250) \\ &= 0^2(.1) + 100^2(.3) + 200^2(.6) = 27{,}000 \end{aligned}$$

Thus,

$$\sigma^2_{Y|X=250} = V(Y|X = 250) = E(Y^2|X = 250) - \mu^2_{Y|X=250} = 27{,}000 - 150^2 = 4500.$$

Taking the square root gives $\sigma_{Y|X=250} = \$67.08$, which is in the right ballpark when we recall that the possible values of Y are 0, 100, and 200. ∎

Example 5.27 (Example 5.25 continued) Suppose a 1-lb can of mixed nuts contains .1 lbs of almonds (i.e., we know that $X = .1$). Given this information, the amount of cashews Y in the can is constrained by $0 \le y \le 1 - x = .9$, and the expected amount of cashews in such a can is

$$E(Y|X = .1) = \int_0^{.9} y \cdot f_{Y|X}(y|.1)dy = \int_0^{.9} y \cdot \frac{2y}{(1 - .1)^2}dy = .6$$

The conditional variance of Y given that $X = .1$ is

$$V(Y|X = .1) = \int_0^{.9} (y - .6)^2 \cdot f_{Y|X}(y|.1)dy = \int_0^{.9} (y - .6)^2 \cdot \frac{2y}{(1 - .1)^2}dy = .045$$

Using the aforementioned shortcut, this can also be calculated in two steps:

$$E(Y^2|X = .1) = \int_0^{.9} y^2 \cdot f_{Y|X}(y|.1)dy$$

$$= \int_0^{.9} y^2 \cdot \frac{2y}{(1 - .1)^2}dy = .405$$

$$\Rightarrow V(Y|X = .1) = .405 - (.6)^2 = .045$$

More generally, conditional on $X = x$ lbs (where $0 < x < 1$), integrals similar to those above can be used to show that the conditional mean amount of cashews is $2(1 - x)/3$, and the corresponding conditional variance is $(1 - x)^2/18$. This formula implies that the variance gets smaller as the weight of almonds (x) in a can approaches 1 lb. Does this make sense? When the weight of almonds is 1 lb, the weight of cashews is *guaranteed* to be 0, implying that the variance is 0. Indeed, Figure 5.2 shows that the set of possible y values narrows to 0 as x approaches 1. ∎

The Laws of Total Expectation and Variance

By the definition of conditional expectation, the rv Y has a conditional mean for every possible value x of the variable X. In Example 5.26, we determined the mean of Y given that $X = 250$, but a different mean would result if we conditioned on $X = 100$. For the continuous rvs in Example 5.27, every value x between 0 and 1 yielded a different conditional mean of Y (and, in fact, we even found a general formula for this conditional expectation). As it turns out, these conditional means can be related back to the *unconditional* mean of Y, i.e., μ_Y. Our next example illustrates the connection.

Example 5.28 Apartments in a certain city have $x = 0$, 1, 2, or 3 bedrooms (0 for a studio apartment), and $y = 1$, 1.5, or 2 bathrooms. The accompanying table gives the proportions of apartments for the various number of bedroom/number of bathroom combinations.

			y		
$p(x, y)$		1	1.5	2	
	0	.10	.00	.00	.1
	1	.20	.08	.02	.3
x	2	.15	.10	.15	.4
	3	.05	.05	.10	.2
		.50	.23	.27	

Let X and Y denote the number of bedrooms and bathrooms, respectively, in a randomly selected apartment in this city. The marginal distribution of Y comes from the column totals in the joint probability table, from which it is easily verified that $E(Y) = 1.385$ and $V(Y) = .179275$. The conditional distributions (pmfs) of Y given that $X = x$ for $x = 0$, 1, 2, and 3 are as follows:

$$x = 0: \quad p_{Y|X=0}(1) = 1 \quad \text{(all studio apartments have one bathroom)}$$
$$x = 1: \quad p_{Y|X=1}(1) = .667, \quad p_{Y|X=1}(1.5) = .267, \quad p_{Y|X=1}(2) = .067$$
$$x = 2: \quad p_{Y|X=2}(1) = .375, \quad p_{Y|X=2}(1.5) = .25, \quad p_{Y|X=2}(2) = .375$$
$$x = 3: \quad p_{Y|X=3}(1) = .25, \quad p_{Y|X=3}(1.5) = .25, \quad p_{Y|X=3}(2) = .50$$

From these conditional pmfs, we obtain the expected value of Y given $X = x$ for each of the four possible x values:

$$E(Y|X = 0) = 1, \quad E(Y|X = 1) = 1.2, \quad E(Y|X = 2) = 1.5, \quad E(Y|X = 3) = 1.625$$

So, on the average, studio apartments have 1 bathroom, one-bedroom apartments have 1.2 bathrooms, 2-bedroom apartments have 1.5 baths, and luxurious 3-bedroom apartments have 1.625 baths.

Now, instead of writing $E(Y|X = x)$ for some specific value x, let's consider the expected number of bathrooms for an apartment of *randomly selected* size, X. This expectation, denoted $E(Y|X)$, is itself a random variable, since it is a function of the random quantity X. Its smallest possible value is 1, which occurs when $X = 0$, and that happens with probability .1 (the sum of probabilities in the first row of the joint probability table). Similarly, the random variable $E(Y|X)$ takes on the value 1.2 with probability $p_X(1) = .3$. Continuing in this manner, the probability distribution of the rv $E(Y|X)$ is as follows:

Value of $E(Y|X)$	1	1.2	1.5	1.625
Probability of value	.1	.3	.4	.2

The expected value of this random variable, denoted $E[E(Y|X)]$, is computed by taking the weighted average of the four values of $E(Y|X = x)$ against the probabilities specified by $p_X(x)$, as suggested by the preceding table:

$$E[E(Y|X)] = 1(.1) + 1.2(.3) + 1.5(.4) + 1.625(.2) = 1.385$$

But this is exactly $E(Y)$, the expected number of bathrooms.   ∎

LAW OF TOTAL EXPECTATION   For any two random variables X and Y,

$$E[E(Y|X)] = E(Y)$$

(This is sometimes referred to as computing $E(Y)$ by means of *iterated expectation.*)

The Law of Total Expectation says that $E(Y)$ is a weighted average of the conditional means $E(Y|X = x)$, where the weights are given by the pmf or pdf of X. It is analogous to the Law of Total Probability, which describes how to find $P(B)$ as a weighted average of conditional probabilities $P(B|A_i)$.

Proof Here is the proof when both rvs are discrete; in the jointly continuous case, simply replace summation by integration and pmfs by pdfs.

$$E[E(Y|X)] = \sum_{x \in D_X} E(Y|X = x)p_X(x) = \sum_{x \in D_X}\sum_{y \in D_Y} y p_{Y|X}(y|x)p_X(x)$$

$$= \sum_{x \in D_X}\sum_{y \in D_Y} y \frac{p(x,y)}{p_X(x)} p_X(x) = \sum_{y \in D_Y} y \sum_{x \in D_X} p(x,y) = \sum_{y \in D_Y} y p_Y(y) = E(Y)   \quad ∎$$

In Example 5.28, the use of iterated expectation to compute $E(Y)$ is unnecessarily cumbersome; working from the marginal pmf of Y is more straightforward. However, there are many situations in which the distribution of a variable Y is only expressed conditional on the value of another variable X. For these so-called *hierarchical models*, the Law of Total Expectation proves very useful.

Example 5.29 A ferry goes from the left bank of a small river to the right bank once an hour. The ferry can accommodate at most two vehicles. The probability that no vehicles show up is .1, that exactly one shows up is .7, and that two or more show up is .2 (but only two can be transported). The fare paid for a vehicle depends upon its weight, and the average fare per vehicle is \$25. What is the expected fare for a single trip made by this ferry?

Let X represent the number of vehicles that show up, and let Y denote the total fare for a single trip. The conditional mean of Y, given X, is $E(Y \mid X) = 25X$. So, by the Law of Total Expectation,

$$E(Y) = E[E(Y|X)] = E[25X] = \sum_{x=0}^{2} [25x \cdot p_X(x)]$$

$$= (0)(.1) + (25)(.7) + (50)(.2) = \$27.50   \quad ∎$$

Now consider computing the *variance* of Y by conditioning on the value of X. There are two contributions to $V(Y)$. The first part is the variance of the random variable $E(Y|X)$. The second part involves the random variable $V(Y|X)$—the variance of Y as a function of X—and in particular the expected value of this random variable.

LAW OF TOTAL VARIANCE   For any two random variables X and Y,

$$V(Y) = V[E(Y|X)] + E[V(Y|X)]$$

Proving the Law of Total Variance requires some slightly clever algebra; see Exercise 90.

Example 5.30 Let's verify the Law of Total Variance for the apartment scenario of Example 5.28. The pmf of the rv $E(Y|X)$ appears in that example, from which its variance is given by

$$V[E(Y|X)] = (1 - 1.385)^2(.1) + (1.2 - 1.385)^2(.3) + (1.5 - 1.385)^2(.4) + (1.625 - 1.385)^2(.2)$$
$$= .0419$$

Recall that 1.385 is the mean of the rv $E(Y|X)$, which, by the Law of Total Expectation, is also $E(Y)$. The second term in the Law of Total Variance involves the variable $V(Y|X)$, which requires determining the conditional variance of Y given $X = x$ for $x = 0, 1, 2, 3$. Using the four conditional distributions displayed in Example 5.28, these are

$$V(Y|X = 0) = 0; \quad V(Y|X = 1) = .0933; \quad V(Y|X = 2) = .1875; \quad V(Y|X = 3) = .171875$$

The rv $V(Y|X)$ takes on these four values with probabilities .1, .3, .4, and .2, respectively (again, these are inherited from the distribution of X). Thus,

$$E[V(Y|X)] = 0(.1) + .0933(.3) + .1875(.4) + .171875(.2) = .137375$$

Combining, $V[E(Y|X)] + E[V(Y \mid X)] = .0419 + .137375 = .179275$. This is exactly $V(Y)$ computed using the marginal pmf of Y in Example 5.28, and the Law of Total Variance is verified for this example. ∎

The computation of $V(Y)$ in Example 5.30 is clearly not efficient; it is much easier, given the joint pmf of X and Y, to determine the variance of Y from its marginal pmf. As with the Law of Total Expectation, the real worth of the Law of Total Variance comes from its application to hierarchical models, where the distribution of one variable (Y, say) is only known conditional on the distribution of another rv.

Example 5.31 In the manufacture of ceramic tiles used for heat shielding, the proportion of tiles that meet the required thermal specifications varies from day to day. Let P denote the proportion of tiles meeting specifications on a randomly selected day, and suppose P can be modeled by the following pdf:

$$f(p) = 9p^8 \quad 0 < p < 1$$

At the end of each day, a random sample of $n = 20$ tiles is selected and each tile is tested. Let Y denote the number of tiles among the 20 that meet specifications; conditional on $P = p$, $Y \sim \text{Bin}(20, p)$. Find the expected number of tiles meeting thermal specifications in a daily sample of 20, and find the corresponding standard deviation.

From the properties of the binomial distribution, we know that $E(Y|P = p) = np = 20p$, so $E(Y|P) = 20P$. Applying the Law of Total Expectation,

$$E(Y) = E[E(Y|P)] = E[20P] = \int_0^1 20p \cdot f(p)dp = \int_0^1 180p^9 dp = 18$$

This is reasonable: since $E(P) = .9$ by integration, the expected proportion of good tiles is 90%, and thus the expected number of good tiles in a random sample of 20 tiles is 18.

Determining the standard deviation of Y requires the two pieces of the Law of Total Variance. First, using the rescaling property of variance,

$$V[E(Y|P)] = V(20P) = 20^2 V(P) = 400V(P)$$

The variance of P can be determined directly from the pdf of P via integration. The result is $V(P) = 9/1100$, so $V[E(Y|P)] = 400(9/1100) = 36/11$. Second, the binomial variance formula $np(1 - p)$ implies that the conditional variance of Y given P is $V(Y|P) = 20P(1 - P)$, so

$$E[V(Y|P)] = E[20P(1 - P)] = \int_0^1 20p(1 - p) \cdot 9p^8 dp = \frac{18}{11}$$

Therefore, by the Law of Total Variance,

$$V(Y) = V[E(Y|P)] + E[V(Y|P)] = \frac{36}{11} + \frac{18}{11} = \frac{54}{11} = 4.909,$$

and the standard deviation of Y is $\sigma_Y = \sqrt{4.909} = 2.22$. This "total" standard deviation accounts for two effects: day-to-day variation in quality as modeled by P (the first term in the variance expression), and random variation in the number of observed good tiles as modeled by the binomial distribution (the second term). ■

Here is an example where the Laws of Total Expectation and Variance are helpful in finding the mean and variance of a random variable that is neither discrete nor continuous.

Example 5.32 The probability of a claim being filed on an insurance policy is .1, and only one claim can be filed. If a claim is filed, the claim amount is exponentially distributed with mean $1000. Recall from Section 3.4 that $\mu = \sigma$ for an exponential rv, so the variance is the square of this value. We want to find the mean and variance of the amount paid. Let X be the number of claims (0 or 1) and let Y be the payment. We know that $E(Y|X = 0) = 0$ and $E(Y|X = 1) = 1000$. Also, $V(Y|X = 0) = 0$ and $V(Y|X = 1) = 1000^2 = 1,000,000$. Here is a table for the distribution of $E(Y|X = x)$ and $V(Y|X = x)$:

x	$P(X = x)$	$E(Y|X = x)$	$V(Y|X = x)$
0	.9	0	0
1	.1	1000	1,000,000

Therefore,

$$E(Y) = E[E(Y|X)] = E(Y|X = 0) \cdot P(X = 0) + E(Y|X = 1) \cdot P(X = 1) = 0(.9) + 1000(.1) = 100$$

The average claim amount across all customers is $100. Next, the variance of the conditional mean is

$$V[E(Y|X)] = (0 - 100)^2(.9) + (1000 - 100)^2(.1) = 90{,}000,$$

and the expected value of the conditional variance is

$$E[V(Y|X)] = 0(.9) + 1{,}000{,}000(.1) = 100{,}000$$

Now apply the Law of Total Variance to get $V(Y)$:

$$V(Y) = V[E(Y|X)] + E[V(Y|X)] = 90{,}000 + 100{,}000 = 190{,}000$$

Taking the square root gives the standard deviation, $\sigma_Y = \$434.89$.

Suppose that we want to compute the mean and variance of Y directly. Notice that X is discrete, but the conditional distribution of Y given $X = 1$ is continuous. The random variable Y itself is neither discrete nor continuous, because it has probability .9 of being 0, but the other .1 of its probability is spread out from 0 to ∞. Such "mixed" distributions may require a little extra effort to evaluate means and variances, although it is not especially hard in this case (because the discrete mass is at 0 and doesn't contribute to expectations):

$$E(Y) = (.9)(0) + (.1)\int_0^\infty y \frac{1}{1000} e^{-y/1000} dy = (.1)(1000) = 100$$

$$E(Y^2) = (.9)^2(0) + (.1)\int_0^\infty y^2 \frac{1}{1000} e^{-y/1000} dy = (.1)(2{,}000{,}000) = 200{,}000$$

$$V(Y) = E(Y^2) - [E(Y)]^2 = 200{,}000 - 10{,}000 = 190{,}000$$

These agree with what we found using the theorems.   ■

Exercises: Section 5.4 (68–90)

68. According to the 2017 CIRP report *The American Freshman*, 36.2% of first-year college students indentify as liberals, 22.4% as conservatives, and 41.4% characterize themselves as middle-of-the-road. Choose two students at random, let X be the number of liberals among the two, and let Y be the number of conservatives among the two.

    a. Using the multinomial distribution from Section 5.1, give the joint probability mass function $p(x, y)$ of X and Y and the corresponding joint probability table.

    b. Determine the marginal probability mass functions by summing $p(x, y)$ numerically. How could these be obtained directly? [*Hint*: What are the univariate distributions of X and Y?]

    c. Determine the conditional probability mass function of Y given $X = x$ for $x = 0$, 1, 2. Compare this to the binomial distribution with $n = 2 - x$ and $p = .224/.638$. Why should this work?

    d. Are X and Y independent? Explain.

    e. Find $E(Y|X = x)$ for $x = 0$, 1, 2. Do this numerically and then compare with the use of the formula for the binomial mean, using the binomial distribution given in part (c).

    f. Determine $V(Y|X = x)$ for $x = 0$, 1, 2. Do this numerically and then compare with the use of the formula for the binomial variance, using the binomial distribution given in part (c).

69. Teresa and Allison each have arrival times uniformly distributed between 12:00 and 1:00. Their times do not influence each other. If Y is the first of the two times and X is the second, on a scale of 0–1, it can be shown that the joint pdf of X and Y is $f(x, y) = 2$ for $0 < y < x < 1$.

    a. Determine the marginal density of X.
    b. Determine the conditional density of Y given $X = x$.
    c. Determine the conditional probability that Y is between 0 and .3, given that X is .5.
    d. Are X and Y independent? Explain.
    e. Determine the conditional mean of Y given $X = x$.
    f. Determine the conditional variance of Y given $X = x$.

70. Refer back to the previous exercise.

    a. Determine the marginal density of Y.
    b. Determine the conditional density of X given $Y = y$.
    c. Determine the conditional mean of X given $Y = y$.
    d. Determine the conditional variance of X given $Y = y$.

71. A pizza place has two phones. On each phone the waiting time until the first call is exponentially distributed with mean one minute. Each phone is not influenced by the other. Let X be the shorter of the two waiting times and let Y be the longer. It can be shown that the joint pdf of X and Y is $f(x, y) = 2e^{-(x+y)}, 0 < x < y < \infty$.

    a. Determine the marginal density of X.
    b. Determine the conditional density of Y given $X = x$.
    c. Determine the probability that Y is greater than 2, given that $X = 1$.
    d. Are X and Y independent? Explain.
    e. Determine the conditional mean of Y given $X = x$.
    f. Determine the conditional variance of Y given $X = x$.

72. A class has 10 mathematics majors, 6 computer science majors, and 4 statistics majors. A committee of two is selected at random to work on a problem. Let X be the number of mathematics majors, and let Y be the number of computer science majors chosen.

    a. Determine the joint probability mass function $p(x, y)$. This generalizes the hypergeometric distribution studied in Section 3.6. Give the joint probability table showing all nine values, of which three should be 0.
    b. Determine the marginal probability mass functions by summing numerically. How could these be obtained directly? [*Hint*: What are the univariate distributions of X and Y?]
    c. Determine the conditional probability mass function of Y given $X = x$ for $x = 0, 1, 2$. Compare with the hypergeometric $h(y; 2 - x, 6, 10)$ distribution. Intuitively, why should this work?
    d. Are X and Y independent? Explain.
    e. Determine $E(Y|X = x)$, $x = 0, 1, 2$. Do this numerically and then compare with the use of the formula for the hypergeometric mean, using the hypergeometric distribution given in part (c).
    f. Determine $V(Y|X = x)$, $x = 0, 1, 2$. Do this numerically and then compare with the use of the formula for the hypergeometric variance, using the hypergeometric distribution given in part (c).

73. A one-foot-long stick is broken at a point X (measured from the left end) chosen randomly uniformly along its length. Then the left part is broken at a point Y chosen randomly uniformly along its length. In other words, X is uniformly distributed between 0 and 1 and, given $X = x$, Y is uniformly distributed between 0 and x.

    a. Determine $E(Y|X = x)$ and then $V(Y|X = x)$.
    b. Determine $f(x,y)$ using $f_X(x)$ and $f_{Y|X}(y|x)$.

c. Determine $f_Y(y)$.

d. Use $f_Y(y)$ from (c) to get $E(Y)$ and $V(Y)$.

e. Use (a) and the Laws of Total Expectation and Variance to get $E(Y)$ and $V(Y)$.

74. A system consisting of two components will continue to operate only as long as both components function. Suppose the joint pdf of the lifetimes (months) of the two components in a system is given by $f(x, y) = c[10 - (x+y)]$ for $x > 0$, $y > 0$, $x+y < 10$.

 a. If the first component functions for exactly 3 months, what is the probability that the second functions for more than 2 months?

 b. Suppose the system will continue to work only as long as both components function. Among 20 of these systems that operate independently of each other, what is the probability that at least half work for more than 3 months?

75. Refer back to Exercise 1 of this chapter.

 a. Given that $X = 1$, determine the conditional pmf of Y—that is, $p_{Y|X}(0|1)$, $p_{Y|X}(1|1)$, and $p_{Y|X}(2|1)$.

 b. Given that two hoses are in use at the self-service island, what is the conditional pmf of the number of hoses in use on the full-service island?

 c. Use the result of part (b) to calculate the conditional probability $P(Y \leq 1|X = 2)$.

 d. Given that two hoses are in use at the full-service island, what is the conditional pmf of the number in use at the self-service island?

76. The joint pdf of pressures for right and left front tires is given in Exercise 11.

 a. Determine the conditional pdf of Y given that $X = x$ and the conditional pdf of X given that $Y = y$.

b. If the pressure in the right tire is found to be 22 psi, what is the probability that the left tire has a pressure of at least 25 psi? Compare this to $P(Y \geq 25)$.

c. If the pressure in the right tire is found to be 22 psi, what is the expected pressure in the left tire, and what is the standard deviation of pressure in this tire?

77. Suppose that X is uniformly distributed between 0 and 1. Given $X = x$, Y is uniformly distributed between 0 and x^2.

 a. Determine $E(Y|X = x)$ and then $V(Y|X = x)$.

 b. Determine $f(x, y)$ using $f_X(x)$ and $f_{Y|X}(y|x)$.

 c. Determine $f_Y(y)$.

78. Refer back to the previous exercise.

 a. Use $f_Y(y)$ from the previous exercise to get $E(Y)$ and $V(Y)$.

 b. Use part (a) of the previous exercise and the Laws of Total Expectation and Variance to get $E(Y)$ and $V(Y)$.

79. David and Peter independently choose at random a number from 1, 2, 3, with each possibility equally likely. Let X be the larger of the two numbers, and let Y be the smaller.

 a. Determine $p(x, y)$.

 b. Determine $p_X(x)$, $x = 1, 2, 3$.

 c. Determine $p_{Y|X}(y|x)$.

 d. Determine $E(Y|X = x)$ for $x = 1, 2, 3$.

 e. Determine $V(Y|X = x)$ for $x = 1, 2, 3$.

80. Refer back to the previous exercise. Find

 a. $E(X)$.

 b. $p_Y(y)$.

 c. $E(Y)$ using $p_Y(y)$.

 d. $E(Y)$ using $E(Y|X)$.

 e. $E(X) + E(Y)$. Why does your answer make intuitive sense?

81. Refer back to the previous two exercises. Find

    a. $p_{X|Y}(x|y)$.
    b. $E(X|Y = y)$ for $y = 1, 2, 3$.
    c. $V(X|Y = y)$ for $y = 1, 2, 3$.

82. Consider three ping-pong balls numbered 1, 2, and 3. Two balls are randomly selected with replacement. If the sum of the two resulting numbers exceeds 4, two balls are again selected. This process continues until the sum is at most 4. Let X and Y denote the last two numbers selected. Possible (X, Y) pairs are $\{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)\}$.

    a. Determine $p_{X,Y}(x,y)$.
    b. Determine $p_{Y|X}(y|x)$.
    c. Determine $E(Y|X = x)$. Is this a linear function of x?
    d. Determine $E(X|Y = y)$. What special property of $p(x, y)$ allows us to get this from (c)?
    e. Determine $V(Y|X = x)$.

83. Let X be a random digit (0, 1, 2, …, 9 are equally likely), and let Y be a random digit not equal to X. That is, the nine digits other than X are equally likely for Y.

    a. Determine $p_X(x)$, $p_{Y|X}(y|x)$, and $p_{X,Y}(x, y)$.
    b. Determine a formula for $E(Y|X = x)$.

84. Consider the situation in Example 5.29, and suppose further that the standard deviation for fares per car is $4.

    a. Find the variance of the rv $E(Y|X)$.
    b. Using Expression (5.6) from the previous section, the conditional variance of Y given $X = x$ is $4^2x = 16x$. Determine the mean of the rv $V(Y|X)$.
    c. Use the Law of Total Variance to find σ_Y, the unconditional standard deviation of Y.

85. This week the number X of claims coming into an insurance office is Poisson with mean 100. The probability that any particular claim relates to automobile insurance is .6, independent of any other claim. If Y is the number of automobile claims, then Y is binomial with X trials, each with "success" probability .6.

    a. Determine $E(Y|X = x)$ and $V(Y|X = x)$.
    b. Use part (a) to find $E(Y)$.
    c. Use part (a) to find $V(Y)$.

86. In the previous exercise, show that the distribution of Y is Poisson with mean 60. [You will need to recognize the Maclaurin series expansion for the exponential function.] Use the knowledge that Y is Poisson with mean 60 to find $E(Y)$ and $V(Y)$.

87. The heights of American men follow a normal distribution with mean 70 in. and standard deviation 3 in. Suppose that the weight distribution (lbs) for men that are x inches tall also has a normal distribution, but with mean $4x - 104$ and standard deviation $.3x - 17$. Let Y denote the weight of a randomly selected American man. Find the (unconditional) mean and standard deviation of Y.

88. A statistician is waiting behind one person to check out at a store. The checkout time for the first person, X, can be modeled by an exponential distribution with some parameter $\lambda > 0$. The statistician observes the first person's checkout time, x; being a statistician, she surmises that her checkout time Y will follow an exponential distribution with mean x.

    a. Determine $E(Y|X = x)$ and $V(Y|X = x)$.
    b. Use the Laws of Total Expectation and Variance to find $E(Y)$ and $V(Y)$.
    c. Write out the joint pdf of X and Y. [Hint: You have $f_X(x)$ and $f_{Y|X}(y|x)$.] Then write an integral expression for the marginal pdf of Y (from which, at least in theory, one could determine the mean and variance of Y). What happens?

89. In the game Plinko on the television game show *The Price is Right*, contestants have the opportunity to earn "chips" (flat, circular disks) that can be dropped down a peg board into slots labeled with cash amounts. Every contestant is given one chip automatically and can earn up to four more chips by correctly guessing the prices of certain small items. If we let p denote the probability a contestant correctly guesses the price of a prize, then the number of chips a contestant earns, X, can be modeled as $X = 1 + N$, where $N \sim \text{Bin}(4, p)$.

   a. Determine $E(X)$ and $V(X)$.
   b. For each chip, the amount of money won on the Plinko board has the following distribution:

Value	$0	$100	$500	$1,000	$10,000
Probability	.39	.03	.11	.24	.23

   Determine the mean and variance of the winnings from a single chip.
   c. Let Y denote the total winnings of a randomly selected contestant. Using results from the previous section, the conditional mean and variance of Y, given a player gets x chips, are μx and $\sigma^2 x$, respectively, where μ and σ^2 are the mean and variance for a single chip computed in (b). Find expressions for the (unconditional) mean and standard deviation of Y. [*Note:* Your answers will be functions of p.]
   d. Evaluate your answers to part (c) for $p = 0, .5$, and 1. Do these answers make sense? Explain.

90. Let X and Y be any two random variables.

   a. Show that $E[V(Y|X)] = E[Y^2] - E\left[\mu_{Y|X}^2\right]$. [*Hint:* Use the variance shortcut formula and apply the Law of Total Expectation to the first term.]
   b. Show that $V(E[Y|X]) = E\left[\mu_{Y|X}^2\right] - (E[Y])^2$. [*Hint:* Use the variance shortcut formula again; this time, apply the Law of Total Expectation to the second term.]
   c. Combine the previous two results to establish the Law of Total Variance.

5.5   The Bivariate Normal Distribution

Perhaps the most useful joint distribution is the **bivariate normal distribution**. Although the formula may seem rather complicated, it is based on a simple quadratic expression in the standardized variables (subtract the mean and then divide by the standard deviation). The bivariate normal pdf is

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \exp\left(-\frac{1}{2(1 - \rho^2)}\left[\left(\frac{x - \mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x - \mu_1}{\sigma_1}\right)\left(\frac{y - \mu_2}{\sigma_2}\right) + \left(\frac{y - \mu_2}{\sigma_2}\right)^2\right]\right)$$

for $-\infty < x < \infty$, $-\infty < y < \infty$. The notation used here for the five parameters reflects the roles they play. Some careful integration shows that μ_1 and σ_1 are the mean and standard deviation, respectively, of X; μ_2 and σ_2 are the mean and standard deviation of Y; and ρ is the correlation coefficient between the two variables. The integration required to do bivariate normal probability calculations is quite difficult. Computer code is available for calculating $P(X \leq x, Y \leq y)$ approximately using numerical integration, and some software packages (e.g., R, SAS, Stata) include this feature.

The density surface in three dimensions looks like a mountain with elliptical cross sections, as shown in Figure 5.7a. The vertical cross sections are all proportional to normal densities. If we set $f(x, y) = c$ to investigate the contours (curves along which the density is constant), this amounts to equating the exponent of the joint pdf to a constant. The contours are then concentric ellipses centered at $(x, y) = (\mu_1, \mu_2)$, as shown in Figure 5.7b.
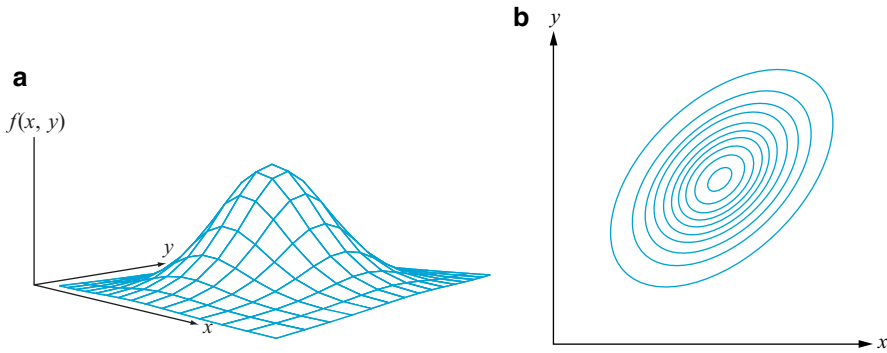
Figure 5.7 (**a**) A graph of the bivariate normal pdf; (**b**) contours of the bivariate normal pdf

If $\rho = 0$, then the bivariate normal pdf simplifies to $f(x,y) = f_X(x) f_Y(y)$, where $X \sim N(\mu_1, \sigma_1)$ and $Y \sim N(\mu_2, \sigma_2)$. That is, X and Y have independent normal distributions. (In this case, the elliptical contours reduce to circles.) Recall that in Section 5.2 we emphasized that independence of X and Y implies $\rho = 0$ but, in general, $\rho = 0$ does not imply independence. However, we have just seen that when X and Y are bivariate normal $\rho = 0$ does imply independence. Therefore, in the bivariate normal case $\rho = 0$ if and only if the two rvs are independent.

Regardless of whether or not $\rho = 0$, the marginal distribution $f_X(x)$ is just a normal pdf with mean μ_1 and standard deviation σ_1:

$$f_X(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-(x-\mu_1)^2/(2\sigma_1^2)}$$

The integration to show this [integrating $f(x, y)$ on y from $-\infty$ to ∞] is rather messy. Likewise, the marginal distribution of Y is $N(\mu_2, \sigma_2)$. These two marginal pdfs are, in fact, just special cases of a much stronger result (whose proof relies on some advanced matrix theory and will not be presented here).

THEOREM Random variables X and Y have a bivariate normal distribution *if and only if* every linear combination of X and Y is normal; i.e., the rv $aX + bY + c$ has a normal distribution for any constants a, b, c (except the case $a = b = 0$).

Example 5.33 Many students applying for college take the SAT, which consists of math and verbal components (the latter is currently called evidence-based reading and writing). Let X and Y denote the math and verbal scores, respectively, for a randomly selected student. According to the College Board, the population of students taking the exam in 2017 had the following results:

$$\mu_1 = 527, \quad \sigma_1 = 107, \quad \mu_2 = 533, \quad \sigma_2 = 100, \quad \rho = .77$$

Suppose that X and Y have approximately (because both X and Y are discrete) a bivariate normal distribution. Let's determine the probability that a student's total score across these two components exceeds 1250, the minimum admission score for a particular university.

Our goal is to calculate $P(X + Y > 1250)$. Using the bivariate normal pdf, the desired probability is a daunting double integral:

$$\frac{1}{2\pi(107)(100)\sqrt{1-.77^2}} \int_{-\infty}^{\infty} \int_{1250-y}^{\infty} e^{-\left\{[(x-527)/107]^2 - 2(.77)(x-527)(y-533)/(107)(100) + [(y-533)/100]^2\right\}/[2(1-.77^2)]} dxdy$$

This is not a practical way to solve this problem! Instead, recognize $X + Y$ as a linear combination of X and Y; by the preceding theorem, $X + Y$ has a normal distribution. The mean and variance of $X + Y$ are calculated using the formulas from Section 5.3:

$$E(X + Y) = E(X) + E(Y) = \mu_1 + \mu_2 = 527 + 533 = 1060$$
$$V(X + Y) = V(X) + V(Y) + 2\text{Cov}(X, Y)$$
$$= \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2 = 107^2 + 100^2 + 2(.77)(107)(100) = 37{,}927$$

Therefore, $P(X + Y > 1250) = 1 - \Phi\left(\frac{1250-1060}{\sqrt{37,927}}\right) \approx 1 - \Phi(.98) = .1635$.

Suppose instead we wish to determine $P(X < Y)$, the probability a student scores lower on math than on reading. If we rewrite this probability as $P(X - Y < 0)$, then we may apply the preceding theorem to the linear combination $X - Y$. With $E(X - Y) = -6$ and $V(X - Y) = 4971$,

$$P(X < Y) = P(X - Y < 0) = \Phi\left(\frac{0 - (-6)}{\sqrt{4971}}\right) \approx \Phi(.09) = .5359 \qquad \blacksquare$$

Independent Normal Random Variables

As alluded to earlier in this section, if X and Y are *independent* normal rvs then the joint distribution of X and Y is trivially bivariate normal (specifically with $\rho = 0$). In Section 5.3, we proved that any linear combination of independent normal rvs is itself normally distributed, which comports with the earlier theorem in this section. In fact, we can generalize to the case of two linear combinations of independent normal rvs.

PROPOSITION   Let U and V be linear combinations of the independent normal rvs $X_1, \ldots, X_n$. Then the joint distribution of U and V is bivariate normal. The converse is also true: if U and V have a bivariate normal distribution, then they can be expressed as linear combinations of independent normal rvs.

The proof uses the methods of the next section together with a little matrix theory.

Example 5.34 How can we simulate bivariate normal rvs with a specified correlation ρ? Let Z_1 and Z_2 be independent standard normal rvs (which can be generated using software, or by applying the Box–Muller method described in Exercise 107), and define two new variables

$$U = Z_1 \quad V = \rho \cdot Z_1 + \sqrt{1 - \rho^2} \cdot Z_2$$

Then U and V are linear combinations of independent normal rvs, so their joint distribution is bivariate normal by the preceding proposition. It can be shown (Exercise 129) that U and V each have mean 0 and standard deviation 1, and Corr$(U, V) = \rho$.

Now suppose we wish to simulate from a bivariate normal distribution with an arbitrary set of parameters μ_1, σ_1, μ_2, σ_2, and ρ. Define X and Y by

$$X = \mu_1 + \sigma_1 U = \mu_1 + \sigma_1 Z_1, \quad Y = \mu_2 + \sigma_2 V = \mu_2 + \sigma_2(\rho Z_1 + \sqrt{1 - \rho^2} Z_2) \qquad (5.7)$$

Since X and Y in Expression (5.7) are linear functions of U and V, it follows from Section 5.2 that Corr$(X, Y) = $ Corr$(U, V) = \rho$. Moreover, since $\mu_U = \mu_V = 0$ and $\sigma_U = \sigma_V = 1$, these linear transformations give X and Y the desired means and standard deviations. So, to simulate a bivariate normal distribution, create a pair of independent standard normal variates z_1 and z_2, and then apply the formulas for X and Y in Expression (5.7). (Notice also that we've just proved the "converse" part of the foregoing proposition.) ■

Conditional Distributions of X and Y

The conditional density of Y given $X = x$ results from dividing the marginal density of X into $f(x, y)$. The algebra is again tedious, but the result is fairly simple.

PROPOSITION   Let X and Y have a bivariate normal distribution. Then the conditional distribution of Y, given $X = x$, is normal with mean and variance

$$\mu_{Y|X=x} = E(Y|X = x) = \mu_2 + \rho\sigma_2 \frac{x - \mu_1}{\sigma_1}$$
$$\sigma_{Y|X=x}^2 = V(Y|X = x) = \sigma_2^2(1 - \rho^2)$$

Notice that the conditional mean of Y is a linear function of x, and the conditional variance of Y doesn't depend on x at all. When $\rho = 0$, the conditional mean is the mean of Y, μ_2, and the conditional variance is just the variance of Y, σ_2^2. In other words, if $\rho = 0$, then the conditional distribution of Y is the same as the unconditional distribution of Y. When ρ is close to 1 or –1 the conditional variance will be much smaller than $V(Y)$, which says that knowledge of X will be very helpful in predicting Y. If ρ is near 0 then X and Y are nearly independent and knowledge of X is not very useful in predicting Y.

Example 5.35 Let X and Y be the heights of a randomly selected mother and her daughter, respectively. A similar situation was one of the first applications of the bivariate normal distribution, by Francis Galton in 1886, and the data was found to fit the distribution very well. Suppose a bivariate normal distribution with mean $\mu_1 = 64$ in. and standard deviation $\sigma_1 = 3$ in. for X and mean $\mu_2 = 65$ in. and standard deviation $\sigma_2 = 3$ in. for Y. Here $\mu_2 > \mu_1$, which is in accord with the increase in height from one generation to the next. Assume $\rho = .4$. Then

$$\mu_{Y|X=x} = \mu_2 + \rho\sigma_2 \frac{x - \mu_1}{\sigma_1} = 65 + .4(3)\frac{x - 64}{3} = 65 + .4(x - 64) = .4x + 39.4$$

$$\sigma^2_{Y|X=x} = V(Y|X=x) = \sigma^2_2(1-\rho^2) = 9\big(1-.4^2\big) = 7.56 \text{ and } \sigma_{Y|X=x} = 2.75.$$

Notice that the conditional variance is 16% less than the variance of Y. Squaring the correlation gives the percentage by which the conditional variance is reduced relative to the variance of Y.   ∎

Regression to the Mean

The formula for the conditional mean can be re-expressed as

$$\frac{\mu_{Y|X=x} - \mu_2}{\sigma_2} = \rho \cdot \frac{x - \mu_1}{\sigma_1}$$

In words, when the formula is expressed in terms of standardized variables, the standardized conditional mean is just ρ times the standardized x. In particular, for the height scenario

$$\frac{\mu_{Y|X=x} - 65}{3} = .4 \cdot \frac{x - 64}{3}$$

If the mother is 5 in. above the mean of 64 in. for mothers, then the daughter's conditional expected height is just 2 in. above the mean for daughters. In this example, with equal standard deviations for Y and X, the daughter's conditional expected height is always closer to its mean than the mother's height is to its mean. One can think of the conditional expectation as falling back toward the mean, and that is why Galton called this *regression to the mean*.

Regression to the mean occurs in many contexts. For example, let X be a baseball player's average for the first half of the season and let Y be the average for the second half. Most of the players with a high X (say, above .300) will not have such a high Y. The same kind of reasoning applies to the "sophomore jinx," which says that if a player has a very good first season, then the player is unlikely to do as well in the second season.

The Multivariate Normal Distribution

The multivariate normal distribution extends the bivariate normal distribution to situations involving models for n random variables X_1, X_2, …, X_n with $n > 2$. The joint density function is quite complicated; the only way to express it compactly is to make use of matrix algebra notation, and probability calculations based on this distribution are extremely complex. Here are some of the most important properties of the distribution:

- The distribution of any linear combination of X_1, X_2, …, X_n is normal.
- The marginal distribution of any X_i is normal.
- The joint distribution of any pair X_i, X_j is bivariate normal.
- The conditional distribution of any X_i, given values of the other $n - 1$ variables, is normal.

Many procedures for the analysis of multivariate data (observations simultaneously on three or more variables) are based on assuming that the data was selected from a multivariate normal distribution. The book by Rencher and Christensen (see the bibliography) provides more information on multivariate analysis and the multivariate normal distribution.

Exercises: Section 5.5 (91–100)

91. For a few years, the SAT consisted of three components: writing, critical reading, and mathematics. Let W = SAT Writing score and X = SAT Critical Reading score for a randomly selected student. According to the College Board, in 2012 W had mean 488 and standard deviation 114, while X had mean 496 and standard deviation 114. Suppose X and W have a bivariate normal distribution with Corr(X, W) = .5.

    a. An English department plans to use $X + W$, a student's total score on the nonmath sections of the SAT, to help determine admission. Determine the distribution of $X + W$.
    b. Calculate $P(X + W > 1200)$.
    c. Suppose the English department wishes to admit only those students who score in the top 10% on this Critical Reading + Writing criterion. What combined score separates the top 10% of students from the rest?

92. Refer to the previous exercise. Let Y = SAT Mathematics score, which had mean 514 and standard deviation 117 in the year 2012. Let $T = W + X + Y$, a student's grand total score on the three components of the SAT.

    a. Find the expected value of T.
    b. Assume Corr(W, Y) = .2. and Corr(X, Y) = .25. Find the variance of T. [*Hint:* Use Expression (5.5) from Section 5.3.]
    c. Suppose W, X, Y have a multivariate normal distribution, in which case T is also normally distributed. Determine $P(T > 2000)$.
    d. What is the 99th percentile of SAT grand total scores, according to this model?

93. Let X = height (inches) and Y = weight (lbs) for an American male. Suppose X and Y have a bivariate normal distribution, the

mean and sd of heights are 70 in and 3 in, the mean and sd of weights are 170 lbs and 20 lbs, and ρ = .9.

    a. Determine the distribution of Y given $X = 68$, i.e., the weight distribution for 5′8″ American males.
    b. Determine the distribution of Y given $X = 70$, i.e., the weight distribution for 5′10″ American males. In what ways is this distribution similar to that of part (a), and how are they different?
    c. Calculate $P(Y < 180 | X = 72)$, the probability that a 6-foot-tall American male weighs less than 180 lb.

94. In electrical engineering, the unwanted "noise" in voltage or current signals is often modeled by a Gaussian (i.e., normal) distribution. Suppose that the noise in a particular voltage signal has a constant mean of 0.9 V, and that two noise instances sampled τ seconds apart have a bivariate normal distribution with covariance equal to $0.04e^{-|\tau|/10}$. Let X and Y denote the noise at times 3 s and 8 s, respectively.

    a. Determine Cov(X, Y).
    b. Determine σ_X and σ_Y. [*Hint:* $V(X) =$ Cov(X, X).]
    c. Determine Corr(X, Y).
    d. Find the probability we observe greater voltage noise at time 3 s than at time 8 s.
    e. Find the probability that the voltage noise at time 3 s is more than 1 V above the voltage noise at time 8 s.

95. For a Calculus I class, the final exam score Y and the average X of the four earlier tests have a bivariate normal distribution with mean $\mu_1 = 73$, standard deviation $\sigma_1 = 12$, mean $\mu_2 = 70$, standard deviation $\sigma_2 = 15$. The correlation is ρ = .71. Determine

    a. $\mu_{Y|X=x}$
    b. $\sigma^2_{Y|X=x}$
    c. $\sigma_{Y|X=x}$

d. $P(Y > 90|X = 80)$, i.e., the probability that the final exam score exceeds 90 given that the average of the four earlier tests is 80.

96. Refer to the previous exercise. Suppose a student's Calculus I grade is determined by $4X + Y$, the total score across five tests.

   a. Find the mean and standard deviation of $4X + Y$.
   b. Determine $P(4X + Y < 320)$.
   c. Suppose the instructor sets the curve in such a way that the top 15% of students, based on total score across the five tests, will receive As. What point total is required to get an A in Calculus I?

97. Let X and Y, reaction times (sec) to two different stimuli, have a bivariate normal distribution with mean $\mu_1 = 20$ and standard deviation $\sigma_1 = 2$ for X and mean $\mu_2 = 30$ and standard deviation $\sigma_2 = 5$ for Y. Assume $\rho = .8$. Determine

   a. $\mu_{Y|X=x}$
   b. $\sigma^2_{Y|X=x}$
   c. $\sigma_{Y|X=x}$
   d. $P(Y > 46 \mid X = 25)$

98. Refer to the previous exercise.

   a. One researcher is interested in $X + Y$, the total reaction time to the two stimuli. Determine the mean and standard deviation of $X + Y$.
   b. If X and Y were independent, what would be the standard deviation of

$X + Y$? Explain why it makes sense that the sd in part (a) is much larger than this.
   c. Another researcher is interested in $Y - X$, the difference in the reaction times to the two stimuli. Determine the mean and standard deviation of $Y - X$.
   d. If X and Y were independent, what would be the standard deviation of $Y - X$? Explain why it makes sense that the sd in part (c) is much smaller than this.

99. Let X and Y be the times for a randomly selected individual to complete two different tasks, and assume that (X, Y) has a bivariate normal distribution with $\mu_1 = 100$, $\sigma_1 = 50$, $\mu_2 = 25$, $\sigma_2 = 5$, $\rho = .4$. From statistical software we obtain $P(X < 100, Y < 25) = .3333$, $P(X < 50, Y < 20) = .0625$, $P(X < 50, Y < 25) = .1274$, and $P(X < 100, Y < 20) = .1274$.

   a. Determine $P(50 < X < 100, 20 < Y < 25)$.
   b. Leave the other parameters the same but change the correlation to $\rho = 0$ (independence). Now re-compute the probability in part (a). Intuitively, why should the original be larger?

100. One of the propositions of this section gives an expression for $E(Y|X = x)$.

   a. By reversing the roles of X and Y give a similar formula for $E(X|Y = y)$.
   b. Both $E(Y|X = x)$ and $E(X|Y = y)$ are linear functions. Show that the product of the two slopes is ρ^2.

5.6   Transformations of Multiple Random Variables

In Chapter 4 we discussed the problem of starting with a single random variable X, forming some function of X, such as $Y = X^2$ or $Y = e^X$, and investigating the distribution of this new random variable Y. We now generalize this scenario by starting with more than a single random variable. Consider as an example a system having a component that can be replaced just once before the system itself expires. Let X_1 denote the lifetime of the original component and X_2 the lifetime of the replacement component. Then any of the following functions of X_1 and X_2 may be of interest to an investigator:

1. The total lifetime, $X_1 + X_2$.
2. The ratio of lifetimes X_1/X_2 (for example, if the value of this ratio is 2, the original component lasted twice as long as its replacement).
3. The ratio $X_1/(X_1 + X_2)$, which represents the proportion of system lifetime during which the original component operated.

The Joint Distribution of Two New Random Variables

Given two random variables X_1 and X_2, consider forming two new random variables $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$. (Since most applications assume that the X_i's are continuous, we restrict ourselves to that case.) Our focus is on finding the joint distribution of these two new variables. The $u_1(\cdot)$ and $u_2(\cdot)$ functions express the new variables in terms of the original ones. The upcoming general result presumes that these functions can be inverted to solve for the original variables in terms of the new ones:

$$X_1 = v_1(Y_1, Y_2), \quad X_2 = v_2(Y_1, Y_2)$$

For example, if

$$y_1 = x_1 + x_2 \quad \text{and} \quad y_2 = \frac{x_1}{x_1 + x_2}$$

then multiplying y_2 by y_1 gives an expression for x_1, and then we can substitute this into the expression for y_1 and solve for x_2:

$$x_1 = y_1 y_2 = v_1(y_1, y_2) \quad x_2 = y_1(1 - y_2) = v_2(y_1, y_2)$$

Finally, let $f(x_1, x_2)$ denote the joint pdf of the two original variables, let $g(y_1, y_2)$ denote the joint pdf of the two *new* variables, and define two sets S and T by

$$S = \{(x_1, x_2) : f(x_1, x_2) > 0\} \quad T = \{(y_1, y_2) : g(y_1, y_2) > 0\}$$

That is, S is the region of positive density for the original variables and T is the region of positive density for the new variables; T is the "image" of S under the transformation.

TRANSFORMATION THEOREM (bivariate case)

Suppose that the partial derivative of each $v_i(y_1, y_2)$ with respect to both y_1 and y_2 exists and is continuous for every $(y_1, y_2) \in T$. Form the 2×2 matrix

$$\mathbf{M} = \begin{pmatrix} \frac{\partial v_1(y_1, y_2)}{\partial y_1} & \frac{\partial v_1(y_1, y_2)}{\partial y_2} \\ \frac{\partial v_2(y_1, y_2)}{\partial y_1} & \frac{\partial v_2(y_1, y_2)}{\partial y_2} \end{pmatrix}$$

The determinant of this matrix, called the *Jacobian*, is

$$\det(\mathbf{M}) = \frac{\partial v_1}{\partial y_1} \cdot \frac{\partial v_2}{\partial y_2} - \frac{\partial v_1}{\partial y_2} \cdot \frac{\partial v_2}{\partial y_1}$$

The joint pdf for the new variables then results from taking the joint pdf $f(x_1, x_2)$ for the original variables, replacing x_1 and x_2 by their expressions in terms of y_1 and y_2, and finally multiplying this by the absolute value of the Jacobian:

$$g(y_1, y_2) = f(v_1(y_1, y_2), v_2(y_1, y_2)) \cdot |\det(\mathbf{M})| \quad (y_1, y_2) \in T$$

The theorem can be rewritten slightly by using the notation

$$\det(\mathbf{M}) = \left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right|$$

Then we have

$$g(y_1, y_2) = f(x_1, x_2) \cdot \left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right|,$$

which is the natural extension of the univariate Transformation Theorem $f_Y(y) = f_X(x) \cdot |dx/dy|$ discussed in Chapter 4.

Example 5.36 Continuing with the component lifetime situation, suppose that X_1 and X_2 are independent, each having an exponential distribution with parameter λ. Let's determine the joint pdf of

$$Y_1 = u_1(X_1, X_2) = X_1 + X_2 \quad \text{and} \quad Y_2 = u_2(X_1, X_2) = \frac{X_1}{X_1 + X_2}$$

We have already inverted this transformation:

$$x_1 = v_1(y_1, y_2) = y_1 y_2 \quad x_2 = v_2(y_1, y_2) = y_1(1 - y_2)$$

The image of the transformation, i.e., the set of (y_1, y_2) pairs with positive density, is $y_1 > 0$ and $0 < y_2 < 1$. The four relevant partial derivatives are

$$\frac{\partial v_1}{\partial y_1} = y_2 \qquad \frac{\partial v_1}{\partial y_2} = y_1 \qquad \frac{\partial v_2}{\partial y_1} = 1 - y_2 \qquad \frac{\partial v_2}{\partial y_2} = -y_1$$

from which the Jacobian is $\det(\mathbf{M}) = -y_1 y_2 - y_1(1 - y_2) = -y_1$.

Since the joint pdf of X_1 and X_2 is

$$f(x_1, x_2) = \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} = \lambda^2 e^{-\lambda(x_1 + x_2)} \quad x_1 > 0, \quad x_2 > 0$$

we have, by the Transformation Theorem,

$$g(y_1, y_2) = \lambda^2 e^{-\lambda y_1} \cdot y_1 = \lambda^2 y_1 e^{-\lambda y_1} \cdot 1 \quad y_1 > 0, \quad 0 < y_2 < 1$$

In the last step, we've factored the joint pdf into two parts: the first part is a gamma pdf with parameters $\alpha = 2$ and $\beta = 1/\lambda$, and the second part is a uniform pdf on $(0, 1)$. Since the pdf factors and the region of positive density is rectangular, we have discovered that

- The distribution of system lifetime $X_1 + X_2$ is gamma (with $\alpha = 2$, $\beta = 1/\lambda$);
- The distribution of the proportion of system lifetime during which the original component functions is uniform on $(0, 1)$; and
- $Y_1 = X_1 + X_2$ and $Y_2 = X_1 / (X_1 + X_2)$ are independent of each other. ∎

In the foregoing example, because the joint pdf factored into one pdf involving y_1 alone and another pdf involving y_2 alone, the individual (i.e., marginal) pdfs of the two new variables were obtained from the joint pdf without any further effort. Often this will not be the case—that is, Y_1 and Y_2 will not be independent. Then to obtain the marginal pdf of Y_1, the joint pdf must be integrated over all values of the second variable.

In fact, in many applications an investigator wishes to obtain the distribution of a single function $Y_1 = u_1(X_1, X_2)$ of the original variables. To accomplish this, a second function $Y_2 = u_2(X_1, X_2)$ is created, the joint pdf is obtained, and then y_2 is integrated out. There are of course many ways to select the second function. The choice should be made so that the transformation can be easily inverted *and* the subsequent integration is straightforward.

Example 5.37 Consider a rectangular coordinate system with a horizontal x_1-axis and a vertical x_2-axis as shown in Figure 5.8a.
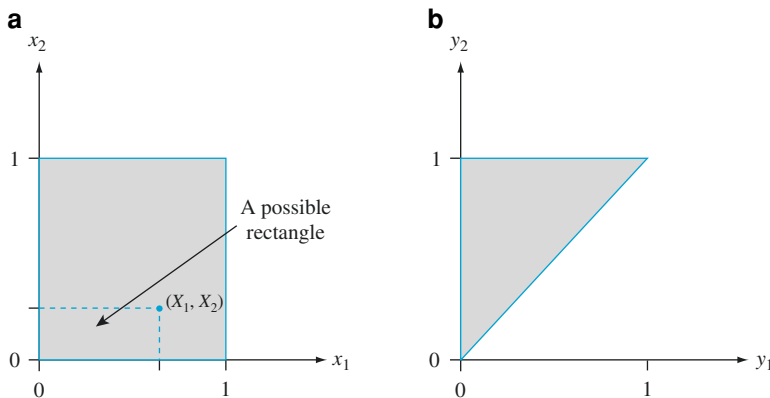
Figure 5.8 Regions of positive density for Example 5.37

First a point (X_1, X_2) is randomly selected, where the joint pdf of X_1, X_2 is

$$f(x_1, x_2) = x_1 + x_2 \quad 0 < x_1 < 1,\ 0 < x_2 < 1$$

Then a rectangle with vertices $(0, 0)$, $(X_1, 0)$, $(0, X_2)$, and (X_1, X_2) is formed as shown in Figure 5.8a. What is the distribution of X_1X_2, the area of this rectangle? Define

$$Y_1 = u_1(X_1, X_2) = X_1X_2 \quad \text{and} \quad Y_2 = u_2(X_1, X_2) = X_2$$

The inverse of this transformation is easily obtained:

$$x_1 = v_1(y_1, y_2) = \frac{y_1}{y_2} \quad \text{and} \quad x_2 = v_2(y_1, y_2) = y_2$$

Notice that because x_2 ($=y_2$) is between 0 and 1 and y_1 is the product of the two x_i's, it must be the case that $0 < y_1 < y_2$. The region of positive density for the new variables is then $T = \{(y_1, y_2):\ 0 < y_1 < y_2,\ 0 < y_2 < 1\}$, the triangular region shown in Figure 5.8b.

Since $\partial v_2 / \partial y_1 = 0$, the product of the two off-diagonal elements in the matrix $\mathbf{M}$ will be 0, so only the two diagonal elements contribute to the Jacobian:

$$\mathbf{M} = \begin{pmatrix} 1/y_2 & y_1/y_2^2 \\ 0 & 1 \end{pmatrix} \qquad \Rightarrow \qquad \det(\mathbf{M}) = \frac{1}{y_2}$$

The joint pdf of the two new variables is now

$$g(y_1, y_2) = f\left(\frac{y_1}{y_2}, y_2\right) \cdot |\det(\mathbf{M})| = \left(\frac{y_1}{y_2} + y_2\right) \cdot \frac{1}{y_2} \quad 0 < y_1 < y_2 < 1$$

Finally, to obtain the marginal pdf of Y_1 alone, we must now fix y_1 at some arbitrary value between 0 and 1, and integrate out y_2. Figure 5.8b shows that for any value of y_1, the values of y_2 range from y_1 to 1:

$$g_1(y_1) = \int_{y_1}^{1} \left(\frac{y_1}{y_2} + y_2\right) \cdot \frac{1}{y_2} dy_2 = 2(1 - y_1) \quad 0 < y_1 < 1$$

This marginal pdf can now be integrated to obtain any desired probability involving the area. For example, integrating from 0 to .5 gives $P(Y_1 < .5) = .75$. ∎

The Joint Distribution of More Than Two New Variables
Consider now starting with three random variables X_1, X_2, and X_3, and forming three new variables Y_1, Y_2, and Y_3. Suppose again that the transformation can be inverted to express the original variables in terms of the new ones:

$$x_1 = v_1(y_1, y_2, y_3), \quad x_2 = v_2(y_1, y_2, y_3), \quad x_3 = v_3(y_1, y_2, y_3)$$

Then the foregoing theorem can be extended to this new situation. The Jacobian matrix has dimension 3×3, with the entry in the ith row and jth column being $\partial v_i / \partial y_j$. The joint pdf of the new variables results from replacing each x_i in the original pdf $f(\cdot)$ by its expression in terms of the y_js and multiplying by the absolute value of the Jacobian.

Example 5.38 Consider $n = 3$ identical components with independent lifetimes X_1, X_2, X_3, each having an exponential distribution with parameter λ. If the first component is used until it fails, replaced by the second one which remains in service until it fails, and finally the third component is used until failure, then the total lifetime of these components is $Y_3 = X_1 + X_2 + X_3$. (This design structure, where one component is replaced by the next in succession, is called a *standby system.*) To find the distribution of total lifetime, let's first define two other new variables: $Y_1 = X_1$ and $Y_2 = X_1 + X_2$ (so that $Y_1 < Y_2 < Y_3$). After finding the joint pdf of all three variables, we integrate out the first two variables to obtain the desired information. Solving for the old variables in terms of the new gives

$$x_1 = y_1 \quad x_2 = y_2 - y_1 \quad x_3 = y_3 - y_2$$

It is obvious by inspection of these expressions that the three diagonal elements of the Jacobian matrix are all 1s and that the elements above the diagonal are all 0s, so the determinant is 1, the product of the diagonal elements. Since

$$f(x_1, x_2, x_3) = \lambda^3 e^{-\lambda(x_1 + x_2 + x_3)} \qquad x_1 > 0, x_2 > 0, x_3 > 0$$

by substitution,

$$g(y_1, y_2, y_3) = \lambda^3 e^{-\lambda y_3} \qquad 0 < y_1 < y_2 < y_3$$

Integrating this joint pdf first with respect to y_1 between 0 and y_2 and then with respect to y_2 between 0 and y_3 (try it!) gives

$$g_3(y_3) = \frac{\lambda^3}{2} y_3^2 e^{-\lambda y_3} \qquad y_3 > 0$$

which is the gamma pdf with $\alpha = 3$ and $\beta = 1/\lambda$. This result is a special case of the last proposition from Section 5.3, stating that the sum of n iid exponential rvs has a gamma distribution with $\alpha = n$. ∎

Exercises: Section 5.6 (101–108)

101. Let X_1 and X_2 be independent, standard normal rvs.

   a. Define $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Determine the joint pdf of Y_1 and Y_2.
   b. Determine the marginal pdf of Y_1. [*Note:* We know the sum of two independent normal rvs is normal, so you can check your answer against the appropriate normal pdf.]
   c. Are Y_1 and Y_2 independent?

102. Consider two components whose lifetimes X_1 and X_2 are independent and exponentially distributed with parameters λ_1 and λ_2, respectively. Obtain the joint pdf of total lifetime $X_1 + X_2$ and the proportion of total lifetime $X_1/(X_1 + X_2)$ during which the first component operates.

103. Let X_1 denote the time (hr) it takes to perform a first task and X_2 denote the time it takes to perform a second one. The second task always takes at least as long to perform as the first task. The joint pdf of these variables is

$$f(x_1, x_2) = 2(x_1 + x_2) \qquad 0 \le x_1 \le x_2 \le 1$$

   a. Obtain the pdf of the total completion time for the two tasks.
   b. Obtain the pdf of the difference $X_2 - X_1$ between the longer completion time and the shorter time.

104. An exam consists of a problem section and a short-answer section. Let X_1 denote the amount of time (hr) that a student spends on the problem section and X_2 represent the amount of time the same student spends on the short-answer section. Suppose the joint pdf of these two times is

$$f(x_1, x_2) = cx_1 x_2$$

$$x_1/3 < x_2 < x_1/2, 0 < x_1 < 1$$

   a. What is the value of c?
   b. If the student spends exactly .25 h on the short-answer section, what is the probability that at most .60 h was spent on the problem section? [*Hint*: First obtain the relevant conditional distribution.]
   c. What is the probability that the amount of time spent on the problem part of the exam exceeds the amount of time spent on the short-answer part by at least .5 h?

d. Obtain the joint distribution of $Y_1 = X_2/X_1$, the ratio of the two times, and $Y_2 = X_2$. Then obtain the marginal distribution of the ratio.

105. Consider randomly selecting a point (X_1, X_2, X_3) in the unit cube according to the joint pdf

$$f(x_1, x_2, x_3) = 8x_1x_2x_3 \quad 0 < x_1 < 1,$$
$$0 < x_2 < 1, \quad 0 < x_3 < 1$$

Then form a rectangular solid whose vertices are $(0, 0, 0)$, $(X_1, 0, 0)$, $(0, X_2, 0)$, $(X_1, X_2, 0)$, $(0, 0, X_3)$, $(X_1, 0, X_3)$, $(0, X_2, X_3)$, and (X_1, X_2, X_3). The volume of this solid is $Y_3 = X_1X_2X_3$. Obtain the pdf of Y_3. [*Hint*: Let $Y_1 = X_1$ and $Y_2 = X_1X_2$.]

106. Let X_1 and X_2 be independent, each having a standard normal distribution. The pair (X_1, X_2) corresponds to a point in a two-dimensional coordinate system. Consider now changing to polar coordinates via the transformation

$$Y_1 = X_1^2 + X_2^2$$

$$Y_2 = \begin{cases} \arctan\left(\frac{X_2}{X_1}\right) & X_1 > 0, X_2 \geq 0 \\ \arctan\left(\frac{X_2}{X_1}\right) + 2\pi & X_1 > 0, X_2 < 0 \\ \arctan\left(\frac{X_2}{X_1}\right) + \pi & X_1 < 0 \\ 0 & X_1 = 0 \end{cases}$$

from which $X_1 = \sqrt{Y_1}\cos(Y_2)$, $X_2 = \sqrt{Y_1}\sin(Y_2)$. Obtain the joint pdf of the new variables and then the marginal distribution of each one. [*Note*: It would be preferable to let $Y_2 = \arctan(X_2/X_1)$, but in

order to insure invertibility of the arctan function, it is defined to take on values only between $-\pi/2$ and $\pi/2$. Our specification of Y_2 allows it to assume any value between 0 and 2π.]

107. The result of the previous exercise suggests how observed values of two independent standard normal variables can be generated by first generating their polar coordinates with an exponential rv with $\lambda = \frac{1}{2}$ and an independent Unif$(0, 2\pi)$ rv: Let U_1 and U_2 be independent Unif$(0, 1)$ rvs, and then let

$$Y_1 = -2\ln(U_1) \quad Y_2 = 2\pi U_2$$
$$Z_1 = \sqrt{Y_1}\cos(Y_2) \quad Z_2 = \sqrt{Y_1}\sin(Y_2)$$

Show that the Z_i's are independent standard normal. [*Note*: This is called the *Box–Muller transformation* after the two individuals who discovered it. Now that statistical software packages will generate almost instantaneously observations from a normal distribution with any mean and variance, it is thankfully no longer necessary for people like you and us to carry out the transformations just described—let the software do it!]

108. Let X_1 and X_2 be independent random variables, each having a standard normal distribution. Show that the pdf of the ratio $Y = X_1/X_2$ is given by $f(y) = 1/[\pi(1 + y^2)]$ for $-\infty < y < \infty$. (This is called the *standard Cauchy distribution*; its density curve is bell-shaped, but the tails are so heavy that μ does not exist.)

5.7   Order Statistics

Many statistical procedures involve ordering the sample observations from smallest to largest and then manipulating these ordered values in various ways. For example, the sample median is either the middle value in the ordered list or the average of the two middle values depending on whether the sample size n is odd or even. The sample range is the difference between the largest and smallest values. And a trimmed mean results from deleting the same number of observations from each end of the ordered list and averaging the remaining values.

Throughout this section, we assume that we have a collection of rvs $X_1, X_2, \ldots, X_n$ with the following properties:

1. The X_i's are independent rvs.
2. Every X_i has the same probability distribution (e.g., they all follow an exponential distribution with the same parameter λ).
3. The distribution shared by the X_i's is continuous, with cumulative distribution function $F(x)$ and density function $f(x)$.

Assumptions 1 and 2 can be paraphrased by saying that the X_i's are a **random sample** from the specified distribution. The continuity assumption in 3 implies that $P(X_i = X_j) = 0$ for $i \neq j$; thus, with probability 1, the n sample observations will all be distinct (no ties). Of course, in practice all measuring instruments have accuracy limitations, so tied values may in fact result.

DEFINITION   The **order statistics** from a random sample are the random variables $Y_1, \ldots Y_n$ given by

Y_1 = the smallest among $X_1, X_2, \ldots, X_n$ (i.e., the sample minimum)
Y_2 = the second smallest among $X_1, X_2, \ldots, X_n$

$$\vdots$$

Y_n = the largest among $X_1, X_2, \ldots, X_n$ (the sample maximum)

Thus, with probability 1, $Y_1 < Y_2 < \cdots < Y_{n-1} < Y_n$.

The sample median is then $Y_{(n+1)/2}$ when n is odd, the sample range is $Y_n - Y_1$, and for $n = 10$ the 20% trimmed mean is $\sum_{i=3}^{8} Y_i / 6$. The order statistics are defined as random variables (hence the use of uppercase letters); observed values are denoted by $y_1, \ldots, y_n$.

The Distributions of Y_n and Y_1

The key idea in obtaining the distribution of the sample maximum Y_n is the observation that Y_n is at most y if and only if every one of the X_i's is at most y. Similarly, the distribution of Y_1 is based on the fact that it will exceed y if and only if all X_i's exceed y.

Example 5.39 Consider 5 identical components connected in parallel, as illustrated in Figure 5.9a.
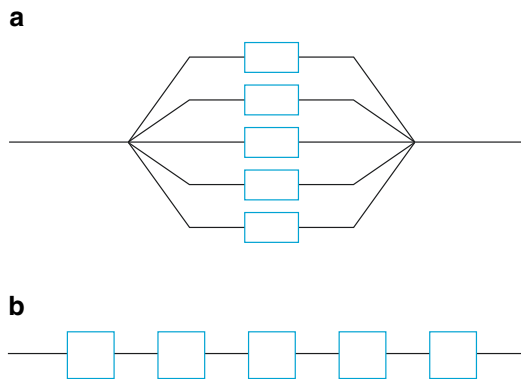
Figure 5.9 Systems of components for Example 5.39: (**a**) parallel connection; (**b**) series connection

Let X_i denote the lifetime, in hours, of the ith component ($i = 1, 2, 3, 4, 5$). Suppose that the X_i's are independent and that each has an exponential distribution with $\lambda = .01$, so the expected lifetime of any particular component is $1/\lambda = 100$ h. Because of the parallel configuration, the system will continue to function as long as at least one component is still working, and will fail as soon as the last component functioning ceases to do so. That is, the system lifetime is Y_5, the largest order statistic in a sample of size 5 from the specified exponential distribution. Now Y_5 will be at most y if and only if every one of the five X_i's is at most y. With $G_5(y)$ denoting the cumulative distribution function of Y_5,

$$
\begin{aligned}
G_5(y) &= P(Y_5 \le y) = P(X_1 \le y \cap X_2 \le y \cap \cdots \cap X_5 \le y) \\
&= P(X_1 \le y) \cdot P(X_2 \le y) \cdot \cdots \cdot P(X_5 \le y) \\
&= [F(y)]^5 = \left[1 - e^{-.01y}\right]^5
\end{aligned}
$$

The pdf of Y_5 can now be obtained by differentiating the cdf with respect to y.

Suppose instead that the five components are connected in series rather than in parallel (Figure 5.9b). In this case the system lifetime will be Y_1, the *smallest* of the five order statistics, since the system will crash as soon as a single one of the individual components fails. Note that system lifetime will exceed y hours if and only if the lifetime of every component exceeds y hours. Thus

$$
\begin{aligned}
G_1(y) &= P(Y_1 \le y) = 1 - P(Y_1 > y) \\
&= 1 - P(X_1 > y \cap X_2 > y \cap \cdots \cap X_5 > y) \\
&= 1 - P(X_1 > y) \cdot P(X_2 > y) \cdot \cdots \cdot P(X_5 > y) \\
&= 1 - \left[e^{-.01y}\right]^5 = 1 - e^{-.05y}
\end{aligned}
$$

This is the form of an exponential cdf with parameter .05. More generally, if the n components in a series connection have lifetimes that are independent, each exponentially distributed with the same parameter λ, then system lifetime will be exponentially distributed with parameter $n\lambda$. The expected system lifetime will then be $1/(n\lambda)$, much smaller than the expected lifetime of an individual component. ∎

An argument parallel to that of the previous example for a general sample size n and an arbitrary pdf $f(x)$ gives the following general results.

PROPOSITION   Let Y_1 and Y_n denote the smallest and largest order statistics, respectively, based on a random sample from a continuous distribution with cdf $F(x)$ and pdf $f(x)$. Then the cdf and pdf of Y_n are

$$
G_n(y) = [F(y)]^n \quad g_n(y) = n[F(y)]^{n-1} \cdot f(y)
$$

The cdf and pdf of Y_1 are

$$
G_1(y) = 1 - [1 - F(y)]^n \quad g_1(y) = n[1 - F(y)]^{n-1} \cdot f(y)
$$

Example 5.40 Let X denote the contents of a one-gallon container, and suppose that its pdf is $f(x) = 2x$ for $0 \leq x \leq 1$ (and 0 otherwise) with corresponding cdf $F(x) = x^2$ on [0, 1]. Consider a random sample of four such containers. The order statistics Y_1 and Y_4 represent the contents of the least-filled container and the most-filled container, respectively. The pdfs of Y_1 and Y_4 are

$$g_1(y) = 4(1 - y^2)^3 \cdot 2y = 8y(1 - y^2)^3 \quad 0 \leq y \leq 1$$
$$g_4(y) = 4(y^2)^3 \cdot 2y = 8y^7 \quad 0 \leq y \leq 1$$

The corresponding density curves appear in Figure 5.10.
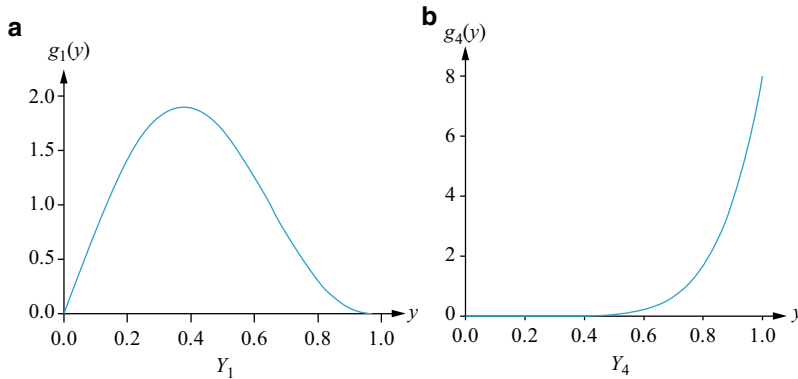
Figure 5.10 Density curves for the order statistics (**a**) Y_1 and (**b**) Y_4 in Example 5.40

Let's determine the expected value of $Y_4 - Y_1$, the difference between the contents of the most-filled container and the least-filled container; $Y_4 - Y_1$ is just the sample range. Apply linearity of expectation:

$$E(Y_4 - Y_1) = E(Y_4) - E(Y_1) = \int_0^1 y \cdot 8y^7 dy - \int_0^1 y \cdot 8y(1 - y^2)^3 dy$$

$$= \frac{8}{9} - \frac{384}{945} = .889 - .406 = .483$$

If random samples of four containers were repeatedly selected and the sample range of contents determined for each one, the long-run average value of the range would be .483 gallons. ∎

The Distribution of the ith Order Statistic

We have already obtained the (marginal) distribution of the largest order statistic Y_n and also that of the smallest order statistic Y_1. A generalization of the argument used previously results in the following proposition; the method of derivation is suggested in Exercise 114.

PROPOSITION   Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a continuous distribution with cdf $F(x)$ and pdf $f(x)$. The pdf of the ith smallest order statistic Y_i is

$$g_i(y) = \frac{n!}{(i - 1)!(n - i)!} [F(y)]^{i-1}[1 - F(y)]^{n-i}f(y) \tag{5.8}$$

An intuitive justification for Expression (5.8) will be given shortly. Notice that it is consistent with the pdf expressions for $g_1(y)$ and $g_n(y)$ given previously; just substitute $i = 1$ and $i = n$, respectively.

Example 5.41   Suppose that component lifetime is exponentially distributed with parameter λ. For a random sample of $n = 5$ components, the expected value of the sample median lifetime is

$$E(Y_3) = \int\limits_0^\infty y \cdot g_3(y)dy = \int\limits_0^\infty y \cdot \frac{5!}{2! \cdot 2!}(1 - e^{-\lambda y})^2(e^{-\lambda y})^2 \cdot \lambda e^{-\lambda y}dy$$

Expanding out the integrand and integrating term by term, the expected value is $.783/\lambda$. The median of the original exponential distribution is, from solving $F(\tilde{\mu}) = .5$, $\tilde{\mu} = -\ln(.5)/\lambda = .693/\lambda$. Thus if sample after sample of five components is selected, the long-run average value of the sample median Y_3 will be somewhat larger than the median value of the individual lifetime distribution. This is because the exponential distribution has a positive skew.   ∎

There's an intuitive "derivation" of Expression (5.8), the general order statistic pdf. Let Δ be a number quite close to 0, and consider the three intervals $(-\infty, y]$, $(y, y + \Delta]$, and $(y + \Delta, \infty)$. For a single X, the probabilities of these three intervals are $p_1 = P(X \leq y) = F(y)$, $p_2 = P(y < X \leq y + \Delta) = \int_y^{y+\Delta} f(x)dx \approx f(y) \cdot \Delta$, $p_3 = P(X > y + \Delta) = 1 - F(y + \Delta)$.

For a random sample of size n, it is very unlikely that two or more X's will fall in the middle interval, since its width is only Δ. The probability that the ith order statistic falls in the middle interval is then approximately the probability that $i - 1$ of the X's are in the first interval, one is in the middle, and the remaining $n - i$ are in the third. This is just a multinomial probability:

$$P(y < Y_i \leq y + \Delta) \approx \frac{n!}{(i - 1)!1!(n - i)!}[F(y_i)]^{i-1} \cdot f(y) \cdot \Delta \cdot [1 - F(y + \Delta)]^{n-i}$$

Dividing both sides by Δ and taking the limit as $\Delta \to 0$ gives exactly Expression (5.8). That is, we may interpret the pdf $g_i(y)$ as loosely specifying that $i - 1$ of the original observations are below y, one is "at" y, and the other $n - i$ are above y.

The Joint Distribution of All *n* Order Statistics

We now develop the joint pdf of $Y_1, Y_2, \ldots, Y_n$. Consider first a random sample X_1, X_2, X_3 of fuel efficiency measurements (mpg). The joint pdf of this random sample is

$$f(x_1, \ x_2, \ x_3) = f(x_1) \cdot f(x_2) \cdot f(x_3)$$

The joint pdf of Y_1, Y_2, Y_3 will be positive only for values of y_1, y_2, y_3 satisfying $y_1 < y_2 < y_3$. What is this joint pdf at the values $y_1 = 28.4$, $y_2 = 29.0$, $y_3 = 30.5$? There are six different ways to obtain these ordered values:

$$X_1 = 28.4 \quad X_2 = 29.0 \quad X_3 = 30.5 \qquad X_1 = 28.4 \quad X_2 = 30.5 \quad X_3 = 29.0$$
$$X_1 = 29.0 \quad X_2 = 28.4 \quad X_3 = 30.5 \qquad X_1 = 29.0 \quad X_2 = 30.5 \quad X_3 = 28.4$$
$$X_1 = 30.5 \quad X_2 = 28.4 \quad X_3 = 29.0 \qquad X_1 = 30.5 \quad X_2 = 29.0 \quad X_3 = 28.4$$

These six possibilities come from the 3! ways to order the three numerical observations once their values are fixed. Thus

$$g(28.4, 29.0, 30.5) = f(28.4) \cdot f(29.0) \cdot f(30.5) + \cdots$$
$$+ f(30.5) \cdot f(29.0) \cdot f(28.4)$$
$$= 3! f(28.4) \cdot f(29.0) \cdot f(30.5)$$

Analogous reasoning with a sample of size n yields the following result:

PROPOSITION   Let $g(y_1, y_2, \ldots, y_n)$ denote the joint pdf of the order statistics $Y_1, Y_2, \ldots, Y_n$ resulting from a random sample of X_i's from a pdf $f(x)$. Then

$$g(y_1, y_2, \ldots, y_n) = n! f(y_1) \cdot f(y_2) \cdot \cdots \cdot f(y_n) \qquad y_1 < y_2 < \cdots < y_n$$

For example, if we have a random sample of component lifetimes and the lifetime distribution is exponential with parameter λ, then the joint pdf of the order statistics is

$$g(y_1, \ldots, y_n) = n! \lambda^n e^{-\lambda(y_1 + \cdots + y_n)} \quad 0 < y_1 < y_2 < \cdots < y_n < \infty$$

Example 5.42 Suppose X_1, X_2, X_3, and X_4 are independent random variables, each uniformly distributed on the interval from 0 to 1. The joint pdf of the four corresponding order statistics Y_1, Y_2, Y_3, and Y_4 is $f(y_1, y_2, y_3, y_4) = 4! \cdot 1$ for $0 < y_1 < y_2 < y_3 < y_4 < 1$. The probability that every pair of X_i's is separated by more than .2 is the same as the probability that $Y_2 - Y_1 > .2$, $Y_3 - Y_2 > .2$, and $Y_4 - Y_3 > .2$. This latter probability results from integrating the joint pdf of the Y_i's over the region $.6 < y_4 < 1$, $.4 < y_3 < y_4 - .2$, $.2 < y_2 < y_3 - .2$, $0 < y_1 < y_2 - .2$:

$$P(Y_2 - Y_1 > .2, Y_3 - Y_2 > .2, Y_4 - Y_3 > .2) = \int_{.6}^{1} \int_{.4}^{y_4 - .2} \int_{.2}^{y_3 - .2} \int_{0}^{y_2 - .2} 4! \, dy_1 dy_2 dy_3 dy_4$$

The inner integration gives $4!(y_2 - .2)$, and this must then be integrated between .2 and $y_3 - .2$. Making the change of variable $z_2 = y_2 - .2$, the integration of z_2 is from 0 to $y_3 - .4$. The result of this integration is $4! \cdot (y_3 - .4)^2/2$. Continuing with the 3rd and 4th integration, each time making an appropriate change of variable so that the lower limit of each integration becomes 0, the result is

$$P(Y_2 - Y_1 > .2, Y_3 - Y_2 > .2, Y_4 - Y_3 > .2) = .4^4 = .0256$$

A more general multiple integration argument for n independent uniform $[0, B]$ rvs shows that the probability that all values are separated by at least d is

$$P(\text{all values are separated by more than } d) = \begin{cases} [1 - (n-1)d/B]^n & 0 \le d \le B/(n-1) \\ 0 & d > B/(n-1) \end{cases}$$

As an application, consider a year that has 365 days, and suppose that the birth time of someone born in that year is uniformly distributed throughout the 365-day period. Then in a group of 10 independently selected people born in that year, the probability that all of their birth times are separated by more than 24 h ($d = 1$ day) is $(1 - 9/365)^{10} = .779$. Thus the probability that at least two of the 10 birth times are separated by at most 24 h is .221. As the group size n increases, it becomes more likely that at least two people have birth times that are within 24 h of each other (but not necessarily on the same day). For $n = 16$, this probability is .467, and for $n = 17$ it is .533. So with as few as 17 people in the group, it is more likely than not that at least two of the people were born within 24 h of each other. Coincidences such as this are not as surprising as one might think. The probability that at least two people are born on the same day (assuming equally likely birthdays) is much easier to calculate than what we have shown here; see The Birthday Problem in Example 2.22.   ∎

The Joint Distribution of Two Order Statistics

Finally, we consider the joint distribution of two order statistics Y_i and Y_j with $i < j$. Consider first $n = 6$ and the two order statistics Y_3 and Y_5. We must then take the joint pdf of all six order statistics, hold y_3 and y_5 fixed, and integrate out y_1, y_2, y_4, and y_6. That is,

$$g(y_3, y_5) = \int_{y_5}^{\infty} \int_{y_3}^{y_5} \int_{-\infty}^{y_3} \int_{y_1}^{y_3} 6! f(y_1) \cdots f(y_6) \, dy_2 \, dy_1 \, dy_4 \, dy_6$$

The result of this integration is

$$g_{3,5}(y_3, y_5) = \frac{6!}{2!1!1!} [F(y_3)]^2 [F(y_5) - F(y_3)]^1 \cdots [1 - F(y_5)]^1 f(y_3) f(y_5)$$

$$-\infty < y_3 < y_5 < \infty$$

The foregoing derivation generalizes as follows.

PROPOSITION   Let $g_{i,j}(y_i, y_j)$ denote the joint pdf of the order statistics Y_i and Y_j, $i < j$, resulting from a random sample of X_i's from a pdf $f(x)$. Then

$$g_{i,j}(y_i, y_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} F(y_i)^{i-1} \cdot$$
$$[F(y_j) - F(y_i)]^{j-i-1} \cdots [1 - F(y_j)]^{n-j} f(y_i) f(y_j)$$
$$\text{for } -\infty < y_i < y_j < \infty$$

This joint pdf can be "derived" intuitively by considering a multinomial probability similar to the argument presented for the marginal pdf of Y_i. In this case, there are five relevant intervals: $(-\infty, y_i]$, $(y_i, y_i + \Delta_1]$, $(y_i + \Delta_1, y_j]$, $(y_j, y_j + \Delta_2]$, and $(y_j + \Delta_2, \infty)$.

Exercises: Section 5.7 (109–121)

109. A friend of ours takes the bus five days per week to her job. The five waiting times until she can board the bus are a random sample from a uniform distribution on the interval from 0 to 10 min.

    a. Determine the pdf and then the expected value of the largest of the five waiting times.
    b. Determine the expected value of the difference between the largest and smallest times.
    c. What is the expected value of the sample median waiting time?
    d. What is the standard deviation of the largest time?

110. Refer back to Example 5.40. Because $n = 4$, the sample median is $(Y_2 + Y_3)/2$. What is the expected value of the sample median, and how does it compare to the median of the population distribution?

111. Referring back to Exercise 109, suppose you learn that the smallest of the five waiting times is 4 min. What is the conditional density function of the largest waiting time, and what is the expected value of the largest waiting time in light of this information?

112. Let X represent a measurement error. It is natural to assume that the pdf $f(x)$ is symmetric about 0, so that the density at a value $-c$ is the same as the density at c (an error of a given magnitude is equally likely to be positive or negative). Consider a random sample of n measurements, where $n = 2k + 1$, so that Y_{k+1} is the sample median. What can be said about $E(Y_{k+1})$? If the X distribution is symmetric about some other value, so that value is the median of the distribution, what does this imply about $E(Y_{k+1})$? [*Hints*: For the first question, symmetry implies that $1 - F(x) = P(X > x) = P(X < -x) = F(-x)$. For the second question, consider $W = X - \tilde{\mu}$; what is the median of the distribution of W?]

113. A store is expecting n deliveries between the hours of noon and 1 p.m. Suppose the arrival time of each delivery truck is uniformly distributed on this one-hour interval and that the times are independent of each other. What are the expected values of the ordered arrival times?

114. The pdf of the second-largest order statistic, Y_{n-1}, can be obtained using reasoning analogous to how the pdf of Y_n was first obtained.

    a. For any number y, $Y_{n-1} \leq y$ if and only if *at least* $n - 1$ of the original X's are $\leq y$. (Do you see why?) Use this fact to derive a formula for the cdf of Y_{n-1} in terms of F, the cdf of the X's. [*Hint*: Separate "at least $n - 1$" into two cases and apply the binomial formula.]
    b. Differentiate part (a) to obtain the pdf of Y_{n-1}. Simplify and verify it matches the formula for $g_{n-1}(y)$ provided in this section.

115. Let X be the amount of time an ATM is in use during a particular one-hour period, and suppose that X has the cdf $F(x) = x^\theta$ for $0 < x < 1$ (where $\theta > 1$). Give expressions involving the gamma function for both the mean and variance of the ith smallest amount of time Y_i from a random sample of n such time periods.

116. The logistic pdf $f(x) = e^{-x}/(1 + e^{-x})^2$ for $-\infty < x < \infty$ is sometimes used to describe the distribution of measurement errors.

    a. Graph the pdf. Does the appearance of the graph surprise you?
    b. For a random sample of size n, obtain an expression involving the gamma function for the moment generating function of the ith smallest order statistic Y_i. This expression can then be differentiated to obtain moments of the order statistics. [*Hint*: Set up the appropriate integral, and then let $u = 1/(1 + e^{-x})$.]

117. An insurance policy issued to a boat owner has a deductible amount of $1000, so the amount of damage claimed must exceed this deductible before there will be a payout. Suppose the amount (1000s of dollars) of a randomly selected claim is a continuous rv with pdf $f(x) = 3/x^4$ for $x > 1$. Consider a random sample of three claims.

    a. What is the probability that at least one of the claim amounts exceeds $5000?
    b. What is the expected value of the largest amount claimed?

118. Conjecture the form of the joint pdf of three order statistics Y_i, Y_j, Y_k ($i < j < k$) in a random sample of size n.

119. Use the intuitive argument sketched in this section to obtain the general formula for the joint pdf of two order statistics given in the last proposition.

120. Consider a sample of size $n = 3$ from the standard normal distribution, and obtain the expected value of the largest order statistic. What does this say about the expected value of the largest order statistic in a sample of this size from *any* normal distribution? [*Hint*: With $\phi(x)$ denoting the standard normal pdf, use the fact that $(d/dx)\phi(x) = -x\phi(x)$ along with integration by parts.]

121. Let Y_1 and Y_n be the smallest and largest order statistics, respectively, from a random sample of size n.

    a. Use the last proposition in this section to determine the joint pdf of Y_1 and Y_n. (Your answer will include the pdf f and cdf F of the original random sample.)
    b. Let $W_1 = Y_1$ and $W_2 = Y_n - Y_1$ (the latter is the sample range). Use the method of Section 5.6 to obtain the joint pdf of W_1 and W_2, and then derive an expression involving an integral for the pdf of the sample range.
    c. For the case in which the random sample is from a uniform distribution on [0, 1], carry out the integration of (b) to obtain an

explicit formula for the pdf of the sample range. [*Hint:* For the Uniform[0, 1] distribution, what are f and F?]

Supplementary Exercises: (122–150)

122. Suppose the amount of rainfall in one region during a particular month has an exponential distribution with mean value 3 in., the amount of rainfall in a second region during that same month has an exponential distribution with mean value 2 in., and the two amounts are independent of each other. What is the probability that the second region gets more rainfall during this month than does the first region?

123. Two messages are to be sent. The time (min) necessary to send each message has an exponential distribution with parameter $\lambda = 1$, and the two times are independent of each other. It costs $2 per minute to send the first message and $1 per minute to send the second. Obtain the density function of the total cost of sending the two messages. [*Hint*: First obtain the cumulative distribution function of the total cost, which involves integrating the joint pdf.]

124. A restaurant serves three fixed-price dinners costing $25, $35, and $50. For a randomly selected couple dining at this restaurant, let $X =$ the cost of the man's dinner and $Y =$ the cost of the woman's dinner. The joint pmf of X and Y is given in the following table:

		y	
$p(x, y)$	25	35	50
x 25	.05	.05	.10
35	.05	.10	.35
50	0	.20	.10

    a. Compute the marginal pmfs of X and Y.
    b. What is the probability that the man's and the woman's dinner cost at most $35 each?

c. Are X and Y independent? Justify your answer.

d. What is the expected total cost of the dinner for the two people?

e. Suppose that when a couple opens fortune cookies at the conclusion of the meal, they find the message "You will receive as a refund the difference between the cost of the more expensive and the less expensive meal that you have chosen." How much does the restaurant expect to refund?

125. A health-food store stocks two different brands of a type of grain. Let $X =$ the amount (lb) of brand A on hand and $Y =$ the amount of brand B on hand. Suppose the joint pdf of X and Y is

$$f(x, y) = kxy \quad x \geq 0, \ y \geq 0, \ 20 \leq x + y \leq 30$$

a. Draw the region of positive density and determine the value of k.

b. Are X and Y independent? Answer by first deriving the marginal pdf of each variable.

c. Compute $P(X + Y \leq 25)$.

d. What is the expected total amount of this grain on hand?

e. Compute $\text{Cov}(X, Y)$ and $\text{Corr}(X, Y)$.

f. What is the variance of the total amount of grain on hand?

126. Let $X_1, X_2, \ldots, X_n$ be random variables denoting n independent bids for an item that is for sale. Suppose each X_i is uniformly distributed on the interval [100, 200]. If the seller sells to the highest bidder, how much can he expect to earn on the sale? [*Hint*: Let $Y = \max(X_1, X_2, \ldots, X_n)$. Find $F_Y(y)$ by using the results of Section 5.7 or else by noting that $Y \leq y$ iff each X_i is $\leq y$. Then obtain the pdf and $E(Y)$.]

127. Suppose a randomly chosen individual's verbal score X and quantitative score Y on a nationally administered aptitude examination have joint pdf

$$f(x, y) = \frac{2}{5}(2x + 3y) \quad 0 \leq x \leq 1, \ 0 \leq y \leq 1$$

You are asked to provide a prediction t of the individual's total score $X + Y$. The error of prediction is the mean squared error $E[(X + Y - t)^2]$. What value of t minimizes the error of prediction?

128. Let X_1 and X_2 be quantitative and verbal scores on one aptitude exam, and let Y_1 and Y_2 be corresponding scores on another exam. If $\text{Cov}(X_1, Y_1) = 5$, $\text{Cov}(X_1, Y_2) = 1$, $\text{Cov}(X_2, Y_1) = 2$, and $\text{Cov}(X_2, Y_2) = 8$, what is the covariance between the two total scores $X_1 + X_2$ and $Y_1 + Y_2$?

129. Let Z_1 and Z_2 be independent standard normal rvs and let

$$U = Z_1 \quad V = \rho \cdot Z_1 + \sqrt{1 - \rho^2} \cdot Z_2$$

a. By definition, U has mean 0 and standard deviation 1. Show that the same is true for V.

b. Use the properties of covariance to show that $\text{Cov}(U, V) = \rho$.

c. Show that $\text{Corr}(U, V) = \rho$.

130. You are driving on a highway at speed X_1. Cars entering this highway after you travel at speeds $X_2, X_3, \ldots$. Suppose these X_i's are independent and identically distributed with pdf $f(x)$ and cdf $F(x)$. Unfortunately there is no way for a faster car to pass a slower one —it will catch up to the slower one and then travel at the same speed. For example, if $X_1 = 52.3$, $X_2 = 37.5$, and $X_3 = 42.8$, then no car will catch up to yours, but the third car will catch up to the second. Let $N =$ the number of cars that ultimately travel at your speed (in your "cohort"), including your own car. Possible values of N are 1, 2, 3, …. Show that the pmf of N is $p(n) = 1/[n(n + 1)]$, and then determine the expected number of cars in your cohort. [*Hint*: $N = 3$ requires that $X_1 < X_2$, $X_1 < X_3$, $X_4 < X_1$.]

131. Suppose the number of children born to an individual has pmf $p(x)$. A *Galton–Watson branching process* unfolds as follows: At time $t = 0$, the population consists of a single individual. Just prior to time $t = 1$, this individual gives birth to X_1 individuals according to the pmf $p(x)$, so there are X_1 individuals in the first generation. Just prior to time $t = 2$, each of these X_1 individuals gives birth independently of the others according to the pmf $p(x)$, resulting in X_2 individuals in the second generation (e.g., if $X_1 = 3$, then $X_2 = Y_1 + Y_2 + Y_3$, where Y_i is the number of progeny of the ith individual in the first generation). This process then continues to yield a third generation of size X_3, and so on.

    a. If $X_1 = 3$, $Y_1 = 4$, $Y_2 = 0$, $Y_3 = 1$, draw a tree diagram with two generations of branches to represent this situation.
    b. Let A be the event that the process ultimately becomes extinct (one way for A to occur would be to have $X_1 = 3$ with none of these three second-generation individuals having any progeny) and let $p^* = P(A)$. Argue that p^* satisfies the equation

    $$p* = \sum (p*)^x \cdot p(x)$$

    That is, $p^* = \psi(p^*)$ where $\psi(s)$ is the probability generating function introduced in Exercise 166 from Chapter 3. [*Hint*: $A = \bigcup_x (A \cap \{X_1 = x\})$, so the Law of Total Probability can be applied. Now given that $X_1 = 3$, A will occur if and only if each of the three separate branching processes starting from the first generation ultimately becomes extinct; what is the probability of this happening?

    c. Verify that one solution to the equation in (b) is $p^* = 1$. It can be shown that this equation has just one other solution, and that the probability of ultimate extinction is in fact the *smaller* of the two roots. If $p(0) = .3$, $p(1) = .5$, and $p(2) = .2$, what is p^*? Is this consistent with the

value of μ, the expected number of progeny from a single individual? What happens if $p(0) = .2$, $p(1) = .5$, and $p(2) = .3$?

132. Let $f(x)$ and $g(y)$ be pdfs with corresponding cdfs $F(x)$ and $G(y)$, respectively. With c denoting a numerical constant satisfying $|c| \le 1$, consider

$$f(x, y) = f(x)g(y)\{1 + c[2F(x) - 1][2G(y) - 1]\}$$

    a. Show that $f(x, y)$ satisfies the conditions necessary to specify a joint pdf for two continuous rvs.
    b. What is the marginal pdf of the first variable X? Of the second variable Y?
    c. For what values of c are X and Y independent?
    d. If $f(x)$ and $g(y)$ are normal pdfs, is the joint distribution of X and Y bivariate normal?

133. The **joint cumulative distribution function** of two random variables X and Y, denoted by $F(x, y)$, is defined by

    $$\begin{aligned} F(x, y) &= P[(X \le x) \cap (Y \le y)] \\ &-\infty < x < \infty, \quad -\infty < y < \infty \end{aligned}$$

    a. Suppose that X and Y are both continuous variables. Once the joint cdf is available, explain how it can be used to determine the probability $P[(X, Y) \in A]$, where A is the rectangular region $\{(x, y): a \le x \le b, c \le y \le d\}$.
    b. Suppose the only possible values of X and Y are 0, 1, 2, … and consider the values $a = 5$, $b = 10$, $c = 2$, and $d = 6$ for the rectangle specified in (a). Describe how you would use the joint cdf to calculate the probability that the pair (X, Y) falls in the rectangle. More generally, how can the rectangular probability be calculated from the joint cdf if a, b, c, and d are all integers?
    c. Determine the joint cdf for the scenario of Example 5.1. [*Hint*: First determine $F(x, y)$ for $x = 100, 250$ and $y = 0, 100$,

and 200. Then describe the joint cdf for various other (x, y) pairs.]

d. Determine the joint cdf for the scenario of Example 5.3 and use it to calculate the probability that X and Y are both between .25 and .75. [*Hint*: For $0 \leq x \leq 1$ and $0 \leq y \leq 1$, $F(x, y) = \int_0^x \int_0^y f(u, v) dv du$.]

e. Determine the joint cdf for the scenario of Example 5.5. [*Hint*: Proceed as in (d), but be careful about the order of integration and consider separately (x, y) points that lie inside the triangular region of positive density and then points that lie outside this region.]

134. A circular sampling region with radius X is chosen by a biologist, where X has an exponential distribution with mean value 10 ft. Plants of a certain type occur in this region according to a (spatial) Poisson process with "rate" .5 plant per square foot. Let Y denote the number of plants in the region.

a. Find $E(Y|X = x)$ and $V(Y|X = x)$
b. Use part (a) to find $E(Y)$.
c. Use part (a) to find $V(Y)$.

135. The number of individuals arriving at a post office to mail packages during a certain period is a Poisson random variable X with mean value 20. Independently of the others, any particular customer will mail either 1, 2, 3, or 4 packages with probabilities .4, .3, .2, and .1, respectively. Let Y denote the total number of packages mailed during this time period.

a. Find $E(Y|X = x)$ and $V(Y|X = x)$.
b. Use part (a) to find $E(Y)$.
c. Use part (a) to find $V(Y)$.

136. Consider a sealed-bid auction in which each of the n bidders has his/her valuation (assessment of inherent worth) of the item being auctioned. The valuation of any particular bidder is not known to the other bidders. Suppose these valuations constitute a random sample $X_1, \ldots, X_n$ from a

distribution with cdf $F(x)$, with corresponding order statistics $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. The *rent* of the winning bidder is the difference between the winner's valuation and the price. The article "Mean Sample Spacings, Sample Size and Variability in an Auction-Theoretic Framework" (*Oper. Res. Lett.* 2004: 103–108) argues that the rent is just $Y_n - Y_{n-1}$ (do you see why?).

a. Suppose that the valuation distribution is uniform on [0, 100]. What is the expected rent when there are $n = 10$ bidders?

b. Referring back to (a), what happens when there are 11 bidders? More generally, what is the relationship between the expected rent for n bidders and for $n + 1$ bidders? Is this intuitive? [*Note*: The cited article presents a counterexample.]

137. Suppose two identical components are connected in parallel, so the system continues to function as long as at least one of the components does so. The two lifetimes are independent of each other, each having an exponential distribution with mean 1000 h. Let W denote system lifetime. Obtain the moment generating function of W, and use it to calculate the expected lifetime.

138. Sandstone is mined from two different quarries. Let X = the amount mined (in tons) from the first quarry each day and Y = the amount mined (in tons) from the second quarry each day. The variables X and Y are independent, with $\mu_X = 12$, $\sigma_X = 4$, $\mu_Y = 10$, $\sigma_Y = 3$.

a. Find the mean and standard deviation of the variable $X + Y$, the total amount of sandstone mined in a day.

b. Find the mean and standard deviation of the variable $X - Y$, the difference in the mines' performances in a day.

c. The manager of the first quarry sells sandstone at $25/ton, while the manager of the second quarry sells sandstone at

$28/ton. Find the mean and standard deviation for the combined amount of money the quarries generate in a day.

d. Assuming X and Y are both normally distributed, find the probability that the quarries generate more than $750 revenue in a day.

139. In cost estimation, the total cost of a project is the sum of component task costs. Each of these costs is a random variable with a probability distribution. It is customary to obtain information about the total cost distribution by adding together characteristics of the individual component cost distributions—this is called the "roll-up" procedure. Since $E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$, the roll-up procedure is valid for mean cost. Suppose that there are two component tasks and that X_1 and X_2 are independent, normally distributed random variables. Is the roll-up procedure valid for the 75th percentile? That is, is the 75th percentile of the distribution of $X_1 + X_2$ the same as the sum of the 75th percentiles of the two individual distributions? If not, what is the relationship between the percentile of the sum and the sum of percentiles? For what percentiles is the roll-up procedure valid in this case?

140. *Random sums.* If X_1, X_2, …, X_n are independent rvs, each with the same mean value μ and variance σ^2, then the methods of Section 5.3 show that $E(X_1 + \cdots + X_n) = n\mu$ and $V(X_1 + X_2 + \cdots + X_n) = n\sigma^2$. In some applications, the number of X_i's under consideration is not a fixed number n but instead a rv N. For example, let N be the number of components of a certain type brought into a repair shop on a particular day and let X_i represent the repair time for the ith component. Then the total repair time is $T_N = X_1 + X_2 + \cdots + X_N$, the sum of a *random* number of rvs.

a. Suppose that N is independent of the X_i's. Use the Law of Total Expectation

to obtain an expression for $E(T_N)$ in terms of μ and $E(N)$.

b. Use the Law of Total Variance to obtain an expression for $V(T_N)$ in terms of μ, σ^2, $E(N)$, and $V(N)$.

c. Customers submit orders for stock purchases at a certain online site according to a Poisson process with a rate of 3 per hour. The amount purchased by any particular customer (in thousands of dollars) has an exponential distribution with mean 30, and purchase amounts are independent of the number of customers. What is the expected total amount ($) purchased during a particular 4-h period, and what is the standard deviation of this total amount?

141. The mean weight of luggage checked by a randomly selected tourist-class passenger flying between two cities on a certain airline is 40 lb, and the standard deviation is 10 lb. The mean and standard deviation for a business-class passenger are 30 lb and 6 lb, respectively.

a. If there are 12 business-class passengers and 50 tourist-class passengers on a particular flight, what are the expected value of total luggage weight and the standard deviation of total luggage weight?

b. If individual luggage weights are independent, normally distributed rvs, what is the probability that total luggage weight is at most 2500 lb?

142. The amount of soft drink that Ann consumes on any given day is independent of consumption on any other day and is normally distributed with $\mu = 13$ oz and $\sigma = 2$. If she currently has two six-packs of 16-oz bottles, what is the probability that she still has some soft drink left at the end of 2 weeks (14 days)? Why should we worry about the validity of the independence assumption here?

143. A student has a class that is supposed to end at 9:00 a.m. and another that is supposed to begin at 9:10 a.m. Suppose the actual ending time of the 9 a.m. class is a

normally distributed rv X_1 with mean 9:02 and standard deviation 1.5 min and that the starting time of the next class is also a normally distributed rv X_2 with mean 9:10 and standard deviation 1 min. Suppose also that the time necessary to get from one classroom to the other is a normally distributed rv X_3 with mean 6 min and standard deviation 1 min. Assuming independence of X_1, X_2, and X_3, what is the probability that the student makes it to the second class before the lecture starts? Why should we worry about the reasonableness of the independence assumption here?

144. This exercise provides an alternative approach to establishing the properties of correlation.

   a. Use the general formula for the variance of a linear combination to write an expression for $V(aX + Y)$. Then let $a = \sigma_Y/\sigma_X$, and show that $\rho \geq -1$. [*Hint*: Variance is always ≥ 0, and $\mathrm{Cov}(X, Y) = \sigma_X \cdot \sigma_Y \cdot \rho$.]
   b. By considering $V(aX - Y)$, conclude that $\rho \leq 1$.
   c. Use the fact that $V(W) = 0$ only if W is a constant to show that $\rho = 1$ only if $Y = aX + b$.

145. A rock specimen from a particular area is randomly selected and weighed two different times. Let W denote the actual weight and X_1 and X_2 the two measured weights. Then $X_1 = W + E_1$ and $X_2 = W + E_2$, where E_1 and E_2 are the two measurement errors. Suppose that the E_i's are independent of each other and of W and that $V(E_1) = V(E_2) = \sigma_E^2$.

   a. Express ρ, the correlation coefficient between the two measured weights X_1 and X_2, in terms of σ_W^2, the variance of actual weight, and σ_X^2, the variance of measured weight.
   b. Compute ρ when $\sigma_W = 1$ kg and $\sigma_E = .01$ kg.

146. Let A denote the percentage of one constituent in a randomly selected rock specimen, and let B denote the percentage of a second constituent in that same specimen. Suppose D and E are measurement errors in determining the values of A and B so that measured values are $X = A + D$ and $Y = B + E$, respectively. Assume that measurement errors are independent of each other and of actual values.

   a. Show that
   $$\mathrm{Corr}(X, Y) = \mathrm{Corr}(A, B) \\ \cdot \sqrt{\mathrm{Corr}(X_1, X_2)} \\ \cdot \sqrt{\mathrm{Corr}(Y_1, Y_2)}$$
   where X_1 and X_2 are replicate measurements on the value of A, and Y_1 and Y_2 are defined analogously with respect to B. What effect does the presence of measurement error have on the correlation?
   b. What is the maximum value of $\mathrm{Corr}(X, Y)$ when $\mathrm{Corr}(X_1, X_2) = .8100$, $\mathrm{Corr}(Y_1, Y_2) = .9025$? Is this disturbing?

147. Let X_1, …, X_n be independent rvs with mean values μ_1, …, μ_n and variances σ_1^2, …, σ_n^2. Consider a function $h(x_1, \ldots, x_n)$, and use it to define a new random variable $Y = h(X_1, \ldots, X_n)$. Under rather general conditions on the h function, if the σ_is are all small relative to the corresponding μ_is, it can be shown that $E(Y) \approx h(\mu_1, \ldots, \mu_n)$ and

$$V(Y) \approx \left(\frac{\partial h}{\partial x_1}\right)^2 \cdot \sigma_1^2 + \cdots + \left(\frac{\partial h}{\partial x_n}\right)^2 \cdot \sigma_n^2$$

where each partial derivative is evaluated at $(x_1, \ldots, x_n) = (\mu_1, \ldots, \mu_n)$. Suppose three resistors with resistances X_1, X_2, X_3 are connected in parallel across a battery with voltage X_4. Then by Ohm's law, the current is

$$Y = X_4\left(\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3}\right)$$

Let $\mu_1 = 10\ \Omega$, $\sigma_1 = 1.0\ \Omega$, $\mu_2 = 15\ \Omega$, $\sigma_2 = 1.0\ \Omega$, $\mu_3 = 20\ \Omega$, $\sigma_3 = 1.5\ \Omega$, $\mu_4 = 120\ V$, $\sigma_4 = 4.0\ V$. Calculate the approximate expected value and standard deviation of the current (suggested by "Random Samplings," *CHEMTECH* 1984: 696–697).

148. A more accurate approximation to $E[h(X_1, \ldots, X_n)]$ in the previous exercise is

$$
E[h(X_1, \ldots, X_n)] \approx h(\mu_1, \ldots, \mu_n) + \frac{1}{2}\sigma_1^2\left(\frac{\partial^2 h}{\partial x_1^2}\right)
$$
$$
+ \cdots + \frac{1}{2}\sigma_n^2\left(\frac{\partial^2 h}{\partial x_n^2}\right)
$$

Compute this for $Y = h(X_1, X_2, X_3, X_4)$ given in the previous exercise, and compare it to the leading term $h(\mu_1, \ldots, \mu_n)$.

149. The following example is based on "Conditional Moments and Independence" (*The American Statistician* 2008: 219). Consider the following joint pdf of two rvs X and Y:

$$
f(x,y) = \frac{e^{-[(\ln x)^2 + (\ln y)^2]/2}}{2\pi xy}[1 + \sin(2\pi \ln x)\sin(2\pi \ln y)]
$$
$$
\text{for } x > 0,\ y > 0
$$

a. Show that the marginal distribution of each rv is lognormal. [*Hint:* When obtaining the marginal pdf of X, make the change of variable $u = \ln(y)$.]

b. Obtain the conditional pdf of Y given that $X = x$. Then show that for every positive integer n, $E(Y^n|X = x) = E(Y^n)$. [*Hint*: Make the change of variable $\ln(y) = u + n$ in the second integrand.]

c. Redo (b) with X and Y interchanged.

d. The results of (b) and (c) suggest intuitively that X and Y are independent rvs. Are they in fact independent?

150. Let Y_0 denote the initial price of a particular security and Y_n denote the price at the end of n additional weeks for $n = 1, 2, 3, \ldots$. Assume that the successive price ratios Y_1/Y_0, Y_2/Y_1, $Y_3/Y_2, \ldots$ are independent of one another and that each ratio has a lognormal distribution with $\mu = .4$ and $\sigma = .8$ (the assumptions of independence and lognormality are common in such scenarios).

a. Calculate the probability that the security price will increase over the course of a week.

b. Calculate the probability that the security price will be higher at the end of the next week, be lower the week after that, and then be higher again at the end of the following week. [*Hint*: What does "higher" say about the ratio Y_{i+1}/Y_i?]

c. Calculate the probability that the security price will have increased by at least 20% over the course of a five-week period. [*Hint*: Consider the ratio Y_5/Y_0, and write this in terms of successive ratios Y_{i+1}/Y_i.]

Statistics and Sampling Distributions

6

Introduction

This chapter helps make the transition between probability and inferential statistics. Given a sample of n observations from a population, we will be calculating *estimates* of the population mean, median, standard deviation, and various other population characteristics (parameters). Prior to obtaining data, there is uncertainty as to which of all possible samples will occur. Because of this, estimates such as $\bar{x}$, $\tilde{x}$, and s will vary from one sample to another. The behavior of such estimates in repeated sampling is described by what are called *sampling distributions*. Any particular sampling distribution will give an indication of how close the estimate is likely to be to the value of the parameter being estimated.

The first two sections use probability results to study sampling distributions. A particularly important result is the *Central Limit Theorem*, which shows how the behavior of the sample mean can be described by a normal distribution when the sample size is large. The last two sections introduce several distributions related to samples from a normal population distribution. Many inferential procedures are based on properties of these sampling distributions.

6.1 Statistics and Their Distributions

The observations in a single sample were denoted in Chapter 1 by x_1, x_2, …, x_n. Consider selecting two different samples of size n from the same population distribution. The x_i's in the second sample will virtually always differ at least a bit from those in the first sample. For example, a first sample of $n = 3$ cars of a particular model might result in fuel efficiencies $x_1 = 30.7$, $x_2 = 29.4$, $x_3 = 31.1$, whereas a second sample may give $x_1 = 28.8$, $x_2 = 30.0$, and $x_3 = 31.1$. Before we obtain data, there is uncertainty about the value of each x_i. Because of this uncertainty, *before* the data becomes available we view each observation as a random variable and denote the sample by X_1, X_2, …, X_n (uppercase letters for random variables).

This variation in observed values in turn implies that the value of any function of the sample observations—such as the sample mean, sample standard deviation, or sample iqr—also varies from sample to sample. That is, prior to obtaining x_1, …, x_n, there is uncertainty as to the value of $\bar{x}$, the value of s, and so on.

Example 6.1 Suppose that material strength for a randomly selected specimen of a particular type has a Weibull distribution with parameter values $\alpha = 2$ (shape) and $\beta = 5$ (scale). The corresponding density curve is shown in Figure 6.1. Formulas from Section 4.5 give
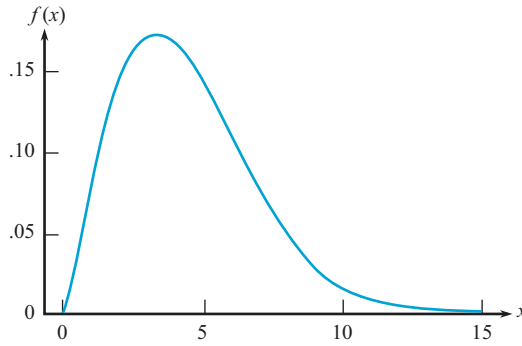
Figure 6.1 The Weibull density curve for Example 6.1

$$\mu = E(X) = 4.4311 \quad \tilde{\mu} = 4.1628 \quad \sigma^2 = V(X) = 5.365 \quad \sigma = 2.316$$

The mean exceeds the median because of the distribution's positive skew.

We used statistical software to generate six different samples, each with $n = 10$, from this distribution (material strengths for six different groups of ten specimens each). The results appear in Table 6.1, followed by the values of the sample mean, sample median, and sample standard deviation for each sample. Notice first that the ten observations in any particular sample are all different from those in any other sample. Second, the six values of the sample mean are all different from each other, as are the six values of the sample median and the six values of the sample standard deviation. The same would be true of the sample 10% trimmed means, sample iqrs, and so on.

Table 6.1 Samples from the Weibull distribution of Example 6.1

	Sample					
	1	2	3	4	5	6
Observation						
1	6.1171	5.07611	3.46710	1.55601	3.12372	8.93795
2	4.1600	6.79279	2.71938	4.56941	6.09685	3.92487
3	3.1950	4.43259	5.88129	4.79870	3.41181	8.76202
4	0.6694	8.55752	5.14915	2.49759	1.65409	7.05569
5	1.8552	6.82487	4.99635	2.33267	2.29512	2.30932
6	5.2316	7.39958	5.86887	4.01295	2.12583	5.94195
7	2.7609	2.14755	6.05918	9.08845	3.20938	6.74166
8	10.2185	8.50628	1.80119	3.25728	3.23209	1.75468
9	5.2438	5.49510	4.21994	3.70132	6.84426	4.91827
10	4.5590	4.04525	2.12934	5.50134	4.20694	7.26081
Mean	4.401	5.928	4.229	4.132	3.620	5.761
Median	4.360	6.144	4.608	3.857	3.221	6.342
SD	2.642	2.062	1.611	2.124	1.678	2.496

Furthermore, the value of the sample mean from any particular sample can be regarded as a *point estimate* ("point" because it is a single number, corresponding to a single point on the number line) of the population mean μ, whose value is known to be 4.4311. None of the estimates from these six samples is identical to what is being estimated. The estimates from the second and sixth samples are much too large, whereas the fifth sample gives a substantial underestimate. Similarly, the sample standard deviation gives a point estimate of the population standard deviation, $\sigma = 2.316$. All six of the resulting estimates are in error by at least a small amount. ∎

In summary, the values of the individual sample observations vary from sample to sample, so in general the value of any quantity computed from sample data, and the value of a sample characteristic used as an estimate of the corresponding population characteristic, will virtually never coincide with what is being estimated.

DEFINITION   A **statistic** is any quantity whose value can be calculated from sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, *a statistic is a random variable* and will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

Thus the sample mean, regarded as a statistic (before a sample has been selected or an experiment has been carried out), is denoted by $\overline{X}$; the calculated value of this statistic from a particular sample is $\bar{x}$. Similarly, S represents the sample standard deviation thought of as a statistic, and its computed value is s.

Any statistic, being a random variable, has a probability distribution. The probability distribution of any particular statistic depends not only on the population distribution (normal, uniform, etc.) and the sample size n but also on the method of sampling. Our next definition describes a sampling method often encountered, at least approximately, in practice.

DEFINITION   The rvs $X_1, X_2, \ldots, X_n$ are said to form a (simple) **random sample** of size n if

1. The X_i's are independent rvs.
2. Every X_i has the same probability distribution.

Such a collection of random variables is also referred to as being **independent and identically distributed (iid)**.

If sampling is either with replacement or from an infinite (conceptual) population, Conditions 1 and 2 are satisfied exactly. These conditions will be approximately satisfied if sampling is without replacement, yet the sample size n is much smaller than the population size N. In practice, if $n/N \leq .05$ (at most 5% of the population is sampled), we can proceed as if the X_i's form a random sample. The virtue of this sampling method is that the probability distribution of any statistic can be more easily obtained than for any other sampling method.

The probability distribution of a statistic is sometimes referred to as its **sampling distribution** to emphasize that it describes how the statistic varies in value across all samples that might be selected. There are two general methods for obtaining information about a statistic's sampling distribution. One method involves calculations based on probability rules, and the other involves carrying out a simulation experiment.

Deriving the Sampling Distribution of a Statistic

Probability rules can be used to obtain the distribution of a statistic provided that it is a "fairly simple" function of the X_i's and either there are relatively few different X values in the population or else the population distribution has a "nice" form. Our next two examples illustrate such situations.

Example 6.2 An online florist offers three different sizes for Mother's Day bouquets: a small arrangement costing \$80 (including shipping), a medium-sized one for \$100, and a large one with a price tag of \$120. If 20% of all purchasers choose the small arrangement, 30% choose medium, and 50% choose large (because they really love Mom!), then the probability distribution of the cost of a single randomly selected flower arrangement is given by

$$\frac{\begin{array}{c|ccc} x & 80 & 100 & 120 \\ \hline p(x) & .2 & .3 & .5 \end{array}}{} \quad \text{with } \mu = 106, \ \sigma^2 = 244 \qquad (6.1)$$

Suppose only two bouquets are sold today. Let X_1 = the cost of the first bouquet and X_2 = the cost of the second. Suppose that X_1 and X_2 are independent, each with the probability distribution shown in (6.1), so that X_1 and X_2 constitute a random sample from the distribution (6.1). Table 6.2 lists possible (x_1, x_2) pairs, the probability of each pair computed using (6.1) and the assumption of independence, and the resulting $\bar{x}$ and s^2 values. (Note that when $n = 2$, $s^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2$.)

Table 6.2 Outcomes, probabilities, and values of $\bar{x}$ and s^2 for Example 6.2

x_1	x_2	$p(x_1, x_2)$	$\bar{x}$	s^2
80	80	(.2)(.2) = .04	80	0
80	100	(.2)(.3) = .06	90	200
80	120	(.2)(.5) = .10	100	800
100	80	(.3)(.2) = .06	90	200
100	100	(.3)(.3) = .09	100	0
100	120	(.3)(.5) = .15	110	200
120	80	(.5)(.2) = .10	100	800
120	100	(.5)(.3) = .15	110	200
120	120	(.5)(.5) = .25	120	0

Now to obtain the probability distribution of $\overline{X}$, the sample average cost per bouquet, we must consider each possible value $\bar{x}$ and compute its probability. For example, $\bar{x} = 100$ occurs three times in the table with probabilities .10, .09, and .10, so

$$P(\overline{X} = 100) = .10 + .09 + .10 = .29$$

Similarly, $s^2 = 800$ appears twice in the table with probability .10 each time, so

$$P(S^2 = 800) = P(X_1 = 80, X_2 = 120) + P(X_1 = 120, X_2 = 80)$$
$$= .10 + .10 = .20$$

The complete sampling distributions of $\overline{X}$ and S^2 appear in (6.2) and (6.3).

$$
\begin{array}{c|ccccc}
\bar{x} & 80 & 90 & 100 & 110 & 120 \\
\hline
p_{\bar{X}}(\bar{x}) & .04 & .12 & .29 & .30 & .25
\end{array}
\tag{6.2}
$$

$$
\begin{array}{c|ccc}
s^2 & 0 & 200 & 800 \\
\hline
p_{S^2}(s^2) & .38 & .42 & .20
\end{array}
\tag{6.3}
$$

Figure 6.2 depicts a probability histogram for both the original distribution of X (6.1) and the $\overline{X}$ distribution (6.2). The figure suggests first that the mean (i.e., expected value) of $\overline{X}$ is equal to the mean \$106 of the original distribution, since both histograms appear to be centered at the same place. Indeed, from (6.2),

$$
E(\overline{X}) = \sum \bar{x} p_{\bar{X}}(\bar{x}) = 80(.04) + \cdots + 120(.25) = 106 = \mu
$$
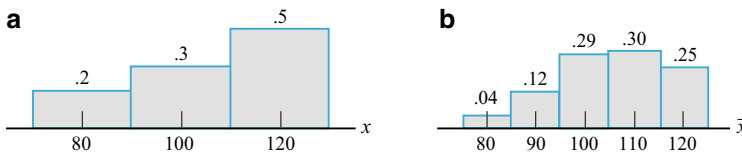
Figure 6.2 Probability histograms for (**a**) the underlying population distribution and (**b**) the sampling distribution of $\overline{X}$ in Example 6.2

Second, it appears that the $\overline{X}$ distribution has smaller spread (variability) than the original distribution, since the values of $\bar{x}$ are more concentrated toward the mean. Again from (6.2),

$$
\begin{aligned}
V(\overline{X}) &= \sum (\bar{x} - \mu_{\overline{X}})^2 p_{\bar{X}}(\bar{x}) = \sum (\bar{x} - 106)^2 p_{\bar{X}}(\bar{x}) \\
&= (80 - 106)^2(.04) + \cdots + (120 - 106)^2(.25) = 122
\end{aligned}
$$

Notice that $V(\overline{X}) = 122 = 244/2 = \sigma^2/2$, exactly half the population variance; that is a consequence of the sample size $n = 2$, and we'll see why in the next section.

Finally, the mean value of S^2 is

$$
E(S^2) = \sum s^2 p_{S^2}(s^2) = 0(.38) + 200(.42) + 800(.20) = 244 = \sigma^2
$$

That is, the $\overline{X}$ sampling distribution is centered at the population mean μ, and the S^2 sampling distribution (histogram not shown) is centered at the population variance σ^2.

If four flower arrangements had been purchased on the day of interest, the sample average cost $\overline{X}$ would be based on a random sample of four X_i's, each having the distribution (6.1). More calculation eventually yields the distribution of $\overline{X}$ for $n = 4$ as

$\bar{x}$	80	85	90	95	100	105	110	115	120
$p_{\overline{X}}(\bar{x})$	.0016	.0096	.0376	.0936	.1761	.2340	.2350	.1500	.0625

From this, $E(\overline{X}) = 106 = \mu$ and $V(\overline{X}) = 61 = \sigma^2/4$. Figure 6.3 is a probability histogram of this distribution.
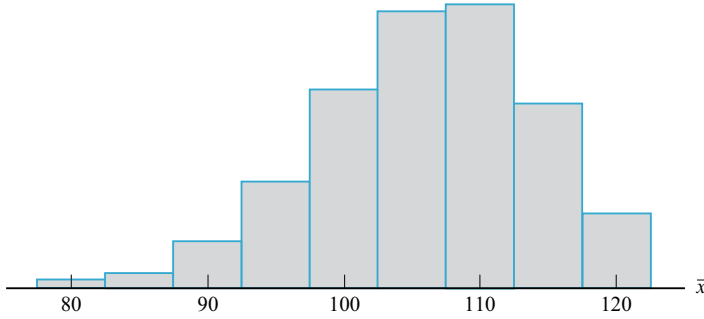
Figure 6.3 Probability histogram for $\overline{X}$ based on $n = 4$ in Example 6.2 ■

Example 6.2 should suggest first of all that the computation of $p_{\overline{X}}(\overline{x})$ and $p_{S^2}(s^2)$ can be tedious. If the original distribution (6.1) had allowed for more than the three possible values 80, 100, and 120, then even for $n = 2$ the computations would have been more involved. The example should also suggest, however, that there are some general relationships between $E(\overline{X}), V(\overline{X}), E(S^2)$, and the mean μ and variance σ^2 of the original distribution. These are stated in the next section. Now consider an example in which the random sample is drawn from a continuous distribution.

Example 6.3 The time that it takes to serve a customer at the cash register in a minimarket is a random variable having an exponential distribution with parameter λ. Suppose X_1 and X_2 are service times for two different customers, assumed independent of each other. Consider the total service time $T_o = X_1 + X_2$ for the two customers, also a statistic. The cdf of T_o is, for $t \geq 0$,

$$F_{T_o}(t) = P(X_1 + X_2 \leq t) = \int\int\limits_{\{(x_1,x_2):x_1+x_2 \leq t\}} f(x_1, x_2) \, dx_1 \, dx_2$$

$$= \int_0^t \int_0^{t-x_1} \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \, dx_2 \, dx_1$$

$$= \int_0^t (\lambda e^{-\lambda x_1} - \lambda e^{-\lambda t}) dx_1 = 1 - e^{-\lambda t} - \lambda t e^{-\lambda t}$$

The region of integration is pictured in Figure 6.4.
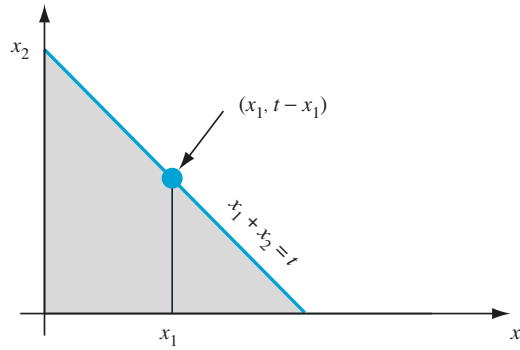
Figure 6.4 Region of integration to obtain cdf of T_o in Example 6.3

The pdf of T_o is obtained by differentiating $F_{T_o}(t)$:

$$f_{T_o}(t) = \lambda^2 t e^{-\lambda t} \quad t \geq 0 \tag{6.4}$$

This is a gamma pdf ($\alpha = 2$ and $\beta = 1/\lambda$). This distribution for T_o can also be derived by convolution or by the moment generating function argument from Section 5.3.

Since $F_{\overline{X}}(\bar{x}) = P(\overline{X} \leq \bar{x}) = P(T_o \leq 2\bar{x}) = F_{T_o}(2\bar{x})$, differentiating with respect to $\bar{x}$ and using (6.4) plus the chain rule gives us the pdf of $\overline{X} = T_o/2$:

$$f_{\overline{X}}(\bar{x}) = 4\lambda^2 \bar{x} e^{-2\lambda\bar{x}} \quad \bar{x} \geq 0 \tag{6.5}$$

The mean and variance of the underlying exponential distribution are $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$. Using Expressions (6.4) and (6.5), it can be verified that $E(\overline{X}) = 1/\lambda, V(\overline{X}) = 1/(2\lambda^2), E(T_o) = 2/\lambda$, and $V(T_o) = 2/\lambda^2$. These results again suggest some general relationships between means and variances of $\overline{X}, T_o$, and the underlying distribution. ∎

Simulation Experiments

The second method of obtaining information about a statistic's sampling distribution is to perform a *simulation experiment*. This method is often used when a derivation via probability rules or properties of distributions is too difficult or complicated to be carried out. Simulations are virtually always done with the aid of computer software. The following characteristics of a simulation experiment must be specified:

1. The statistic of interest ($\overline{X}$, S, a particular trimmed mean, etc.)
2. The population distribution (normal with $\mu = 100$ and $\sigma = 15$, uniform with lower limit $A = 5$ and upper limit $B = 10$, etc.)
3. The sample size n (e.g., $n = 10$ or $n = 50$)
4. The number of replications k (e.g., $k = 10{,}000$).

Then use a computer to obtain k different random samples, each of size n, from the designated population distribution. For each such sample, calculate the value of the statistic and construct a histogram of the k calculated values. This histogram gives the *approximate* sampling distribution of the statistic. The larger the value of k, the better the approximation will tend to be (the actual sampling distribution emerges as $k \to \infty$). In practice, $k = 10{,}000$ may be enough for a "fairly simple" statistic and population distribution, but modern computers allow for a much larger number of replications.

Example 6.4 Consider a simulation experiment in which the population distribution is quite skewed. Figure 6.5 shows the density curve for lifetimes of a certain type of electronic control. This is actually a lognormal distribution with $E[\ln(X)] = 3$ and $V[\ln(X)] = 0.16$; that is, $\ln(X)$ is normal with mean 3 and standard deviation 0.4.
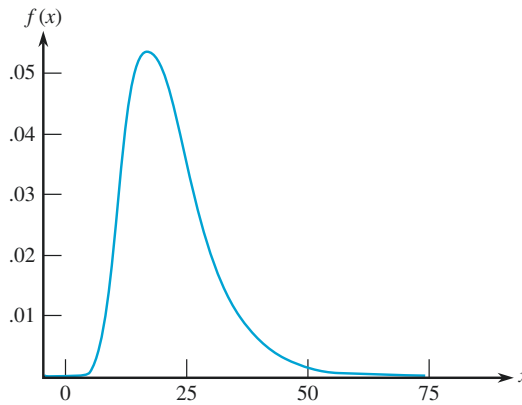
Figure 6.5 Density curve for the simulation experiment of Example 6.4:
a lognormal distribution with $E(X) = 21.76$ and $V(X) = 82.14$

Imagine the statistic of interest is the sample mean, $\overline{X}$. For any given sample size n, we repeat the following procedure k times:

- Generate values $x_1, \ldots, x_n$ from a lognormal distribution with the specified parameter values; equivalently, generate $y_1, \ldots, y_n$ from a $N(3, 0.4)$ distribution and apply the transformation $x = e^y$ to each value.
- Calculate and store the sample mean $\bar{x}$ of the n x-values.

We performed this simulation experiment at four different sample sizes: $n = 5, 10, 20,$ and 30. The experiment utilized $k = 1000$ replications (a very modest value) for each sample size. The resulting histograms along with a normal probability plot from R for the 1000 $\bar{x}$ values based on $n = 30$ are shown in Figure 6.6 on the next page.

The first thing to notice about the histograms is that each one is centered approximately at the mean of the population being sampled, $\mu_X = e^{3+0.16/2} \approx 21.76$. Had the histograms been based on an unending sequence of $\bar{x}$ values, their centers would have been exactly at the population mean.

Second, note the spread of the histograms relative to each other. The smaller the value of n, the greater the extent to which the sampling distribution spreads out about the mean value. This is why the histograms for $n = 20$ and $n = 30$ are based on narrower class intervals than those for the two smaller sample sizes. For the larger sample sizes, most of the $\bar{x}$ values are quite close to μ_X. This is the effect of averaging. When n is small, a single unusual x value can result in an $\bar{x}$ value far from the center. With a larger sample size, any unusual x values, when averaged in with the other sample values, still tend to yield an $\bar{x}$ value close to μ_X. Combining these insights yields an intuitively-appealing result: $\overline{X}$ *based on a large n tends to be closer to μ than does* $\overline{X}$ *based on a small n.*

Third and finally, consider the shapes of the histograms. Recall from Figure 6.5 that the population from which the samples were drawn is quite skewed. But as the sample size n increases, the distribution of $\overline{X}$ appears to become progressively less skewed. In particular, when $n = 30$ the

distribution of the 1000 $\bar{x}$ values appears to be approximately normal, a fact validated by the normal probability plot in Figure 6.6e. We will discover in the next section that this is part of a much broader phenomenon known as the Central Limit Theorem: *as the sample size n increases, the sampling distribution of $\overline{X}$ becomes increasingly normal, irrespective of the population distribution from which values were sampled.*
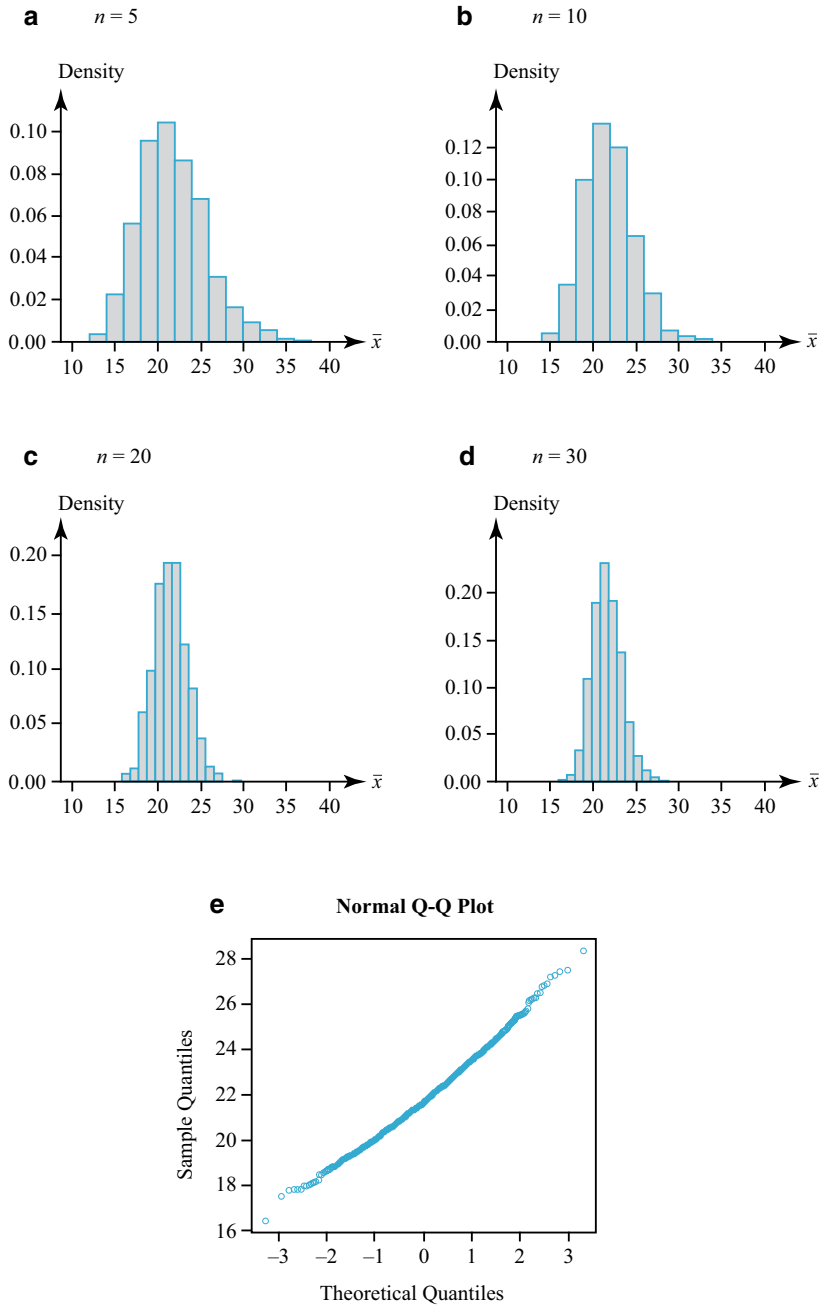
Figure 6.6 Results of the simulation experiment of Example 6.4: (**a**) $\overline{X}$ histogram for $n = 5$; (**b**) $\overline{X}$ histogram for $n = 10$; (**c**) $\overline{X}$ histogram for $n = 20$; (**d**) $\overline{X}$ histogram for $n = 30$; (**e**) normal probability plot for $n = 30$ (from R) ∎

Example 6.5 The 2017 study described in Example 4.23 determined that the variable X = *proximal grip distance for female surgeons* follows a normal distribution with mean 6.58 cm and standard deviation 0.50 cm. Consider the statistic Q_1 = the sample 25th percentile (equivalently, the lower quartile). To investigate the sampling distribution of Q_1 we repeated the following procedure k = 1000 times:

- Generate a sample $x_1, \ldots, x_n$ from the $N(6.58, 0.50)$ distribution.
- Calculate and store the lower quartile, q_1, of the n resulting x values.

The results of two such simulation experiments—one for n = 5, another for n = 40—are shown in Figure 6.7. Similar to $\overline{X}$'s behavior in the previous example, we see that the sampling distribution of Q_1 has greater variability for small n than for large n. Both sampling distributions appear to be centered roughly at 6.5 cm, which is perhaps not surprising: the 25th percentile of the *population* distribution is

$$\eta_{.25} = \mu + \Phi^{-1}(.25) \cdot \sigma = 6.83 + (-0.675)(0.50) \approx 6.49\,\text{cm}$$
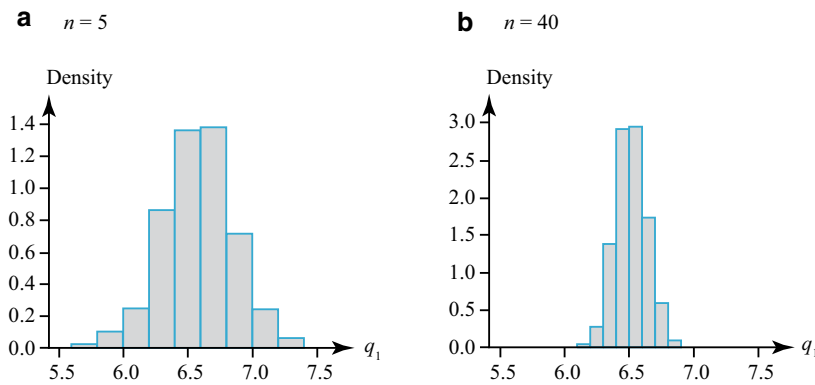
Figure 6.7 Sample histograms of Q_1 based on 1000 samples, each consisting of n observations: (**a**) n = 5, (**b**) n = 40

In fact, even with an infinite set of replications (i.e., the "true" sampling distribution), the mean of Q_1 is not exactly $\eta_{.25}$, but that difference decreases as n increases. ∎

Exercises: Section 6.1 (1–10)

1. A particular brand of dishwasher soap is sold in three sizes: 25, 40, and 65 oz. 20% of all purchasers select a 25-oz box, 50% select a 40-oz box, and the remaining 30% choose a 65-oz box. Let X_1 and X_2 denote the package sizes selected by two independently selected purchasers.

   a. Determine the sampling distribution of $\overline{X}$, calculate $E(\overline{X})$, and compare to μ.
   b. Determine the sampling distribution of the sample variance S^2, calculate $E(S^2)$, and compare to σ^2.

2. There are two traffic lights on the way to work. Let X_1 be the number of lights that

are red, requiring a stop, and suppose that the distribution of X_1 is as follows:

x_1	0	1	2
$p(x_1)$	.2	.5	.3

$\mu = 1.1$, $\sigma^2 = .49$

Let X_2 be the number of lights that are red on the way home; X_2 is independent of X_1. Assume that X_2 has the same distribution as X_1, so that X_1, X_2 is a random sample of size $n = 2$.

a. Let $T_o = X_1 + X_2$, and determine the probability distribution of T_o.

b. Calculate μ_{T_o}. How does it relate to μ, the population mean?

c. Calculate $\sigma^2_{T_o}$. How does it relate to σ^2, the population variance?

3. It is known that 80% of all Brand A MP3 players work in a satisfactory manner throughout the warranty period (are "successes"). Suppose that $n = 10$ players are randomly selected. Let $X =$ the number of successes in the sample. The statistic X/n is the sample proportion (fraction) of successes. Obtain the sampling distribution of this statistic. [*Hint*: One possible value of X/n is .3, corresponding to $X = 3$. What is the probability of this value (what kind of random variable is X)?]

4. A box contains ten sealed envelopes numbered 1, …, 10. The first five contain no money, the next three each contain $5, and there is a $10 bill in each of the last two. A sample of size 3 is selected *with* replacement (so we have a random sample), and you get the largest amount in any of the envelopes selected. If X_1, X_2, and X_3 denote the amounts in the selected envelopes, the statistic of interest is $M =$ the maximum of X_1, X_2, and X_3.

a. Obtain the probability distribution of this statistic.

b. Describe how you would carry out a simulation experiment to compare the distributions of M for various sample sizes. How would you guess the distribution would change as n increases?

5. Let X be the number of packages being mailed by a randomly selected customer at a shipping facility. Suppose the distribution of X is as follows:

x	1	2	3	4
$p(x)$	.4	.3	.2	.1

a. Consider a random sample of size $n = 2$ (two customers), and let $\overline{X}$ be the sample mean number of packages shipped. Obtain the sampling distribution of $\overline{X}$.

b. Refer to part (a) and calculate $P(\overline{X} \le 2.5)$.

c. Again consider a random sample of size $n = 2$, but now focus on the statistic $R =$ the sample range (difference between the largest and smallest values in the sample). Obtain the sampling distribution of R. [*Hint*: Calculate the value of R for each outcome and use the probabilities from part (a).]

d. If a random sample of size $n = 4$ is selected, what is $P(\overline{X} \le 1.5)$? [*Hint*: You should not have to list all possible outcomes, only those for which $\bar{x} \le 1.5$.]

6. A company maintains three offices in a region, each staffed by two employees. Information concerning yearly salaries (1000s of dollars) is as follows:

Office	1	1	2	2	3	3
Employee	1	2	3	4	5	6
Salary	29.7	33.6	30.2	33.6	25.8	29.7

a. Suppose two of these employees are randomly selected from among the six (without replacement). Determine the sampling distribution of the sample mean salary $\overline{X}$.

b. Suppose one of the three offices is randomly selected. Let X_1 and X_2 denote the salaries of the two employees. Determine the sampling distribution of $\overline{X}$.

c. How does $E(\overline{X})$ from parts (a) and (b) compare to the population mean salary μ?

7. The number of dirt specks on a randomly selected square yard of polyethylene film of a certain type has a Poisson distribution with a mean value of 2 specks per square yard. Consider a random sample of $n = 5$ film specimens, each having area 1 square yard, and let $\overline{X}$ be the resulting sample mean number of dirt specks. Obtain the first 21 probabilities in the $\overline{X}$ sampling distribution. [*Hint*: What does a moment generating function argument say about the distribution of $X_1 + \cdots + X_5$?]

8. Suppose the amount of liquid dispensed by a machine is uniformly distributed with lower limit $A = 8$ oz and upper limit $B = 10$ oz. Describe how you would carry out simulation experiments to compare the sampling distribution of the sample iqr for sample sizes $n = 5$, 10, 20, and 30.

9. Carry out a simulation experiment using a statistical computer package or other software to study the sampling distribution of $\overline{X}$ when the population distribution is Weibull with $\alpha = 2$ and $\beta = 5$, as in Example 6.1. Consider the four sample sizes $n = 5$, 10, 20, and 30, and in each case use at least 1000 replications. For which of these sample sizes does the $\overline{X}$ sampling distribution appear to be approximately normal?

10. Carry out a simulation experiment using a statistical computer package or other software to study the sampling distribution of $\overline{X}$ when the population distribution is lognormal with $E[\ln(X)] = 3$ and $V[\ln(X)] = 1$. Consider the four sample sizes $n = 10$, 20, 30, and 50, and in each case use at least 1000 replications. For which of these sample sizes does the $\overline{X}$ sampling distribution appear to be approximately normal?

6.2   The Distribution of Sample Totals, Means, and Proportions

Throughout this section, we will be primarily interested in the properties of two particular rvs derived from random samples: the sample total T_o and the sample mean $\overline{X}$:

$$T_o = X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i, \quad \overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{T_o}{n}$$

The importance of the sample mean $\overline{X}$ springs from its use in drawing conclusions about the population mean μ. Some of the most frequently used inferential procedures are based on properties of the sampling distribution of $\overline{X}$. A preview of these properties appeared in the calculations and simulation experiments of the previous section, where we noted relationships between $E(\overline{X})$ and μ and also among $V(\overline{X})$, σ^2, and n.

PROPOSITION   Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value μ and standard deviation σ. Then

1. $E(T_o) = n\mu$

2. $V(T_o) = n\sigma^2$ and $\sigma_{T_o} = \sqrt{n}\sigma$

3. If the X_i's are normally distributed, then T_o is also normally distributed.

1. $E(\overline{X}) = \mu$

2. $V(\overline{X}) = \dfrac{\sigma^2}{n}$ and $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$

3. If the X_i's are normally distributed, then $\overline{X}$ is also normally distributed.

Proof From the main theorem of Section 5.3, the expected value of a sum is the sum of the individual expected values; moreover, when the variables in the sum are independent, the variance of the sum is the sum of the individual variances:

$$E(T_o) = E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = \mu + \cdots + \mu = n\mu$$
$$V(T_o) = V(X_1 + \cdots + X_n) = V(X_1) + \cdots + V(X_n) = \sigma^2 + \cdots + \sigma^2 = n\sigma^2$$
$$\sigma_{T_o} = \sqrt{n\sigma^2} = \sqrt{n}\sigma$$

The corresponding results for $\overline{X}$ can be derived by writing $\overline{X} = \frac{1}{n} \cdot T_o$ and using basic rescaling properties, such as $E(cY) = cE(Y)$. Property 3 is a consequence of the more general result from Section 5.3 that *any* linear combination of independent normal rvs is normal. ∎

According to Property 1, the distribution of $\overline{X}$ is centered precisely at the mean of the population from which the sample has been selected. If the sample mean is used to compute an estimate (educated guess) of the population mean μ, there will be no systematic tendency for the estimate to be too large or too small.

Property 2 shows that the $\overline{X}$ distribution becomes more concentrated about μ as the sample size n increases, because its standard deviation decreases. In marked contrast, the distribution of T_o becomes more spread out as n increases. Averaging moves probability in toward the middle, whereas totaling spreads probability out over a wider and wider range of values. The expression $\sigma/\sqrt{n}$ for the standard deviation of $\overline{X}$ is called the **standard error of the mean**, and it indicates the typical amount by which a value of $\overline{X}$ will deviate from the true mean, μ (in contrast, σ itself represents the typical difference between an *individual* X_i and μ).

When σ is unknown, as is usually the case when μ is unknown and we are trying to estimate it, we may substitute the sample standard deviation, s, of our sample into the standard error formula and say that an observed value of $\overline{X}$ will typically differ by about $s/\sqrt{n}$ from μ. This is the estimated standard error formula presented in Sections 3.8 and 4.8.

Finally, Property 3 says that $\overline{X}$ and T_o are both normally distributed when the population distribution is normal. In particular, probabilities such as $P(a \leq \overline{X} \leq b)$ and $P(c \leq T_o \leq d)$ can be obtained simply by standardizing, with the appropriate means and standard deviations provided by Properties 1 and 2. Figure 6.8 illustrates the $\overline{X}$ part of the proposition.
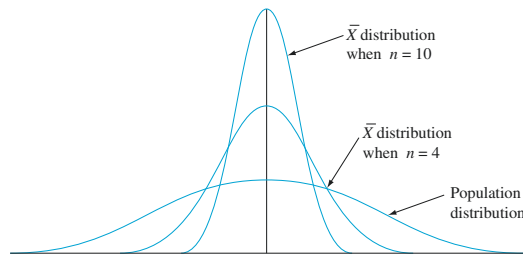
Figure 6.8 A normal population distribution and $\overline{X}$ sampling distributions

Example 6.6 The amount of time that a patient spends in a certain outpatient surgery center is a random variable with a mean value of 4.5 h and a standard deviation of 1.4 h. Let $X_1, \ldots, X_{25}$ be the times for a random sample of 25 patients. Then the expected total time for the 25 patients is

$E(T_o) = n\mu = 25(4.5) = 112.5$ h, whereas the expected sample mean amount of time is $E(\overline{X}) = \mu = 4.5$ h. The standard deviations of T_o and $\overline{X}$ are

$$\sigma_{T_o} = \sqrt{n}\sigma = \sqrt{25}(1.4) = 7 \text{ h}$$
$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.4}{\sqrt{25}} = .28 \text{ h}$$

Suppose further that such patient times follow a normal distribution; i.e., $X_i \sim N(4.5, 1.4)$. Then the total time spent by 25 randomly selected patients in this center is also normal: $T_o \sim N(112.5, 7)$. The probability their total time exceeds five days (120 h) is

$$P(T_o > 120) = 1 - P(T_o \le 120) = 1 - \Phi\left(\frac{120 - 112.5}{7}\right) = 1 - \Phi(1.07) = .8577$$

This same probability can be reframed in terms of $\overline{X}$: for 25 patients, a total time of 120 h equates to an average time of 120/25 = 4.8 h, and since $\overline{X} \sim N(4.5, .28)$,

$$P(\overline{X} > 4.8) = 1 - \Phi\left(\frac{4.8 - 4.5}{.28}\right) = 1 - \Phi(1.07) = .8577 \qquad \blacksquare$$

Example 6.7 Resistors used in electronics manufacturing are labeled with a "nominal" resistance as well as a percentage tolerance. For example, a 330-Ω resistor with a 5% tolerance is anticipated to have an actual resistance between 313.5 and 346.5 Ω. Consider five such resistors, randomly selected from the population of all resistors with those specifications, and model the resistance of each by a uniform distribution on [313.5, 346.5]. If these are connected in series, the resistance R of the system is given by $R = X_1 + \cdots + X_5$, where the X_i's are the iid uniform resistances.

   A random variable uniformly distributed on $[A, B]$ has mean $(A + B)/2$ and standard deviation $(B - A)/\sqrt{12}$. For our uniform model, the mean resistance is $E(X_i) = (313.5 + 346.5)/2 = 330$ Ω, the nominal resistance, with a standard deviation of $(346.5 - 313.5)/\sqrt{12} = 9.526 \, \Omega$. The system's resistance has mean and standard deviation

$$E(R) = n\mu = 5(330) = 1650 \, \Omega, \quad \sigma_R = \sqrt{n}\sigma = \sqrt{5}(9.526) = 21.3 \, \Omega$$

But what is the probability distribution of R? Is R also uniformly distributed? Determining the exact pdf of R is difficult (it requires four convolutions). And the mgf of R, while easy to obtain, is not recognizable as coming from any particular family of known distributions. Instead, we resort to a simulation of R, the results of which appear in Figure 6.9. For 10,000 iterations, five independent uniform variates on [313.5, 346.5] were created and summed; see Section 4.8 for information on simulating values from a uniform distribution. The histogram in Figure 6.9 clearly indicates that R is not uniform; in fact, if anything, R appears (from the simulation, anyway) to be approximately normally distributed!

Histogram of R
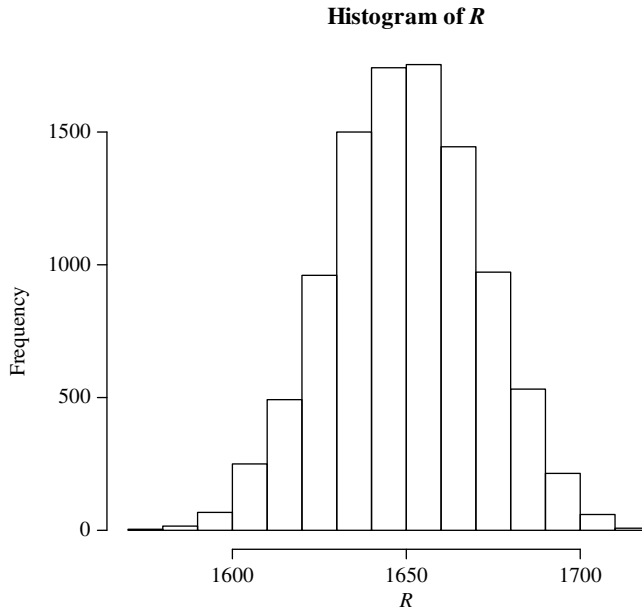
Figure 6.9 Simulated distribution of the random variable R in Example 6.7 ∎

The Central Limit Theorem

When iid X_i's are normally distributed, so are T_o and $\overline{X}$ for every sample size n. The simulation results from Example 6.7 suggest that even when the population distribution is not normal, summing (or averaging) produces a distribution more bell-shaped than the one being sampled. Upon reflection, this is quite intuitive: in order for R to be near $5(346.5) = 1732.5$, its theoretical maximum, all five randomly selected resistors would have to exert resistances at the high end of their common range (i.e., every X_i would have to be near 346.5). Thus, R-values near 1732.5 are unlikely, and the same applies to R's theoretical minimum of $5(313.5) = 1567.5$. On the other hand, there are many ways for R to be near the mean value of 1650: all five resistances in the middle, two low and one middle and two high, and so on. Thus, R is more likely to be "centrally" located than out at the extremes. (This is analogous to the well-known fact that rolling a pair of dice is far more likely to result in a sum of 7 than 2 or 12, because there are more ways to obtain 7.)

This general pattern of behavior for sample totals and sample means is formalized by the most important theorem of probability, the *Central Limit Theorem* (CLT).

CENTRAL LIMIT THEOREM (CLT) Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and standard deviation σ. Then, in the limit as $n \to \infty$, the standardized versions of $\overline{X}$ and T_o have the standard normal distribution. That is,

$$\lim_{n \to \infty} P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le z\right) = P(Z \le z) = \Phi(z)$$

and

$$\lim_{n\to\infty} P\left(\frac{T_o - n\mu}{\sqrt{n}\sigma} \le z\right) = P(Z \le z) = \Phi(z)$$

where Z is a standard normal rv. It is customary to say that $\overline{X}$ and T_o are **asymptotically normal**, and that their standardized versions **converge in distribution** to Z. Thus when n is sufficiently large, $\overline{X}$ has approximately a normal distribution with $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. Equivalently, for large n the sum T_o has approximately a normal distribution with $\mu_{T_o} = n\mu$ and $\sigma_{T_o} = \sqrt{n}\sigma$.

Figure 6.10 illustrates the Central Limit Theorem. A partial proof of the CLT appears in the appendix to this chapter. It is shown that, if the moment generating function exists, then the mgf of the standardized $\overline{X}$ (and of T_o) approaches the standard normal mgf. With the aid of an advanced probability theorem, this implies the CLT statement about convergence of probabilities.
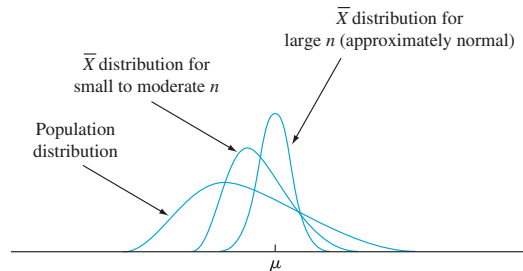
Figure 6.10   The Central Limit Theorem for $\overline{X}$ illustrated

A practical difficulty in applying the CLT is in knowing when n is "sufficiently large." The problem is that the accuracy of the approximation for a particular n depends on the shape of the original underlying distribution being sampled. If the underlying distribution is symmetric and there is not much probability far out in the tails, then the approximation will be good even for a small n, whereas if it is highly skewed or has "heavy" tails, then a large n will be required. For example, if the distribution is uniform on an interval, then it is symmetric with no probability in the tails, and the normal approximation is very good for n as small as 10 (in Example 6.9, even for $n = 5$, the distribution of the sample total appeared rather bell-shaped). However, at the other extreme, a distribution can have such fat tails that its mean fails to exist and the Central Limit Theorem does not apply, so no n is big enough. A popular, although frequently somewhat conservative, convention is that the Central Limit Theorem may be safely applied when $n > 30$. Of course, there are exceptions, but this rule applies to most distributions of real data.

Example 6.8   When a batch of a certain chemical product is prepared, the amount of a particular impurity in the batch is a random variable with mean value 4.0 g and standard deviation 1.5 g. If 50 batches are independently prepared, what is the (approximate) probability that the sample average amount of impurity $\overline{X}$ is between 3.5 and 3.8 g? According to the convention mentioned above, $n = 50$ is large enough for the CLT to be applicable. The sample mean $\overline{X}$ then has approximately a normal distribution with mean value $\mu_{\overline{X}} = 4.0$ and $\sigma_{\overline{X}} = 1.5/\sqrt{50} = .2121$, so

$$P(3.5 \leq \overline{X} \leq 3.8) \approx P\left(\frac{3.5 - 4.0}{.2121} \leq Z \leq \frac{3.8 - 4.0}{.2121}\right)$$
$$= \Phi(-.94) - \Phi(-2.36) = .1645 \qquad \blacksquare$$

Example 6.9 Suppose the number of times a randomly selected customer of a large bank uses the bank's ATM during a particular period is a random variable with a mean value of 3.2 and a standard deviation of 2.4. Among 100 randomly selected customers, how likely is it that the sample mean number of times the bank's ATM is used exceeds 4? Let X_i denote the number of times the ith customer in the sample uses the bank's ATM. Notice that X_i is a discrete rv, but the CLT is not limited to continuous random variables. Also, although the fact that the standard deviation of this nonnegative variable is quite large relative to the mean value suggests that its distribution is positively skewed, the large sample size implies that $\overline{X}$ does have approximately a normal distribution. Using $\mu_{\overline{X}} = 3.2$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 2.4/\sqrt{100} = .24$,

$$P(\overline{X} > 4) \approx P\left(Z > \frac{4 - 3.2}{.24}\right) = 1 - \Phi(3.33) = .0004 \qquad \blacksquare$$

Example 6.10 Consider the distribution shown in Figure 6.11 for the amount purchased (rounded to the nearest dollar) by a randomly selected customer at a particular gas station. (A similar distribution for purchases in Britain (in £) appeared in the article "Data Mining for Fun and Profit," *Stat. Sci.* 2000: 111–131; there were big spikes at the values 10, 15, 20, 25, and 30.) The distribution is obviously quite nonnormal.
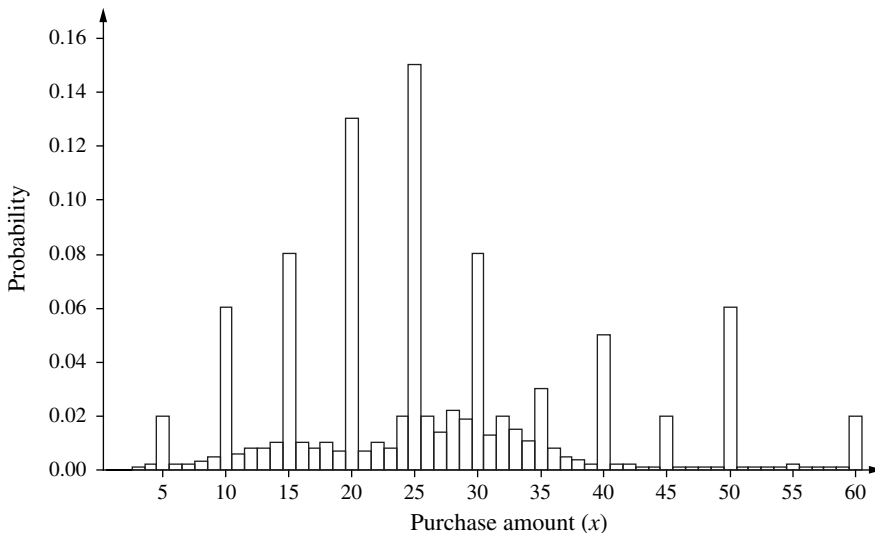
Figure 6.11 Probability distribution of X = amount of gasoline purchased (\$) in Example 6.10

We asked R to select 1000 different samples, each consisting of $n = 15$ observations, and calculate the value of the sample mean $\overline{X}$ for each one. Figure 6.12 is a histogram of the resulting 1000 values; this is the approximate sampling distribution of $\overline{X}$ under the specified circumstances. This distribution is clearly approximately normal even though the sample size is not all that large. As further evidence

for normality, Figure 6.13 shows a normal probability plot of the 1000 $\bar{x}$ values; the linear pattern is very prominent. It is typically not nonnormality in the central part of the population distribution that causes the CLT to fail, but instead very substantial skewness or extremely heavy tails.
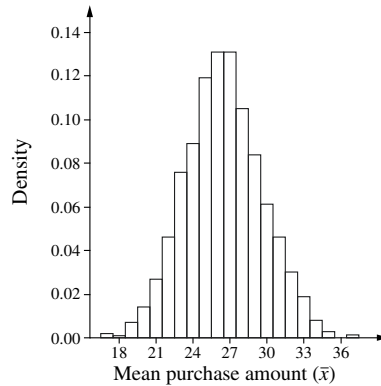
Figure 6.12   Approximate sampling distribution of the sample mean amount purchased when $n = 15$ and the population distribution is as shown in Figure 6.11
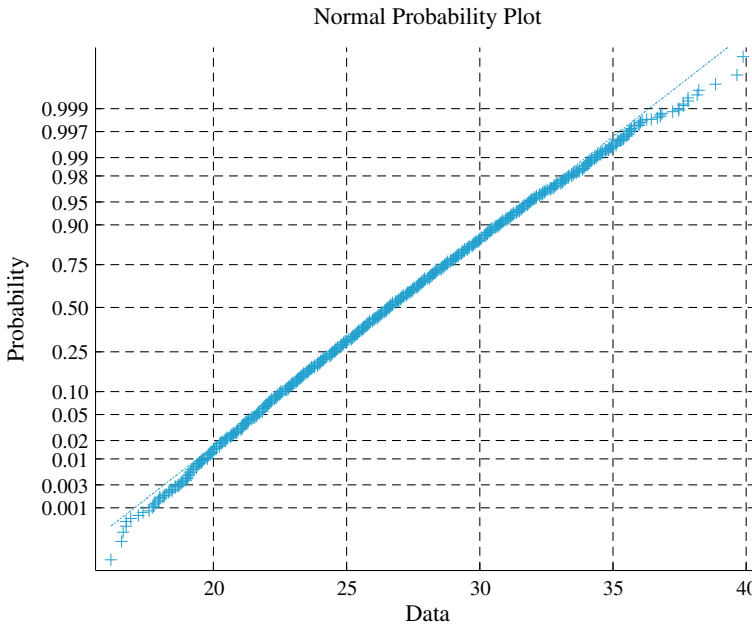
Figure 6.13   Normal probability plot of the 1000 $\bar{x}$ values based on samples of size $n = 15$ ∎

The CLT can also be generalized so it applies to nonidentically-distributed independent random variables and certain linear combinations. Roughly speaking, if n is large and no individual term is likely to contribute too much to the overall value, then asymptotic normality prevails (see Exercise 68). It can also be generalized to sums of variables which are not independent, provided the extent of dependence between most pairs of variables is not too strong.

Other Applications of the Central Limit Theorem

The CLT can be used to justify the normal approximation to the binomial distribution discussed in Chapter 4. Recall that a binomial variable X is the number of successes in a binomial experiment consisting of n independent success/failure trials with $p = P(\text{success})$ for any particular trial. Define new rvs $X_1, X_2, \ldots, X_n$ by

$$X_i = \begin{cases} 1 & \text{if the } i\text{th trial results in a success} \\ 0 & \text{if the } i\text{th trial results in a failure} \end{cases} \quad (i = 1, \ldots, n)$$

Because the trials are independent and $P(\text{success})$ is constant from trial to trial, the X_i's are iid (a random sample from a Bernoulli distribution). When the X_i's are summed, a 1 is added for every success that occurs and a 0 for every failure so $X = X_1 + \cdots + X_n$, their total. The sample mean of the X_i's is $\overline{X} = X/n$, the sample proportion of successes, which in previous discussions we have denoted $\hat{P}$. The CLT then implies that if n is sufficiently large, both X and $\hat{P}$ are approximately normal when n is large. We summarize properties of the $\hat{P}$ distribution in the following corollary; Statements 1 and 2 were derived in Section 3.5.

COROLLARY   Consider an event A in the sample space of some experiment with $p = P(A)$. Let $X =$ the number of times A occurs when the experiment is repeated n independent times, and define

$$\hat{P} = \hat{P}(A) = \frac{X}{n}$$

Then

1. $E(\hat{P}) = p$
2. $V(\hat{P}) = \dfrac{p(1-p)}{n}$ and $\sigma_{\hat{P}} = \sqrt{\dfrac{p(1-p)}{n}}$
3. As n increases, the distribution of $\hat{P}$ approaches a normal distribution.

In practice, Property 3 is taken to say that $\hat{P}$ is approximately normal, provided that $np \geq 10$ and $n(1-p) \geq 10$.

The necessary sample size for this approximation depends on the value of p: When p is close to .5, the distribution of each Bernoulli X_i is reasonably symmetric (see Figure 6.14), whereas the distribution is quite skewed when p is near 0 or 1. Using the approximation only if both $np \geq 10$ and $n(1-p) \geq 10$ ensures that n is large enough to overcome any skewness in the underlying Bernoulli distribution.
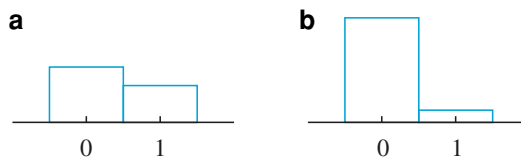
Figure 6.14   Two Bernoulli distributions: (**a**) $p = .4$ (reasonably symmetric); (**b**) $p = .1$ (very skewed)

Example 6.11 A computer simulation in the style of Section 2.6 is used to determine the probability that a complex system of components operates properly throughout the warranty period. Unknown to the investigator, the true probability is $P(A) = .18$. If 10,000 simulations of the underlying process are run, what is the chance the estimated probability $\hat{P}(A)$ will be within .01 of the true probability $P(A)$?

Apply the preceding corollary, with $n = 10,000$ and $p = P(A) = .18$. The expected value of $\hat{P}(A)$ is $p = .18$, and the standard deviation is $\sigma_{\hat{P}} = \sqrt{.18(.82)/10,000} = .00384$. Since $np = 1800 \geq 10$ and $n(1 - p) = 8200 \geq 10$, a normal distribution can safely be used to approximate the distribution of $\hat{P}(A)$. This sample proportion is within .01 of the true probability if and only if $.17 < \hat{P}(A) < .19$, so the desired likelihood is approximately

$$P(.17 < \hat{P} < .19) \approx P\left(\frac{.17 - .18}{.00384} < Z < \frac{.19 - .18}{.00384}\right) = \Phi(2.60) - \Phi(-2.60) = .9906 \quad \blacksquare$$

The normal distribution serves as a reasonable approximation to the binomial pmf when n is large because the binomial distribution is *additive*; i.e., a binomial rv can be expressed as the sum of other, iid rvs. Other additive distributions include the Poisson, negative binomial, gamma, and (of course) normal distributions; some of these were discussed at the end of Section 5.3. In particular, CLT justifies normal approximations to the following distributions:

- Poisson, when μ is large
- Negative binomial, when r is large
- Gamma, when α is large

As a final application of the CLT, first recall from Section 4.5 that X has a lognormal distribution if $\ln(X)$ has a normal distribution.

PROPOSITION   Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution for which only positive values are possible $[P(X_i > 0) = 1]$. Then if n is sufficiently large, the product $Y = X_1 \ X_2 \cdot \cdots \cdot \ X_n$ has approximately a lognormal distribution; that is, $\ln(Y)$ has a normal distribution.

To verify this, note that

$$\ln(Y) = \ln(X_1) + \ln(X_2) + \cdots + \ln(X_n)$$

Since $\ln(Y)$ is a sum of independent and identically distributed rvs [the $\ln(X_i)$'s], it is approximately normal when n is large, so Y itself has approximately a lognormal distribution. As an example of the applicability of this result, it has been argued that the damage process in plastic flow and crack propagation is a multiplicative process, so that variables such as percentage elongation and rupture strength have approximately lognormal distributions.

The Law of Large Numbers

In the simulation sections of Chapters 2–4, we described how a sample proportion $\hat{P}$ could estimate a true probability p, and a sample mean $\overline{X}$ served to approximate a theoretical expected value μ. Moreover, in both cases the precision of the estimation improves as the number of simulation runs, n,

increases. We would like to be able to say that our estimates "converge" to the correct answers in some sense. Such a convergence statement is justified by another important theoretical result, called the *Law of Large Numbers*.

To begin, recall the first proposition in this section: If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ and standard deviation σ, then $E(\overline{X}) = \mu$ and $V(\overline{X}) = \sigma^2/n$. As n increases, the expected value of $\overline{X}$ remains at μ but the variance approaches zero; that is, $E[(\overline{X} - \mu)]^2 = V(\overline{X}) = \sigma^2/n \to 0$. We say that $\overline{X}$ converges *in mean square* to μ because the mean of the squared difference between $\overline{X}$ and μ goes to zero. This is one form of the Law of Large Numbers.

Another form of convergence states that as the sample size n increases, $\overline{X}$ is increasingly unlikely to differ by any set amount from μ. More precisely, let ε be a positive number close to 0, such as .01 or .001, and consider $P(|\overline{X} - \mu| \geq \varepsilon)$, the probability that $\overline{X}$ differs from μ by at least ε (at least .01, at least .001, etc.). We will prove shortly that, no matter how small the value of ε, this probability will approach zero as $n \to \infty$. Because of this, statisticians say that $\overline{X}$ *converges to μ in probability*.

The two forms of the Law of Large Numbers are summarized in the following theorem.

LAW OF LARGE NUMBERS
If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ, then $\overline{X}$ converges to μ

1. in mean square: $E[(\overline{X} - \mu)^2] \to 0$ as $n \to \infty$
2. in probability: $P(|\overline{X} - \mu| \geq \varepsilon) \to 0$ as $n \to \infty$ for any $\varepsilon > 0$.

Proof The proof of Statement 1 appears a few paragraphs above. For Statement 2, recall Chebyshev's inequality (Exercises 45 and 163 in Chapter 3), which states that for any rv Y, $P(|Y - \mu_Y| \geq k\sigma_Y) \leq 1/k^2$ for any $k \geq 1$ (i.e., the probability that Y is at least k standard deviations away from its mean is at most $1/k^2$). Let $Y = \overline{X}$, so $\mu_Y = E(\overline{X}) = \mu$ and $\sigma_Y = \sigma_{\overline{X}} = \sigma/\sqrt{n}$. Now, for any $\varepsilon > 0$, determine the value of k such that $\varepsilon = k\sigma_Y = k\sigma/\sqrt{n}$; solving for k yields $k = \varepsilon\sqrt{n}/\sigma$, which for sufficiently large n will exceed 1. Apply Chebyshev's inequality:

$$P(|Y - \mu_Y| \geq k\sigma_Y) \leq \frac{1}{k^2} \Rightarrow P\left(|\overline{X} - \mu| \geq \frac{\varepsilon\sqrt{n}}{\sigma} \cdot \frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{(\varepsilon\sqrt{n}/\sigma)^2}$$

$$\Rightarrow P(|\overline{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2 n} \to 0 \text{ as } n \to \infty$$

That is, $P(|\overline{X} - \mu| \geq \varepsilon) \to 0 \to 0$ as $n \to \infty$ for any $\varepsilon > 0$. ∎

Convergence of $\overline{X}$ to μ in probability actually holds even if the variance σ^2 does not exist (a heavy-tailed distribution) as long as μ is finite. But then Chebyshev's inequality cannot be used, and the proof is much more complicated.

In statistical language, the Law of Large Numbers states that $\overline{X}$ is a *consistent estimator* of μ. Other statistics are also consistent estimators of the corresponding parameters. For example, it can be shown that the sample proportion $\hat{P}$ is a consistent estimator of the population proportion p (Exercise 24), and the sample variance $S^2 = \sum (X_i - \overline{X})^2/(n-1)$ is a consistent estimator of the population variance σ^2.

Exercises: Section 6.2 (11–27)

11. The inside diameter of a randomly selected piston ring is a random variable with mean value 12 cm and standard deviation .04 cm.

    a. If $\overline{X}$ is the sample mean diameter for a random sample of $n = 16$ rings, where is the sampling distribution of $\overline{X}$ centered, and what is the standard deviation of the $\overline{X}$ distribution?

    b. Answer the questions posed in part (a) for a sample size of $n = 64$ rings.

    c. For which of the two random samples, the one of part (a) or the one of part (b), is $\overline{X}$ more likely to be within .01 cm of 12 cm? Explain your reasoning.

12. Refer to the previous exercise. Suppose the distribution of diameter is normal.

    a. Calculate $P(11.99 \leq \overline{X} \leq 12.01)$ when $n = 16$.

    b. How likely is it that the sample mean diameter exceeds 12.01 when $n = 25$?

13. The National Health Statistics Reports dated Oct. 22, 2008 stated that for a sample size of 277 18-year-old American males, the sample mean waist circumference was 86.3 cm. A somewhat complicated method was used to *estimate* various population percentiles, resulting in the following values:

5th	10th	25th	50th	75th	90th	95th
69.6	70.9	75.2	81.3	95.4	107.1	116.4

    a. Is it plausible that the waist size distribution is at least approximately normal? Explain your reasoning. If your answer is no, conjecture the shape of the population distribution.

    b. Suppose that the population mean waist size is 85 cm and that the population standard deviation is 15 cm. How likely is it that a random sample of 277 individuals will result in a sample mean waist size of at least 86.3 cm?

    c. Referring back to (b), suppose now that the population mean waist size is 82 cm (closer to the median than the mean).

Now what is the (approximate) probability that the sample mean will be at least 86.3? In light of this calculation, do you think that 82 is a reasonable value for μ?

14. There are 40 students in an elementary statistics class. On the basis of years of experience, the instructor knows that the time needed to grade a randomly chosen first examination paper is a random variable with an expected value of 6 min and a standard deviation of 6 min.

    a. If grading times are independent and the instructor begins grading at 6:50 p.m. and grades continuously, what is the (approximate) probability that he is through grading before the 11:00 p.m. TV news begins?

    b. If the sports report begins at 11:10, what is the probability that he misses part of the report if he waits until grading is done before turning on the TV?

15. The tip percentage at a restaurant has a mean value of 18% and a standard deviation of 6%.

    a. What is the approximate probability that the sample mean tip percentage for a random sample of 40 bills is between 16 and 19%?

    b. If the sample size had been 15 rather than 40, could the probability requested in part (a) be calculated from the given information?

16. The time taken by a randomly selected applicant for a mortgage to fill out a certain form has a normal distribution with mean value 10 min and standard deviation 2 min. If five individuals fill out a form on one day and six on another, what is the probability that the sample average amount of time taken on each day is at most 11 min?

17. The lifetime of a type of battery is normally distributed with mean value 10 h and standard deviation 1 h. There are four

batteries in a package. What lifetime value is such that the total lifetime of all batteries in a package exceeds that value for only 5% of all packages?

18. Let X represent the amount of gasoline (gallons) purchased by a randomly selected customer at a gas station. Suppose that the mean value and standard deviation of X are 11.5 and 4.0, respectively.

   a. In a sample of 50 randomly selected customers, what is the approximate probability that the sample mean amount purchased is at least 12 gallons?

   b. In a sample of 50 randomly selected customers, what is the approximate probability that the total amount of gasoline purchased is at most 600 gallons?

   c. What is the approximate value of the 95th percentile for the total amount purchased by 50 randomly selected customers?

19. Suppose that the fracture angle under pure compression of a randomly selected specimen of fiber reinforced polymer-matrix composite material is normally distributed with mean value 53 and standard deviation 1 (suggested in the article "Stochastic Failure Modelling of Unidirectional Composite Ply Failure," *Reliability Engr. Syst. Safety* 2012: 1–9; this type of material is used extensively in the aerospace industry).

   a. If a random sample of 4 specimens is selected, what is the probability that the sample mean fracture angle is at most 54? Between 53 and 54?

   b. How many such specimens would be required to ensure that the first probability in (a) is at least .999?

20. The first assignment in a statistical computing class involves running a short program. If past experience indicates that 40% of all students will make no programming errors, compute the (approximate) probability that in a class of 50 students

   a. At least 25 will make no errors. [*Hint*: Normal approximation to the binomial.]

   b. Between 15 and 25 (inclusive) will make no errors.

21. The number of parking tickets issued in a certain city on any given weekday has a Poisson distribution with parameter $\mu = 50$. What is the approximate probability that

   a. Between 35 and 70 tickets are given out on a particular day? [*Hint*: When μ is large, a Poisson rv has approximately a normal distribution.]

   b. The total number of tickets given out during a 5-day week is between 225 and 275?

   c. Use software to obtain the exact probabilities in (a) and (b), and compare to the approximations.

22. Suppose the distribution of the time X (in hours) spent by students at a certain university on a particular project is gamma with parameters $\alpha = 50$ and $\beta = 2$. Because α is large, it can be shown that X has approximately a normal distribution. Use this fact to compute the probability that a randomly selected student spends at most 125 h on the project.

23. The Central Limit Theorem says that $\overline{X}$ is approximately normal if the sample size is large. More specifically, the theorem states that the standardized $\overline{X}$ has a limiting standard normal distribution. That is, the rv $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ has a distribution approaching the standard normal. Can you reconcile this with the Law of Large Numbers?

24. Assume a sequence of independent trials, each with probability p of success. Use the Law of Large Numbers to show that the proportion of successes approaches p as the number of trials becomes large.

25. Let Y_n be the largest order statistic in a sample of size n from the uniform distribution on $[0, \theta]$. Show that Y_n converges in probability to θ, that is, that $P(|Y_n - \theta| \geq \varepsilon) \to 0$ as n approaches ∞. [*Hint*: The pdf of the largest

order statistic appears in Section 5.7, so the relevant probability can be obtained by integration (Chebyshev's inequality is not relevant here).]

26. A friend commutes by bus to and from work 6 days/week. Suppose that waiting time is uniformly distributed between 0 and 10 min, and that waiting times going and returning on various days are independent of each other. What is the approximate probability that total waiting time for an entire week is at most 75 min?

27. It can be shown that if Y_n converges in probability to a constant τ, then $h(Y_n)$ converges to $h(\tau)$ for any function $h(\cdot)$ that is continuous at τ. Use this to obtain a consistent estimator for the rate parameter λ of an exponential distribution. [*Hint*: How does μ for an exponential distribution relate to the exponential parameter λ?]

6.3   The χ^2, *t*, and *F* Distributions

The previous section explored the sampling distribution of the sample mean, $\overline{X}$, with particular attention to the special case when our sample $X_1, \ldots, X_n$ is drawn from a normally distributed population. In this section, we introduce three distributions closely related to the normal: the chi-squared (χ^2), *t*, and *F* distributions. These distributions will then be used in the next section to describe the sampling variability of several statistics on which important inferential procedures are based.

The Chi-Squared Distribution

DEFINITION   For a positive integer v, let $Z_1, \ldots, Z_v$ be iid standard normal random variables. Then the **chi-squared distribution with v degrees of freedom (df)** is defined to be the distribution of the rv

$$X = Z_1^2 + \cdots + Z_v^2$$

This will sometimes be denoted by $X \sim \chi_v^2$.

Our first goal is to determine the pdf of this distribution. We start with the $v = 1$ case, where we may write $X = Z_1^2$. As in previous chapters, let $\Phi(z)$ and $\phi(z)$ denote the cdf and pdf, respectively, of the standard normal distribution. Then the cdf of X, for $x > 0$, is given by

$$F(x) = P(X \leq x) = P(Z_1^2 \leq x) = P(-\sqrt{x} \leq Z_1 \leq \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x})$$
$$= \Phi(\sqrt{x}) - [1 - \Phi(\sqrt{x})] = 2\Phi(\sqrt{x}) - 1$$

Above, we've used the symmetry property $\Phi(-z) = 1 - \Phi(z)$ of the standard normal distribution. Differentiate to obtain the pdf for $x > 0$:

$$f(x) = F'(x) = 2\Phi'(\sqrt{x}) \cdot \frac{1}{2\sqrt{x}} - 0 = \phi(\sqrt{x}) \cdot \frac{1}{\sqrt{x}} = \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{x})^2/2} \cdot \frac{1}{\sqrt{x}} = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} \quad (6.6)$$

We have established the χ_1^2 pdf. But this expression looks familiar: comparing (6.6) to the gamma pdf in Expression (4.6), and recalling that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, we find that the χ_1^2 distribution is exactly the same as the gamma distribution with parameters $\alpha = 1/2$ and $\beta = 2$!

To generalize to any number of degrees of freedom, recall that the moment generating function of the gamma distribution is $M(t) = (1 - \beta t)^{-\alpha}$. So, the mgf of a χ_1^2 rv—that is, the mgf of Z^2 when $Z \sim N(0, 1)$—is $M(t) = (1 - 2t)^{-1/2}$. Using the definition of the chi-squared distribution and properties of mgfs, we find that for $X \sim \chi_v^2$,

$$M_X(t) = M_{Z_1^2}(t) \cdot \cdots \cdot M_{Z_v^2}(t) = (1 - 2t)^{-1/2} \cdot \cdots \cdot (1 - 2t)^{-1/2} = (1 - 2t)^{-v/2},$$

which we recognize as the mgf of the gamma distribution with $\alpha = v/2$ and $\beta = 2$. By the uniqueness of mgfs, we have established the following distributional result.

PROPOSITION   The chi-squared distribution with v degrees of freedom is the gamma distribution with $\alpha = v/2$ and $\beta = 2$. In particular, the pdf of the χ_v^2 distribution is

$$f(x; v) = \frac{1}{2^{v/2}\Gamma(v/2)} x^{(v/2)-1} e^{-x/2} \quad x > 0$$

Moreover, if $X \sim \chi_v^2$ then $E(X) = v$, $V(X) = 2v$, and $M_X(t) = (1 - 2t)^{-v/2}$.

The mean and variance stated in the proposition follow from properties of the gamma distribution:

$$\mu = \alpha\beta = \frac{v}{2} \cdot 2 = v, \quad \sigma^2 = \alpha\beta^2 = \frac{v}{2} \cdot 2^2 = 2v$$

Figure 6.15 shows graphs of the chi-squared pdf for 1, 2, 3, and 5 degrees of freedom. Notice that the pdf is unbounded near $x = 0$ for 1 df and the pdf is exponentially decreasing for 2 df. Indeed, the chi-squared for 2 df is exponential with mean 2, $f(x) = \frac{1}{2}e^{-x/2}$ for $x > 0$. If $v > 2$ the pdf is unimodal with a peak at $x = v - 2$, as shown in Exercise 31. The distribution is skewed, but it becomes more symmetric as the number of degrees of freedom increases, and for large df values the distribution is approximately normal (see Exercise 29).
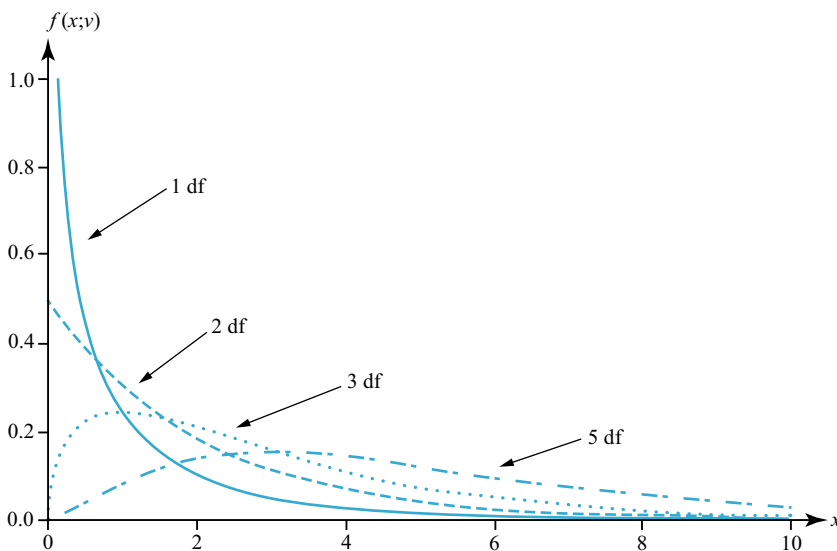
Figure 6.15   The chi-squared pdf for 1, 2, 3, and 5 DF

Without software, it is difficult to integrate a chi-squared pdf, so Table A.5 in the appendix has critical values for chi-squared distributions. For example, the second row of the table is for 2 df, and under the heading .01 the value 9.210 indicates that $P(\chi_2^2 > 9.210) = .01$. We will use the notation $\chi_{.01,2}^2 = 9.210$. In general $\chi_{\alpha,v}^2 = c$ means that $P(\chi_v^2 > c) = \alpha$. Instructions for chi-squared computations using R appear at the end of this section.

Example 6.12   The article "Reliability analysis of LED-based electronic devices" (*Proc. Engr.* 2012: 260–269) uses chi-squared distributions to model the lifecycle, in thousands of hours, of certain LED lamps. In one particular setting, the authors suggest a parameter value of $v = 8$ df. Let X represent this χ_8^2 rv. The mean and standard deviation of X are $E(X) = v = 8$ thousand hours and $SD(X) = \sqrt{2v} = \sqrt{16} = 4$ thousand hours.

We can use the gamma cdf, as illustrated in Chapter 4, to determine probabilities concerning X, because the χ_8^2 distribution is the same as the gamma distribution with $\alpha = 8/2 = 4$ and $\beta = 2$. For instance, the probability an LED lamp of this type has a lifecycle between 6 and 10 thousand hours is

$$P(6 \leq X \leq 10) = G(10/2; 4) - G(6/2; 4) = G(5; 4) - G(3; 4)$$
$$= .735 - .353 = .382$$

Next, what values define the "middle 95%" of lifecycle values for these LED lamps? We desire the .025 and .975 quantiles of the χ_8^2 distribution; from Appendix Table A.5, they are

$$\chi_{.975,8}^2 = 2.180 \quad \text{and} \quad \chi_{.025,8}^2 = 17.534$$

That is, the middle 95% of lifecycle values ranges from 2.180 to 17.534 h.   ■

Given the definition of the chi-squared distribution, the following properties should come as no surprise. Proofs of both statements rely on moment generating functions (Exercises 32 and 33).

PROPOSITION   1. If $X_3 = X_1 + X_2$, X_1 and X_2 are independent, $X_1 \sim \chi_{v_1}^2$, and $X_2 \sim \chi_{v_2}^2$, then $X_3 \sim \chi_{v_1 + v_2}^2$.
2. If $X_3 = X_1 + X_2$, X_1 and X_2 are independent, $X_1 \sim \chi_{v_1}^2$, and $X_3 \sim \chi_{v_3}^2$ with $v_3 > v_1$, then $X_2 \sim \chi_{v_3 - v_1}^2$.

Statement 1 says that the chi-squared distribution is an *additive* distribution; we saw in Chapter 5 that the normal and Poisson distributions are also additive. Statement 2 establishes a "subtractive" property of chi-squared, which will be critical in the next section for establishing the sampling distribution of the sample variance S^2.

The t Distribution

DEFINITION Let Z be a standard normal rv and let Y be a χ^2_v rv independent of Z. Then the **t distribution with v degrees of freedom (df)** is defined to be the distribution of the ratio

$$T = \frac{Z}{\sqrt{Y/v}}$$

We will sometimes abbreviate this distribution by $T \sim t_v$.

With some careful calculus, we can obtain the t pdf.

PROPOSITION The pdf of a random variable T having a t distribution with v degrees of freedom is

$$f(t) = \frac{1}{\sqrt{\pi v}} \frac{\Gamma((v+1)/2)}{\Gamma(v/2)} \frac{1}{(1+t^2/v)^{(v+1)/2}} \quad -\infty < t < \infty$$

Proof A t_v variable is defined in terms of a standard normal Z and a χ^2_v variable Y. They are independent, so their joint pdf $f(y, z)$ is the product of their individual pdfs. We first find the cdf of T and then differentiate to obtain the pdf:

$$F(t) = P(T \le t) = P\left(\frac{Z}{\sqrt{Y/v}} \le t\right) = P\left(Z \le t\sqrt{\frac{Y}{v}}\right) = \int_0^\infty \int_{-\infty}^{t\sqrt{y/v}} f(y,z)\,dz\,dy$$

Differentiating with respect to t using the Fundamental Theorem of Calculus,

$$f(t) = \frac{d}{dt}F(t) = \int_0^\infty \frac{\partial}{\partial t} \int_{-\infty}^{t\sqrt{y/v}} f(y,z)\,dz\,dy = \int_0^\infty f\left(y, t\sqrt{\frac{y}{v}}\right) \cdot \sqrt{\frac{y}{v}}\,dy$$

Now substitute the joint pdf—that is, the product of the marginal pdfs of Y and Z—and integrate:

$$f(t) = \int_0^\infty \frac{y^{v/2-1}}{2^{v/2}\Gamma(v/2)} e^{-y/2} \cdot \frac{1}{\sqrt{2\pi}} e^{-[t\sqrt{y/v}]^2/2} \cdot \sqrt{\frac{y}{v}}\,dy$$

$$= \frac{1}{2^{v/2}\Gamma(v/2)\sqrt{2\pi v}} \int_0^\infty y^{(v+1)/2-1} e^{-1[1/2+t^2/(2v)]y}\,dy$$

The integral can be evaluated using Expression (4.5) from the section on the gamma distribution:

$$f(t) = \frac{1}{2^{v/2}\Gamma(v/2)\sqrt{2\pi v}} \cdot \frac{\Gamma((v+1)/2)}{[1/2 + t^2/(2v)]^{v/2+1/2}}$$

$$= \frac{\Gamma((v+1)/2)}{\sqrt{\pi v}\Gamma(v/2)} \frac{1}{(1+t^2/v)^{(v+1)/2}}, \quad -\infty < t < \infty \qquad \blacksquare$$

The pdf has a maximum at 0 and decreases symmetrically as $|t|$ increases. As v becomes large, the t pdf approaches the standard normal pdf, as shown in Exercise 36. It makes sense that the t distribution would be close to the standard normal for large v, because $T = Z/\sqrt{\chi_v^2/v}$, and χ_v^2/v converges to 1 by the Law of Large Numbers, as shown in Exercise 30.

Figure 6.16 shows t density curves for $v = 1, 5,$ and 20 along with the standard normal (z) curve. Notice how fat the tails are for 1 df, as compared to the standard normal. However, as the number of df increases, the t pdf becomes more like the standard normal. For 20 df there is not much difference.
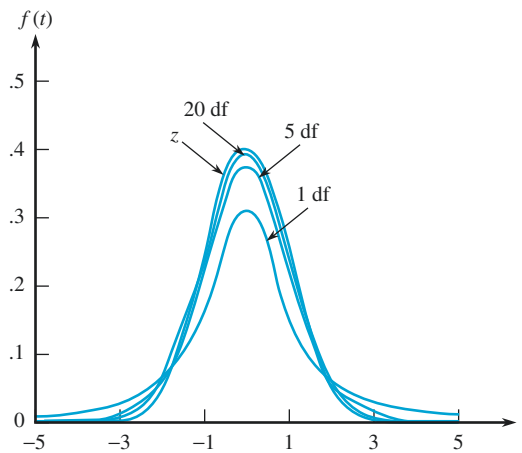
Figure 6.16 Comparison of t curves to the z curve

Integration of the t pdf is difficult without software, so values of upper-tail areas are given in Table A.7. For example, the value in the column labeled 2 and the row labeled 3.0 is .048, meaning that $P(T > 3.0) = .048$ when $T \sim t_2$. We write this as $t_{.048,2} = 3.0$. In general we write $t_{\alpha,v} = c$ if $P(T > c) = \alpha$ when $T \sim t_v$. A tabulation of these t critical values (i.e., $t_{\alpha,v}$) for frequently used tail areas α appears in Table A.6.

Using $\Gamma(1/2) = \sqrt{\pi}$, we obtain the pdf for the t distribution with 1 df as $f(t) = 1/[\pi(1+t^2)]$, which is also known as the *Cauchy distribution*. This distribution has such heavy tails that the mean does not exist (Exercise 37).

The mean and variance of a t variable can be obtained directly from the pdf, but it's instructive to derive them through the definition in terms of independent standard normal and chi-squared variables, $T = Z/\sqrt{Y/v}$. Recall from Section 5.2 that $E(UV) = E(U)E(V)$ if U and V are independent and the expectations of U and V both exist. Thus,

$$E(T) = E(Z)E(1/\sqrt{Y/v}) = E(Z)v^{1/2}E(Y^{-1/2})$$

Of course, $E(Z) = 0$, so $E(T) = 0$ if $E(Y^{-1/2})$ exists. Let's compute $E(Y^k)$ for *any* k if Y is chi-squared, using Expression (4.5):

$$E(Y^k) = \int_0^\infty y^k \frac{y^{(v/2)-1}}{2^{v/2}\Gamma(v/2)} e^{-y/2} dy = \frac{1}{2^{v/2}\Gamma(v/2)} \int_0^\infty y^{(k+v/2)-1} e^{-y/2} dy$$

$$= \frac{1}{2^{v/2}\Gamma(v/2)} \cdot 2^{k+v/2}\Gamma(k+v/2) = \frac{2^k \Gamma(k+v/2)}{\Gamma(v/2)} \quad \text{for} \quad k+v/2 > 0 \qquad (6.7)$$

If $k + v/2 \le 0$, the integral does not converge and $E(Y^k)$ does not exist. When $k = -\frac{1}{2}$, we require that $v > 1$ for the integral to converge. Thus, the mean of a t variable fails to exist if $v = 1$ and the mean is indeed 0 otherwise.

For the variance of T we need $E(T^2) = E(Z^2) \cdot E[1/(Y/v)] = 1 \cdot vE(Y^{-1})$. Using $k = -1$ in Expression (6.7), we obtain, with the help of the property $\Gamma(a + 1) = a\Gamma(a)$,

$$E(Y^{-1}) = \frac{2^{-1}\Gamma(-1+v/2)}{\Gamma(v/2)} = \frac{2^{-1}}{v/2-1} = \frac{1}{v-2} \quad \Rightarrow \quad V(T) = v \cdot \frac{1}{v-2} = \frac{v}{v-2}$$

provided that $-1 + v/2 > 0$, or $v > 2$. For 1 or 2 df the variance of T does not exist. For $v > 2$, the variance always exceeds 1, and for large df the variance is close to 1. This is appropriate because any t curve spreads out more than the z curve, but for large df the t curve approaches the z curve.

The *F* Distribution

DEFINITION   Let Y_1 and Y_2 be independent chi-squared random variables with v_1 and v_2 degrees of freedom, respectively. The **F distribution with v_1 numerator degrees of freedom and v_2 denominator degrees of freedom** is defined to be the distribution of the ratio

$$F = \frac{Y_1/v_1}{Y_2/v_2}, \qquad (6.8)$$

This distribution will sometimes be denoted F_{v_1,v_2}.

The pdf of a random variable having an F distribution is

$$f(x; v_1, v_2) = \frac{\Gamma[(v_1+v_2)/2]}{\Gamma(v_1/2)\Gamma(v_2/2)} \left(\frac{v_1}{v_2}\right)^{v_1/2} \cdot \frac{x^{v_1/2-1}}{(1+v_1 x/v_2)^{(v_1+v_2)/2}} \quad x > 0$$

Its derivation (Exercise 40) is similar to the derivation of the t pdf. Figure 6.17 shows the F density curves for several choices of v_1 and $v_2 = 10$.
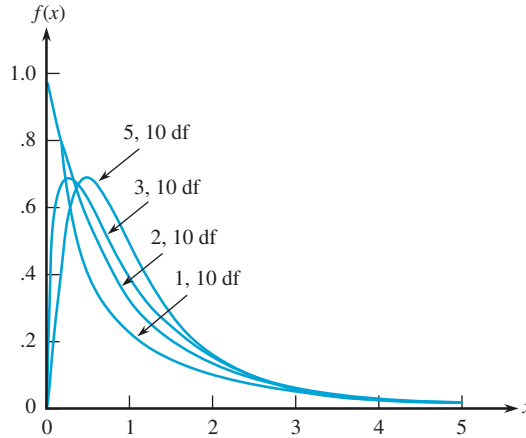
Figure 6.17 *F* density curves

The mean of the *F* distribution can be obtained with the help of Equation (6.8): $E(F) = v_2/(v_2 - 2)$ if $v_2 > 2$, and it does not exist if $v_2 \leq 2$ (Exercise 41).

What happens to *F* if the degrees of freedom are large? If v_2 is large, then the denominator of Expression (6.8) will be close to 1 (see Exercise 30), and approximately the *F* will be just the numerator chi-squared over its degrees of freedom. Similarly, if both v_1 and v_2 are large, then both the numerator and denominator will be close to 1, and the *F* ratio therefore will be close to 1.

Except for a few special choices of degrees of freedom, integration of the *F* pdf is difficult without software, so *F* critical values (values that capture specified *F* distribution tail areas) are given in Table A.8. For example, the value in the column labeled 1 and the row labeled 2 and .100 is 8.53, meaning that $P(F_{1,2} > 8.53) = .100$. We can express this as $F_{.1,1,2} = 8.53$, where $F_{\alpha,v_1,v_2} = c$ means that $P(F_{v_1,v_2} > c) = \alpha$.

That same table can also be used to determine some lower-tail areas. Since $1/F = (X_2/v_2)/(X_1/v_1)$, the reciprocal of an *F* variable also has an *F* distribution, but with the degrees of freedom reversed, and this can be used to obtain lower-tail critical values. For example, $.100 = P(F_{1,2} > 8.53) = P(1/F_{1,2} < 1/8.53) = P(F_{2,1} < .117)$. This can be written as $F_{.9,2,1} = .117$ because $.9 = P(F_{2,1} > .117)$. In general,

$$F_{p,v_1,v_2} = \frac{1}{F_{1-p,v_2,v_1}}$$

Finally, recalling the definition $T = Z/\sqrt{X/v}$ of a t_v rv, it follows that

$$T^2 = \frac{Z^2}{X/v} \sim \frac{\chi_1^2/1}{\chi_v^2/v} = F_{1,v}$$

That is, $t_v^2 = F_{1,v}$. In theory, we can use this to obtain tail areas. For example,

$$.100 = P(F_{1,2} > 8.53) = P(T_2^2 > 8.53) = P(|T_2| > \sqrt{8.53}) = 2P(T_2 > 2.92),$$

and therefore $.05 = P(T_2 > 2.92)$. We previously determined that $.048 = P(T_2 > 3.0)$, which is very nearly the same statement. In terms of our notation, $t_{.05,2} = \sqrt{F_{.10,1,2}}$, and we can similarly show that in general $t_{\alpha,v} = \sqrt{F_{2\alpha,1,v}}$ if $0 < \alpha < .5$.

Chi-squared, t, and F Calculations with Software

Although we have made several references in this section to the statistical tables in the appendix, software alleviates the need to rely on such tables. The commands for the cdf and quantile functions of the χ^2, t, and F distributions in R are presented in Table 6.3.

Table 6.3 R code for chi-squared, t, and F calculations

	Chi-squared	t	F
cdf	pchisq(x, v)	pt(x, v)	pf(x, v_1, v_2)
Quantile	qchisq(p, v)	qt(p, v)	qf(p, v_1, v_2)

Critical values can be computed by substituting $p = 1 - \alpha$ into the quantile functions. For instance, the $\alpha = .1$ critical value of the $F_{1,2}$ distribution, $F_{.1,1,2} = 8.53$, can be obtained with qf(.9, 1, 2) in R.

Exercises: Section 6.3 (28–50)

28. a. Use Table A.5 to find $\chi^2_{.05,2}$.
    b. Verify the answer to (a) by integrating the pdf.
    c. Verify the answer to (a) by using software.

29. Why should χ^2_v be approximately normal for large v? What theorem applies here, and why?

30. Apply the Law of Large Numbers to show that χ^2_v/v approaches 1 as v becomes large.

31. Show that the χ^2_v density function has a maximum at $v - 2$ if $v > 2$.

32. Show that if X_1 and X_2 are independent, $X_1 \sim \chi^2_{v_1}$, and $X_2 \sim \chi^2_{v_2}$, then $X_1 + X_2 \sim \chi^2_{v_1 + v_2}$. [*Hint*: Use mgfs.]

33. a. Show that if X_1 and X_2 are independent, $X_1 \sim \chi^2_{v_1}$, and $X_1 + X_2 \sim \chi^2_{v_3}$ with $v_3 > v_1$, then $X_2 \sim \chi^2_{v_3 - v_1}$. [*Hint*: Use mgfs.]
    b. In the setting of part (a), can we allow $v_3 < v_1$? The answer is no: show that if X_1 and X_2 are independent, $X_1 \sim \chi^2_{v_1}$, and $X_1 + X_2 \sim \chi^2_{v_3}$, then $v_3 \geq v_1$. [*Hint*: Calculate the variance of $X_1 + X_2$.]

34. a. Use Table A.6 to find $t_{.102,1}$.
    b. Verify the answer to part (a) by integrating the pdf.
    c. Verify the answer to part (a) using software.

35. a. Use Table A.6 to find $t_{.005,10}$.
    b. Use Table A.8 to find $F_{.01,1,10}$ and relate this to the value you obtained in part (a).
    c. Verify the answer to part (b) using software.

36. Show that the t pdf approaches the standard normal pdf for large df values. [*Hint*: $\Gamma(x + 1/2)/[\sqrt{x}\Gamma(x)] \rightarrow 1$ and $(1 + a/x)^x \rightarrow e^a$ as $x \rightarrow \infty$.]

37. Show directly from the pdf that the mean of a t_1 (Cauchy) random variable does not exist.

38. Show that the ratio of two independent standard normal random variables has the t_1 distribution. [*Hint*: Split the domain of the denominator into positive and negative parts.]

39. a. Use Table A.8 to find $F_{.1,2,4}$.
    b. Verify the answer to part (a) using the pdf.
    c. Verify the answer to part (a) using software.

40. Derive the F pdf by applying the method used to derive the t pdf.

41. Let X have an F distribution with v_1 numerator df and v_2 denominator df.
    a. Determine the mean value of X.
    b. Determine the variance of X.

42. Is $E(F_{v_1,v_2}) = E(\chi^2_{v_1}/v_1)/E(\chi^2_{v_2}/v_2)$? Explain.

43. Show that $F_{p,v_1,v_2} = 1/F_{1-p,v_2,v_1}$.

44. a. Use Table A.6 to find $t_{.25,10}$.
    b. Use (a) to find the median of the $F_{1,10}$ distribution.

c. Verify the answer to part (b) using software.

45. Show that if X has a gamma distribution and $c > 0$ is a constant, then cX has a gamma distribution. In particular, if X is chi-squared distributed, then cX has a gamma distribution.

46. Suppose $T \sim t_9$. Determine the distribution of $1/T^2$.

47. Let Z_1, Z_2, X_1, X_2, X_3 be independent rvs with each $Z_i \sim N(0, 1)$ and each $X_i \sim N(0, 5)$. Construct a variable involving the Z_i's and X_i's which has an $F_{3,2}$ distribution.

48. Let $Z_1, Z_2, \ldots, Z_{10}$ be independent standard normal. Use these to construct

   a. A χ_4^2 random variable.
   b. A t_4 random variable.
   c. An $F_{4,6}$ random variable.
   d. A Cauchy random variable.
   e. An exponential random variable with mean 2.

f. An exponential random variable with mean 1.

g. A gamma random variable with mean 1 and variance $\frac{1}{2}$. [*Hint*: Use part (a) and Exercise 45.]

49. a. Use Exercise 29 to approximate $P(\chi_{50}^2 > 70)$, and compare the result with the answer given by software, .03237.

   b. Use the formula from Table A.5, $\chi_{\alpha,v}^2 \approx v\big(1 - 2/(9v) + z_\alpha\sqrt{2/(9v)}\big)^3$, to approximate $P(\chi_{50}^2 > 70)$, and compare with part (a).

50. The difference of two independent normal variables itself has a normal distribution. Is it true that the difference between two independent chi-squared variables has a chi-squared distribution? Explain.

6.4 Distributions Based on Normal Random Samples

Let $X_1, \ldots, X_n$ be a random sample from a normally distributed population. We saw previously that the sampling distribution of the sample mean, $\overline{X}$, is then also normal. In this section, we develop the sampling distribution of the sample *variance* S^2, the joint distribution of $\overline{X}$ and S^2, and the distributions of other important statistics when sampling from a normal distribution. The χ^2, t, and F distributions of Section 6.3 will feature centrally in this section, and the results established here will serve as the backbone for many of the statistical inference procedures in the second half of this book.

The Joint Sampling Distribution of $\overline{X}$ and S^2

For a random sample $X_1, \ldots, X_n$, the sample variance S^2 is defined as a rv by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

This can be used to calculate an estimate of σ^2 when the population mean μ is unknown. This is the same formula presented in Section 1.4, but now we acknowledge that S^2, like any statistic, will vary in value from sample to sample. To establish the sampling distribution of S^2 when sampling from a normal population, we first need the following critical result.

THEOREM   If $X_1, X_2, \ldots, X_n$ form a random sample from a normal distribution, then $\overline{X}$ and S^2 are independent.

Proof Consider the covariances between the sample mean and the deviations from the sample mean. Using the linearity of the covariance operator,

$$\text{Cov}(X_i - \overline{X}, \overline{X}) = \text{Cov}(X_i, \overline{X}) - \text{Cov}(\overline{X}, \overline{X})$$

$$= \text{Cov}\left(X_i, \frac{1}{n}\sum_{k=1}^{n} X_k\right) - V(\overline{X})$$

$$= \frac{1}{n}\text{Cov}(X_i, X_i) + \frac{1}{n}\sum_{k \neq i}\text{Cov}(X_i, X_k) - V(\overline{X})$$

$$= \frac{1}{n}V(X_i) + 0 - V(\overline{X}) = \frac{1}{n}\sigma^2 - \frac{\sigma^2}{n} = 0$$

The middle term in the third line is zero because independence of the X_i's implies that $\text{Cov}(X_i, X_k) = 0$ for $k \neq i$. This shows that $\overline{X}$ is uncorrelated with all the deviations of the observations from their mean. In general, this would not imply independence, but *in the special case of the bivariate normal distribution, being uncorrelated is equivalent to independence.* Both $\overline{X}$ and $X_i - \overline{X}$ are linear combinations of the independent normal observations, so their joint distribution is bivariate normal, as discussed in Section 5.5. Because the sample variance S^2 is composed of the deviations $X_i - \overline{X}$, we conclude that $\overline{X}$ and S^2 are independent.   ∎

To better understand the foregoing independence property, consider selecting sample after sample of size n from a particular population distribution, calculating $\bar{x}$ and s for each sample, and then plotting the resulting $(\bar{x}, s)$ pairs. Figure 6.18a shows the result for 1000 samples of size $n = 5$ from a standard normal population distribution. The elliptical pattern, with axes parallel to the coordinate axes, suggests no relationship between $\bar{x}$ and s, that is, independence of the statistics $\overline{X}$ and S (equivalently, $\overline{X}$ and S^2). However, this independence fails for data from a nonnormal distribution. Figure 6.18b illustrates what happens for samples of size 5 from an exponential distribution with mean 1. This plot shows a strong relationship between the two statistics, which is what might be expected for data from a highly skewed distribution.
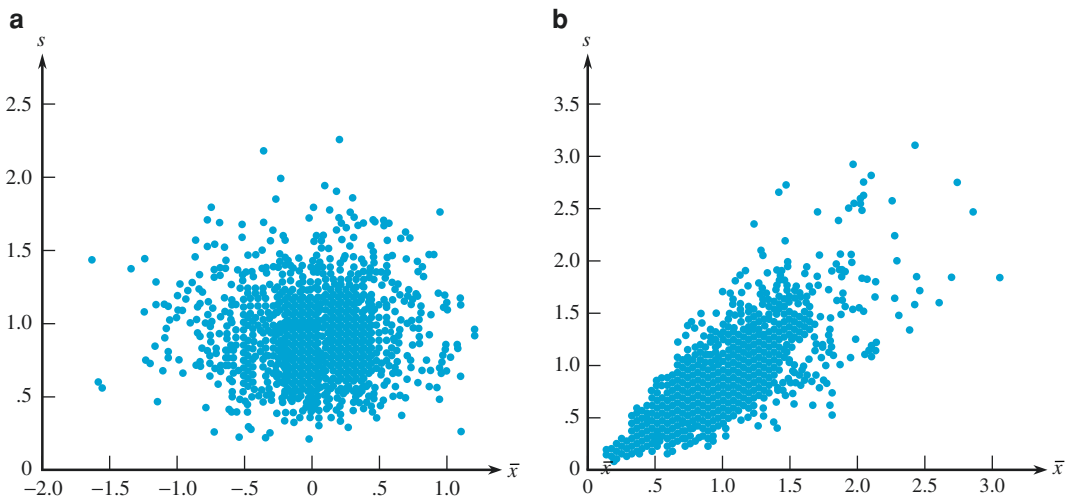
Figure 6.18 Plot of $(\bar{x}, s)$ pairs for (**a**) samples from a normal distribution; (**b**) samples from a nonnormal distribution

We are now ready to derive the sampling distribution of S^2 when sampling from a normally distributed population. Notice that we'll then know the *joint* distribution of $\overline{X}$ and S^2, since it was established in Section 6.2 that $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$ and we just proved that these two statistics are independent.

PROPOSITION | If X_1, X_2, ..., X_n form a random sample from a $N(\mu,\ \sigma)$ distribution, then $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Proof To begin, write

$$\sum (X_i - \mu)^2 = \sum [(X_i - \overline{X}) + (\overline{X} - \mu)]^2$$
$$= \sum (X_i - \overline{X})^2 + 2(\overline{X} - \mu) \sum (X_i - \overline{X}) + \sum (\overline{X} - \mu)^2$$

The middle term on the second line vanishes (do you see why?). Dividing through by σ^2, we obtain

$$\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 + \sum \left(\frac{\overline{X} - \mu}{\sigma}\right)^2 = \sum \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 + n\left(\frac{\overline{X} - \mu}{\sigma}\right)^2$$

This can be re-written as

$$\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum \frac{(X_i - \overline{X})^2}{\sigma^2} + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$
$$\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \tag{6.9}$$

If $X_i \sim N(\mu,\ \sigma)$, then $(X_i - \mu)/\sigma$ is a standard normal rv. So, the left-hand side of (6.9) is the sum of squares of n iid standard normal rvs, which by definition has a χ^2_n distribution. At the same time, the rightmost term in (6.9) is the square of the standardized version of $\overline{X}$. So, it's distributed as Z^2 with $Z \sim N(0, 1)$, which by definition is χ^2_1. And, critically, the two terms on the right-hand side of (6.9) are independent, because S^2 and $\overline{X}$ are independent. Therefore, from the "subtractive" property of the chi-squared distribution in Section 6.3 (with $v_3 = n$ and $v_1 = 1$), we conclude that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, as claimed. ∎

Intuitively, the degrees of freedom make sense because s^2 is built from the deviations $(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_n - \bar{x})$, which sum to zero:

$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$$

The last deviation is determined by the first $(n-1)$ deviations, so it is reasonable that s^2 has only $(n-1)$ degrees of freedom. The degrees of freedom help to explain why the definition of s^2 has $(n-1)$ and not n in the denominator.

Knowing that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, it can be shown (see Exercise 52) that the expected value of S^2 is σ^2, and also that the variance of S^2 approaches 0 as n becomes large.

A *t*-Distributed Statistic

In Section 6.3, we defined the *t* distribution as a particular ratio involving a normal rv and a chi-squared rv. From the definition it is not obvious how the *t* distribution can be applied to data, but the next result puts the distribution in more directly usable form. This result was originally discovered in 1908 by William Sealy Gosset, a statistician at the Guinness Brewery in Dublin, Ireland.

GOSSET'S THEOREM   If $X_1, X_2, ..., X_n$ is a random sample from a $N(\mu, \sigma)$ distribution, then the rv

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has the *t* distribution with $(n - 1)$ degrees of freedom, t_{n-1}.

Proof   Re-express the fraction in a slightly messier way:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{\frac{(n - 1)S^2}{\sigma^2}/(n - 1)}}$$

The numerator on the right-hand side is standard normal. The denominator is the square root of a χ^2_{n-1} variable, divided by its degrees of freedom. This chi-squared variable is independent of the numerator, so by definition the ratio has the *t* distribution with $n - 1$ degrees of freedom.   ∎

It's worth comparing the two rvs

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{6.10}$$

When $X_1, \ldots, X_n$ are iid normal rvs, then Z has a standard normal distribution. By contrast, the rv T—obtained by replacing σ with S in the expression for Z in (6.10)—has a t_{n-1} distribution. Replacing the constant σ with the rv S results in T having greater variability than Z, which is consistent with the comparison between the *t* distributions and the standard normal distribution described in Section 6.3 (look back at Figure 6.16).

An *F*-Distributed Statistic

Suppose that we have a random sample of *m* observations from the normal population $N(\mu_1, \sigma_1)$ and an independent random sample of *n* observations from a second normal population $N(\mu_2, \sigma_2)$. Then for the sample variance S_1^2 from the first group we know $(m - 1)S_1^2/\sigma_1^2$ is χ^2_{m-1}, and similarly for the second group $(n - 1)S_2^2/\sigma_2^2$ is χ^2_{n-1}. Thus, according to the definition of the *F* distribution given in (6.8),

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\dfrac{(m - 1)S_1^2/\sigma_1^2}{m - 1}}{\dfrac{(n - 1)S_2^2/\sigma_2^2}{n - 1}} \sim F_{m-1, n-1} \tag{6.11}$$

The F distribution, via Expression (6.11), will be used in Chapter 10 to compare the variances from two independent groups. Also, for several independent groups, in Chapter 11 we will use the F distribution to see if the differences among sample means are bigger than would be expected by chance.

Exercises: Section 6.4 (51–58)

51. Show that when sampling from a normal distribution, the sample variance S^2 has a gamma distribution, and identify the parameters α and β. [*Hint:* See Exercise 45.]

52. Knowing that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ for a normal random sample,
    a. Show that $E(S^2) = \sigma^2$.
    b. Show that $V(S^2) = 2\sigma^4/(n-1)$. What happens to this variance as n gets large?
    c. Apply Expression (6.7) to show that

    $$E(S) = \sigma \frac{\sqrt{2}\Gamma(n/2)}{\sqrt{n-1}\Gamma[(n-1)/2]}.$$

    Then show that $E(S) = \sigma\sqrt{2/\pi}$ if $n = 2$. Is $E(S) = \sigma$ for normal data?

53. Suppose $X_1, \ldots, X_{13}$ form a random sample from a normal distribution with mean 5 and standard deviation 8.
    a. Calculate $P(\overline{X} < 9.13)$.
    b. Calculate $P\left(\sum (X_i - \overline{X})^2 < 1187\right)$.
       [*Hint:* How does this relate to S^2?]
    c. Calculate
       $P\left(\overline{X} < 9.13 \cap \sum (X_i - \overline{X})^2 < 1187\right)$.
    d. In this context, construct a rv that has a t distribution, and identify its df.

54. In the unusual situation that the population mean μ is known but the population variance σ^2 is not, we might consider the following alternative statistic for estimating σ^2:

    $$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2$$

    a. Show that $E(\hat{\sigma}^2) = \sigma^2$ regardless of whether the X_i's are normally distributed (but still assuming they comprise a random sample from *some* population distribution).

    b. Now assume the X_i's are normally distributed. Determine a scaling constant c so that the rv $c \cdot \hat{\sigma}^2$ has a chi-squared distribution, and identify the number of degrees of freedom.
    c. Determine the variance of $\hat{\sigma}^2$ assuming the X_i's are normally distributed.

55. Suppose $X_1, \ldots, X_{27}$ are iid $N(5, 4)$ rvs. Let $\overline{X}$ and S denote their sample mean and sample standard deviation, respectively. Calculate $P(|\overline{X} - 5| > 0.4S)$.

56. It was established in this section that $\overline{X}$ and S^2 are independent rvs when sampling from a normal population. Is the same true for $\overline{X}$ and $\hat{\sigma}^2 = (1/n)\sum (X_i - \mu)^2$, the estimator from Exercise 54? Let's find out.
    a. Let $X \sim N(\mu, \sigma)$.
       Determine $\text{Cov}(X - \mu, (X - \mu)^2)$ and $\text{Cov}(X, (X - \mu)^2)$. [*Hint:* Use the covariance shortcut formula.]
    b. Use part (a) to show that $\overline{X}$ and $\hat{\sigma}^2$ are uncorrelated. Does it follow that $\overline{X}$ and $\hat{\sigma}^2$ are independent?
    c. The proof of the independence of $\overline{X}$ and S^2 relied critically on the fact that $\text{Cov}(\overline{X}, X_i - \overline{X}) = 0$. Calculate $\text{Cov}(\overline{X}, X_i - \mu)$. Based on this result, does it appear that $\overline{X}$ and $\hat{\sigma}^2$ are independent?

57. Suppose we have a sample of size n from a $N(\mu, \sigma)$ distribution. Define rvs Z and T as in Expression (6.10).
    a. Calculate $P(-2 \le Z \le 2)$ for $n = 5, 10$, and 15. How does sample size affect your answer?
    b. Calculate $P(-2 \le T \le 2)$ for $n = 5, 10$, and 15. How does sample size affect your answer?

58. Suppose that we have a random sample of size m from a $N(\mu_1, \sigma_1)$ distribution and an

independent random sample of size n from a $N(\mu_2, \sigma_2)$ distribution. To assess whether the two *populations* have equal variances— a requirement of several procedures later in this book—we consider the ratio of the two *sample* variances:

$$R = \frac{S_1^2}{S_2^2}$$

a. If the two populations indeed have equal variances—that is, if $\sigma_1^2 = \sigma_2^2$— then what is the distribution of R?

b. A common convention for accepting that the population variances might be equal is that the larger sample standard deviation should be no more than twice the smaller. Express that condition in terms of R.

c. For the specific case $m = 10$ and $n = 15$, calculate the probability of the condition in part (b), assuming that the two population variances are indeed equal.

d. If the population variances really are equal, but the sample sizes are now $m = 50$ and $n = 60$, will the probability in part (c) be higher or lower? Why?

Supplementary Exercises: (59–68)

59. A small high school holds its graduation ceremony in the gym. Because of seating constraints, students are limited to a maximum of four tickets to graduation for family and friends. The vice principal knows that historically 30% of students want four tickets, 25% want three, 25% want two, 15% want one, and 5% want none.

a. Let X = the number of tickets requested by a randomly selected graduating student, and assume the historical distribution applies to this rv. Find the mean and standard deviation of X.

b. Let T_o = the total number of tickets requested by the 150 students graduating

this year. Assuming all 150 students' requests are independent, determine the mean and standard deviation of T_o.

c. The gym can seat a maximum of 500 guests. Calculate the (approximate) probability that all students' requests can be accommodated. [*Hint:* Express this probability in terms of T_o. What distribution does T_o have?]

60. Suppose that for a certain individual, calorie intake at breakfast is a random variable with expected value 500 and standard deviation 50, calorie intake at lunch is random with expected value 900 and standard deviation 100, and calorie intake at dinner is a random variable with expected value 2000 and standard deviation 180. Assuming that intakes at different meals are independent of each other, what is the probability that average calorie intake per day over the next (365-day) year is at most 3500? [*Hint*: Let X_i, Y_i, and Z_i denote the three calorie intakes on day i. Then total intake is given by $\sum (X_i + Y_i + Z_i)$.]

61. Suppose the proportion of rural voters in a certain state who favor a particular gubernatorial candidate is .45 and the proportion of suburban and urban voters favoring the candidate is .60. If a sample of 200 rural voters and 300 urban and suburban voters is obtained, what is the approximate probability that at least 250 of these voters favor this candidate?

62. Let μ denote the true pH of a chemical compound. A sequence of n independent sample pH determinations will be made. Suppose each sample pH is a random variable with expected value μ and standard deviation. 1. How many determinations are required if we wish the probability that the sample average is within .02 of the true pH to be at least .95? What theorem justifies your probability calculation?

63. A large university has 500 single employees who are covered by its dental plan. Suppose the number of claims filed during

the next year by such an employee is a Poisson rv with mean value 2.3. Assuming that the number of claims filed by any such employee is independent of the number filed by any other employee, what is the approximate probability that the total number of claims filed is at least 1200?

64. Consider independent and identically distributed random variables X_1, X_2, X_3, … where each X_i has a discrete uniform distribution on the integers 0, 1, 2, …, 9; that is, $P(X_i = k) = 1/10$ for $k = 0, 1, 2, …, 9$. Now form the sum

$$U_n = \sum_{i=1}^{n} \frac{1}{(10)^i} X_i$$
$$= .1X_1 + .01X_2 + \cdots + (.1)^n X_n$$

Intuitively, this is just the first n digits in the decimal expansion of a random number on the interval [0, 1]. Show that as $n \to \infty$, U_n converges in distribution to an rv U uniformly distributed on [0, 1], i.e. that $P(U_n \le u) \to P(U \le u)$, by showing that the moment generating function of U_n converges to the moment generating function of U.
[The argument for this appears on p. 52 of the article "A Few Counter Examples Useful in Teaching Central Limit Theorems," *The American Statistician*, Feb. 2013.]

65. The Empirical Rule from Chapter 4 states that roughly 68% of a standard normal distribution is within ±1 of its center, 95% within ±2, and 99.7% within ±3.

   a. For the t_2 distribution, determine what percent of the total area is within ±1, ±2, and ±3 of its center.
   b. For the t_2 distribution, determine how far you must go out to capture 68, 95, and 99.7% of the total area under the pdf.

66. a. Show that if $X \sim F_{v_1, v_2}$, then the distribution of $v_1 \cdot X$ approaches $\chi^2_{v_1}$ as $v_2 \to \infty$. [*Hint*: Apply Exercise 30.] What is the limiting distribution of X itself as $v_1 \to \infty$?

   b. Show that if $X \sim F_{v_1, v_2}$, then the distribution of v_2/X approaches $\chi^2_{v_2}$ as $v_1 \to \infty$. What is the limiting distribution of X itself as $v_1 \to \infty$?

67. Suppose that we have a random sample of 10 observations from a $N(\mu_1, \sigma_1)$ distribution and an independent random sample of 12 observations from a $N(\mu_2, \sigma_2)$ distribution. Let S_1^2 and S_2^2 denote the sample variances of these two random samples.

   a. Determine

$$P\left(2.90 \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} \le 8.12 \frac{\sigma_1^2}{\sigma_2^2}\right)$$

   b. Define a rv $\hat{\sigma}_1^2 = \frac{1}{10} \sum_{i=1}^{10} (X_i - \mu_1)^2$ from the first random sample and define $\hat{\sigma}_2^2$ similarly for the second random sample. Determine

$$P\left(2.19 \frac{\sigma_1^2}{\sigma_2^2} \le \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \le 4.30 \frac{\sigma_1^2}{\sigma_2^2}\right)$$

68. Let X_1, X_2, … be a sequence of independent, but not necessarily identically distributed random variables, and let $T_o = X_1 + \cdots + X_n$. *Lyapunov's Theorem* states that the standardized rv $(T_o - \mu_{T_o})/\sigma_{T_o}$ converges to a $N(0, 1)$ distribution as $n \to \infty$, provided that

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} E(|X_i - \mu_i|^3)}{\sigma_{T_o}^3} = 0$$

where $\mu_i = E(X_i)$. This limit is sometimes referred to as the *Lyapunov condition* for convergence.

   a. Assuming $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$, write expressions for μ_{T_o} and σ_{T_o}.
   b. Show that the Lyapunov condition is automatically met when the X_i's are iid. [*Hint*: Let $\tau = E(|X_i - \mu_i|^3)$, which we assume is finite, and observe that τ is the same for every X_i. Then simplify the limit.]

c. Let X_1, X_2, … be independent random variables, with X_i having an exponential distribution with mean i. Show that $X_1 + \cdots + X_n$ has an approximately normal distribution as n increases.

d. An online trivia game presents progressively harder questions to players; specifically, the probability of answering the ith question correctly is $1/i$. Assume any player's successive answers are independent, and let T_o denote the number of questions a player has right out of the first n. Show that T_o has an approximately normal distribution for large n.

Appendix: Proof of the Central Limit Theorem

First, here is a restatement of the theorem. Let X_1, X_2, …, X_n be a random sample from a distribution with mean μ and standard deviation σ. Then, if Z is a standard normal random variable,

$$\lim_{n \to \infty} P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = P(Z \leq z) = \Phi(z)$$

The theorem says that the distribution of the standardized $\overline{X}$ approaches the standard normal distribution. Our proof is for the special case in which the moment generating function exists, which implies also that all its derivatives exist and that they are continuous. We will show that the mgf of the standardized $\overline{X}$ approaches the mgf of the standard normal distribution. Convergence of the mgf implies convergence of the distribution, though we will not prove that here (the mathematics is beyond the scope of this book).

To simplify the proof slightly, define new rvs by $W_i = (X_i - \mu)/\sigma$ for $i = 1, 2,…, n$, the standardized versions of the X_i. Then $X_i = \mu + \sigma W_i$, from which $\overline{X} = \mu + \sigma \overline{W}$ and we may write the standardized $\overline{X}$ expression as

$$Y = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{(\mu + \sigma\overline{W}) - \mu}{\sigma/\sqrt{n}} = \sqrt{n} \cdot \overline{W} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i$$

Let $M_W(t)$ denote the common mgf of the W_i's (since the X_i's are iid, so are the W_i's). We will obtain the mgf of Y in terms of $M_W(t)$; we then want to show that the mgf of Y converges to the mgf of a standard normal random variable, $M_Z(t) = e^{t^2/2}$.

From the mgf properties in Section 5.3, we have the following:

$$M_Y(t) = M_{W_1 + \cdots + W_n}(t/\sqrt{n}) = [M_W(t/\sqrt{n})]^n$$

For the limit, we will use the fact that $M_W(0) = 1$, a basic property of all mgfs. And, critically, because the W_i's are standardized rvs, $E(W_i) = 0$ and $V(W_i) = 1$, from which we also have $M_W'(0) = E(W) = 0$ and $M_W''(0) = E(W^2) = V(W) + [E(W)]^2 = 1$.

To determine the limit as $n \to \infty$, we take a natural logarithm, make the substitution $x = 1/\sqrt{n}$, then apply L'Hôpital's Rule twice:

$$\lim_{n\to\infty} \ln[M_Y(t)] = \lim_{n\to\infty} n \ln[M_W(t/\sqrt{n})]$$

$$= \lim_{x\to 0} \frac{\ln[M_W(tx)]}{x^2} \quad \text{substitute } x = 1/\sqrt{n}$$

$$= \lim_{x\to 0} \frac{M'_W(tx) \cdot t/M_W(tx)}{2x} \quad \text{L'Hôpital's Rule}$$

$$= \lim_{x\to 0} \frac{tM'_W(tx)}{2xM_W(tx)}$$

$$= \lim_{x\to 0} \frac{t^2 M''_W(tx)}{2M_W(tx) + 2xtM'_W(tx)} \quad \text{L'Hôpital's Rule}$$

$$= \frac{t^2 M''_W(0)}{2M_W(0) + 2(0)tM'_W(0)} = \frac{t^2(1)}{2(1) + 0} = \frac{t^2}{2}$$

You can verify for yourself that at each use of L'Hôpital's Rule, the preceding fraction had the indeterminate 0/0 form. Finally, since the logarithm function and its inverse are continuous, we may conclude that $M_Y(t) \to e^{t^2/2}$, which completes the proof.   ∎

Point Estimation

Introduction

Given a parameter of interest, such as a population mean μ or population proportion p, the objective of point estimation is to use a sample to compute a number that represents, in some sense, a "good guess" for the true value of the parameter. The resulting number is called a *point estimate*. In Section 7.1, we present some general concepts of point estimation. In Section 7.2, we describe and illustrate two important methods for obtaining point estimates: the method of moments and the method of maximum likelihood.

Obtaining a point estimate entails calculating the value of a statistic such as the sample mean $\overline{X}$ or sample proportion $\hat{P}$. We should therefore be concerned that the chosen statistic utilizes all the relevant information available about the parameter of interest. The idea of "no information loss" is made precise by the concept of *sufficiency*, which is developed in Section 7.3. Finally, Section 7.4 further explores the meaning of *efficient* estimation and properties of maximum likelihood estimators.

7.1 Concepts and Criteria for Point Estimation

Statistical inference is frequently directed toward drawing some type of conclusion about one or more parameters (population characteristics). To do so requires that an investigator obtain sample data from each of the populations under study. Conclusions can then be based on the computed values of various sample quantities. For example, let μ (a parameter) denote the average salary of all alumni from a certain university. A random sample of $n = 250$ alumni might be chosen and the salary for each one determined, resulting in observed values $x_1, x_2, \ldots, x_{250}$. The sample mean salary $\bar{x}$ could then be used to draw a conclusion about the value of μ. Similarly, if σ is the standard deviation of the alumni salary distribution (population sd, another parameter), the value of the sample standard deviation s can be used to infer something about σ.

Recall from the previous chapter that before data is available, the sample observations are considered random variables (rvs) $X_1, X_2, \ldots, X_n$. It follows that any function of the X_i's—that is, any statistic—such as the sample mean $\overline{X}$ or sample standard deviation S is also a random variable. That is, its value will generally vary from one sample to another, and before a particular sample is selected, there is uncertainty as to what value the statistic will assume. The same is true if available data consists of more than one sample. For example, we can represent the salaries of m statistics alumni and n computer science alumni by $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, respectively. The difference between the

two sample mean salaries is $\overline{X} - \overline{Y}$, the natural statistic for making inferences about $\mu_1 - \mu_2$, the difference between the population mean salaries.

When discussing general concepts and methods of inference, it is convenient to have a generic symbol for the parameter of interest. We will use the Greek letter θ for this purpose.

DEFINITION   A **point estimate** of a parameter θ is a single number that can be regarded as a sensible value for θ. A point estimate is obtained by selecting a suitable statistic and determining its value from the given sample data. The selected statistic is called the **point estimator** of θ.

Suppose, for example, that the parameter of interest is $\mu =$ the true average battery life (in hours) for a certain type of cell phone under continuous use. A random sample of $n = 3$ phones might yield observed lifetimes $x_1 = 5.0$, $x_2 = 6.4$, $x_3 = 5.9$. The computed value of the sample mean lifetime is $\overline{x} = 5.77$, and it is reasonable to regard 5.77 h as a plausible value of μ, our "best guess" for the value of μ based on the available sample information. The point *estimator* used was the statistic $\overline{X}$, and the point *estimate* of μ was $\overline{x} = 5.77$. If the three observed lifetimes had instead been $x_1 = 5.6$, $x_2 = 4.5$, and $x_3 = 6.1$, use of the same *estimator* $\overline{X}$ would have resulted in a different point *estimate*, $\overline{x} = (5.6 + 4.5 + 6.1)/3 = 5.40$ h.

The symbol $\hat{\theta}$ ("theta hat") is customarily used to denote the point estimate resulting from a given sample; we shall also use it to denote the estimator, as an uppercase $\hat{\Theta}$ is somewhat awkward to write. Thus $\hat{\mu} = \overline{X}$ is read as "the point estimator of μ is the sample mean $\overline{X}$." The statement "the point estimate of μ is 5.77 h" can be written concisely as $\hat{\mu} = \overline{x} = 5.77$. Notice that in writing a statement like $\hat{\theta} = 72.5$, there is no indication of how this point estimate was obtained (i.e., what statistic was used). We recommend that both the estimator/statistic and the resulting estimate be reported.

Example 7.1 An automobile manufacturer has developed a new type of bumper, which is supposed to absorb impacts with less damage than previous bumpers. The manufacturer has used this bumper in a sequence of 25 controlled crashes against a wall, each at 10 mph, using one of its compact car models. Let $X =$ the number of crashes that result in no visible damage to the automobile (a "success"). The parameter to be estimated is $p =$ the proportion of *all* such crashes that result in no visible damage; equivalently, $p = P$(no visible damage in a crash). If X is observed to be $x = 15$, the most reasonable estimator and estimate are

$$\text{estimator} = \hat{P} = \frac{X}{n} \quad \text{estimate} = \hat{p} = \frac{x}{n} = \frac{15}{25} = .60 \qquad \blacksquare$$

If for each parameter of interest there were only one reasonable point estimator, there would not be much to point estimation. In most problems, though, there will be more than one reasonable estimator.

Example 7.2 Many communities have added fluoride to drinking water since the 1940s, but the solubility of sodium fluoride in particular is important to many industries. The article "A Review of Sodium Fluoride Solubility in Water" (*J. Chem. Engr. Data* 2017: 1743–1748) provides the following $n = 16$ values for the solubility of sodium fluoride (millimoles of NaF per kilogram of H_2O, mmol/kg) at 25 °C:

| 956 | 974 | 980 | 980 | 982 | 983 | 983 | 985 |
| 985 | 985 | 987 | 987 | 995 | 999 | 1000 | 1007 |

One goal in the article was to estimate μ = the true mean solubility of NaF at 25 °C. A dotplot of the sample data suggests a symmetric measurement distribution, so μ could also represent the true median solubility. The given observations are assumed to be the result of a random sample $X_1, X_2, \ldots,$ X_{16} from this symmetric distribution. Consider the following estimators and resulting estimates for μ:

a. Estimator = $\overline{X}$, estimate = $\bar{x} = \sum x_i / n = 15{,}768/16 = 985.5 \, \text{mmol/kg}$
b. Estimator = $\tilde{X}$, estimate = $\tilde{x} = (985 + 985)/2 = 985 \, \text{mmol/kg}$
c. Estimator = $\overline{X}_e = [\min(X_i) + \max(X_i)]/2$ = the *midrange* (i.e., the average of the two extreme values), estimate = $[\min(x_i) + \max(x_i)]/2 = (956 + 1007)/2 = 981.5 \, \text{mmol/kg}$
d. Estimator = $\overline{X}_{\text{tr}(6.25)}$, the 6.25% trimmed mean (discard the smallest and largest values of the sample and then average), estimate = $\bar{x}_{\text{tr}(6.25)} = (15{,}768 - 956 - 1007)/14 = 986.1 \, \text{mmol/kg}$.

Each one of the different estimators (a)–(d) uses a different measure of the center of the sample to estimate μ. Which of the estimates is closest to the true value? This question cannot be answered without already knowing the true value. However, a question that *can* be addressed is, "Which estimator, when used on *other* samples of X_i's, will tend to produce estimates closest to the true value?"   ■

Example 7.3 Continuing the previous example, suppose we also want to estimate the population variance σ^2. A natural estimator is the sample variance:

$$\hat{\sigma}^2 = S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$$

The corresponding point estimate is

$$\hat{\sigma}^2 = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum (x_i - 985.5)^2}{16-1} = \frac{(956 - 985.5)^2 + \cdots + (1007 - 985.5)^2}{15} = 135.87$$

A point estimate of σ would then be $\hat{\sigma} = s = \sqrt{135.87} = 11.66 \, \text{mmol/kg}$.

An alternative estimator would result from using divisor n instead of $n - 1$ (i.e., the average squared deviation):

$$\hat{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{n} \quad \text{estimate} = \frac{(956 - 985.5)^2 + \cdots + (1007 - 985.5)^2}{16} = 127.38$$

We will indicate shortly why many statisticians prefer S^2 to the estimator with divisor n.   ■

Assessing Estimators: Accuracy and Precision

When a particular statistic is selected to estimate an unknown parameter, two criteria often used to assess the quality of that estimator are its accuracy and its precision. Loosely speaking, an estimator is *accurate* if it has no systematic tendency to overestimate or underestimate the value of the parameter, across repeated values of the estimator calculated from different samples. An estimator is *precise* if those same repeated values are typically "close together," so that two statisticians using the same estimator (but two different random samples) are liable to get similar point estimates. The notions of accuracy and precision are made more rigorous by the following definitions.

DEFINITION   A point estimator $\hat{\theta}$ is said to be an **unbiased estimator** of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ. The difference $E(\hat{\theta}) - \theta$ is called the **bias** of $\hat{\theta}$ (and equals 0 if $\hat{\theta}$ is unbiased).

The **standard error** of $\hat{\theta}$ is its standard deviation, $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$. If the standard error itself involves unknown parameters whose values can be estimated, substitution of these estimates into $\sigma_{\hat{\theta}}$ yields the **estimated standard error** of $\hat{\theta}$. The estimated standard error can be denoted by either $\hat{\sigma}_{\hat{\theta}}$ or by $s_{\hat{\theta}}$.

Unbiasedness requires that the sampling distribution of the estimator be centered at the value of θ, whatever that value might be. Thus if $\theta = 50$, the mean value of the estimator must be 50, if $\theta = .25$ the mean value must be .25, and so on. The bias of an estimator $\hat{\theta}$ quantifies its accuracy by measuring how far, on the average, $\hat{\theta}$ differs from θ. The standard error of $\hat{\theta}$ quantifies its precision by measuring the variability of $\hat{\theta}$ across different possible realizations (i.e., different random samples). Intuitively its value describes the "typical" deviation between an estimate and the mean value of the estimator. It is important to note that both bias and standard error are properties of an *estimator* (the random variable), such as $\overline{X}$, and not of any specific value or *estimate*, $\overline{x}$.

Figure 7.1 illustrates bias and standard error for three potential estimators of a population parameter θ. Figure 7.1a shows the distribution of an estimator $\hat{\theta}_1$ whose expected value is very close to θ but whose distribution is quite dispersed. Hence, $\hat{\theta}_1$ has low bias but relatively high standard error. In contrast, the distribution of $\hat{\theta}_2$ displayed in Figure 7.1b is very concentrated but is "off target": the values of $\hat{\theta}_2$ across different random samples will systematically overestimate θ. So, $\hat{\theta}_2$ has low standard error but high bias. The "ideal" estimator is illustrated in Figure 7.1c: $\hat{\theta}_3$ has a mean roughly equal to θ, so it has low bias, and it also has a relatively small standard error.
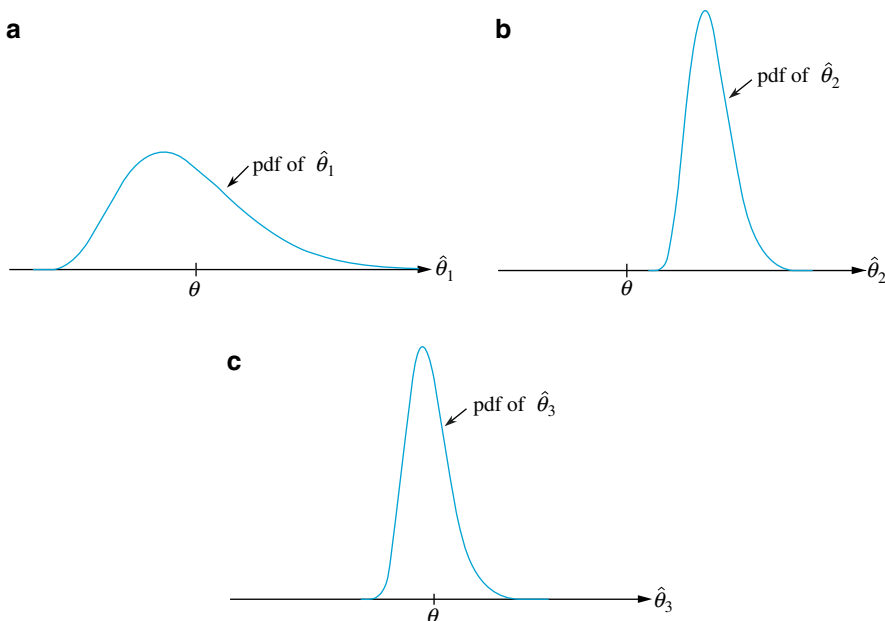
Figure 7.1   Three potential types of estimators: (**a**) accurate, but not precise; (**b**) precise, but not accurate; (**c**) both accurate and precise

It may seem as though it is necessary to know the value of θ (in which case estimation is unnecessary!) to decide whether an estimator $\hat{\theta}$ is unbiased. This is not usually the case, though, as we'll see in the next several examples.

Example 7.4 In Example 7.1, the sample proportion $\hat{P} = X/n$ was used as an estimator of $p =$ the true proportion of successes in all possible crash tests. Because X, the number of sample successes, has a $\text{Bin}(n, p)$ distribution, the mean of $\hat{P}$ is

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$$

Thus $\hat{P}$ is unbiased regardless of the value of p and the sample size n. The standard error of the estimator is

$$\sigma_{\hat{P}} = \sqrt{V(\hat{P})} = \sqrt{V\left(\frac{X}{n}\right)} = \sqrt{\frac{1}{n^2}V(X)} = \sqrt{\frac{1}{n^2}np(1-p)} = \sqrt{\frac{p(1-p)}{n}}$$

Since p is unknown (else why estimate?), we could substitute $\hat{p} = x/n$ into $\sigma_{\hat{P}}$, yielding the *estimated* standard error $\hat{\sigma}_{\hat{P}} = \sqrt{\hat{p}(1-\hat{p})/n}$. When $n = 25$ and $\hat{p} = .6$, this gives $\hat{\sigma}_{\hat{P}} = \sqrt{(.6)(.4)/25} = .098$. Alternatively, since the largest value of $p(1-p)$ is attained when $p = .5$, an upper bound on the standard error is $\sqrt{(.5)(.5)/n} = 1/(2\sqrt{n})$. Notice that the precision of the estimator $\hat{P}$ improves (i.e., its standard error decreases) as the sample size n increases. ∎

Example 7.5 In the solubility study of Example 7.2, suppose we use the estimator $\overline{X}$ to estimate μ. Properties of $\overline{X}$ derived in Chapter 6 include

$$E(\overline{X}) = \mu \quad \text{and} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}},$$

where σ denotes the standard deviation of the population distribution of solubility measurements (another parameter whose value is unknown). Thus, the sampling distribution of $\overline{X}$ is centered at μ—i.e., $\overline{X}$ is an unbiased estimator of μ—regardless of its value and the sample size n. As with the sample proportion, the standard error of the sample mean decreases (that is, its precision improves) with increasing sample size.

Since the value of σ is almost always unknown, we can estimate the standard error of $\overline{X}$ by $\hat{\sigma}_{\overline{X}} = s/\sqrt{n}$, where s denotes the sample standard deviation. For the 16 observations presented in Example 7.2, $s = 11.66$. The estimated standard error is then $s/\sqrt{n} = 11.66/\sqrt{16} = 2.92$. This quantity indicates that, based on the available data, we believe our estimate of μ, $\bar{x} = 985.5 \text{ mmol/kg}$, is liable to differ by about ± 2.92 mmol/kg from the actual value of μ. ∎

Example 7.6 Suppose that X, the reaction time (s) to a stimulus, has a uniform distribution on the interval from 0 to an unknown upper limit θ. An investigator wants to estimate θ on the basis of a random sample $X_1,\ X_2,\ ...,\ X_n$ of reaction times. Since θ is the largest possible reaction time in the entire population, consider as a first estimator the largest sample reaction time: $\hat{\theta}_b = \max(X_1,\ ...,\ X_n)$. If $n = 5$ and $x_1 = 1.7$, $x_2 = 4.2$, $x_3 = 2.4$, $x_4 = 3.9$, $x_5 = 1.3$, the point estimate of θ is $\hat{\theta}_b = \max(1.7, 4.2, 2.4, 3.9, 1.3) = 4.2$ s.

For an *unbiased* estimator, some samples will yield estimates that exceed θ and other samples will yield estimates smaller than θ—otherwise θ could not possibly be the center of the estimator's distribution. However, our proposed estimator $\hat{\theta}_b$ will *never* overestimate θ—the largest sample value cannot exceed the largest population value—and will underestimate θ unless the largest sample value equals θ. This intuitive argument shows that $\hat{\theta}_b$ is a biased estimator (hence the subscript b). More precisely, using results on ordered values from a random sample (Section 5.7), it can be shown (see Exercise 62) that

$$E(\hat{\theta}_b) = \frac{n}{n+1} \cdot \theta < \theta \quad \text{and} \quad V(\hat{\theta}_b) = \frac{n\theta^2}{(n+1)^2(n+2)}$$

The bias of $\hat{\theta}_b$ is given by $E(\hat{\theta}_b) - \theta = n\theta/(n+1) - \theta = -\theta/(n+1)$. Because the bias is negative, we say that $\hat{\theta}_b$ is *biased low*, meaning that it systematically *under*estimates the true value of θ. Thankfully, the bias approaches 0 as n increases and is negligible for large n. The standard error of $\hat{\theta}_b$ can be estimated by substituting the known value of $\hat{\theta}_b$ for the unknown θ in the square root of the variance formula above.

It is easy to modify $\hat{\theta}_b$ to obtain an unbiased estimator of θ. Consider the estimator

$$\hat{\theta}_u = \frac{n+1}{n} \cdot \hat{\theta}_b = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n)$$

Using this estimator on the data gives the estimate $(6/5)(4.2) = 5.04$ s. The fact that $(n + 1)/n > 1$ implies that $\hat{\theta}_u$ will overestimate θ for some samples and underestimate it for others. The mean value of this estimator is

$$E(\hat{\theta}_u) = E\left[\frac{n+1}{n} \cdot \hat{\theta}_b\right] = \frac{n+1}{n} \cdot E[\hat{\theta}_b]$$
$$= \frac{n+1}{n} \cdot \frac{n}{n+1} \theta = \theta$$

Thus, by definition, $\hat{\theta}_u$ is an unbiased estimator of θ. If $\hat{\theta}_u$ is used repeatedly on different samples to estimate θ, some estimates will be too large and others will be too small, but in the long run there will be no systematic tendency to underestimate or overestimate θ. ∎

Mean Squared Error

Rather than consider bias and variance (accuracy and precision) separately, another popular way to quantify the idea of $\hat{\theta}$ being close to θ is to consider the squared error $(\hat{\theta} - \theta)^2$. For some samples, $\hat{\theta}$ will be quite close to θ and the resulting squared error will be very small, whereas the squared error will be quite large whenever a sample produces an estimate $\hat{\theta}$ that is far from the target. An omnibus

measure of quality is the mean squared error (expected squared error), which entails averaging the squared error over all possible samples and resulting estimates.

DEFINITION The **mean squared error (MSE)** of an estimator $\hat{\theta}$ is $E[(\hat{\theta} - \theta)^2]$.

For the estimators whose distributions are displayed in Figure 7.1a and b, the mean squared error is comparatively large, since there are many values of $\hat{\theta}$ in those two distributions that are quite some distance from θ. On the other hand, the estimator $\hat{\theta}_3$ in Figure 7.1c has much lower MSE. In fact, mean squared error penalizes an estimator for having either high bias (poor accuracy) *or* high variance (poor precision), as indicated by the following proposition.

PROPOSITION For any estimator $\hat{\theta}$ of a parameter θ,

$$\text{MSE} = V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 = \text{variance of estimator} + (\text{bias})^2$$

In particular, for any unbiased estimator of θ, its MSE and variance are equal.

The proof of this result is a simple application of the variance shortcut formula and is left as an exercise (Exercise 23).

Example 7.7 (Example 7.4 continued) Consider once again estimating a population proportion of "successes" p. We have already established that the sample proportion $\hat{P} = X/n$ is an unbiased estimator of p with variance equal to $p(1 - p)/n$. Hence, its mean squared error is

$$E[(\hat{P} - p)^2] = V(\hat{P}) + 0^2 = \frac{p(1 - p)}{n}$$

Now consider the alternative estimator $\tilde{P} = (X + 2)/(n + 4)$; that is, add two successes and two failures to the sample and then calculate the new sample proportion of successes. One intuitive justification for this estimator is that

$$\left|\frac{X}{n} - .5\right| = \left|\frac{X - .5n}{n}\right| \quad \text{while} \quad \left|\frac{X + 2}{n + 4} - .5\right| = \left|\frac{X - .5n}{n + 4}\right|,$$

from which we see that $\tilde{P}$ is always somewhat closer to .5 than is $\hat{P}$. (It seems particularly reasonable to move the estimate toward .5 when the number of successes in the sample is close to 0 or n. For example, if there are no successes at all in the sample, is it sensible to estimate the population proportion of successes as zero, especially if n is small?)

The bias of $\tilde{P}$ is

$$E\left(\frac{X + 2}{n + 4}\right) - p = \frac{E(X) + 2}{n + 4} - p = \frac{np + 2}{n + 4} - p = \frac{2/n - 4p/n}{1 + 4/n}$$

This bias is not zero unless $p = .5$. However, as n increases the numerator approaches zero and the denominator approaches 1, so the bias approaches zero. The variance of $\tilde{P}$ is

$$V\left(\frac{X+2}{n+4}\right) = \frac{V(X+2)}{(n+4)^2} = \frac{V(X)}{(n+4)^2} = \frac{np(1-p)}{(n+4)^2} = \frac{p(1-p)}{n+8+16/n}$$

This variance approaches zero as the sample size increases. Finally, the mean squared error of $\tilde{P}$ is

$$\text{MSE} = \frac{p(1-p)}{n+8+16/n} + \left(\frac{2/n - 4p/n}{1+4/n}\right)^2$$

So how does the mean squared error of the usual estimator $\hat{P}$ compare to that of the alternative estimator $\tilde{P}$? If one MSE were smaller than the other for all values of p, then we could say that one estimator is always preferred to the other (using MSE as our criterion). But as Figure 7.2 shows, this is not the case at least for the sample sizes $n = 10$ and $n = 100$, and in fact is not true for any other sample size.
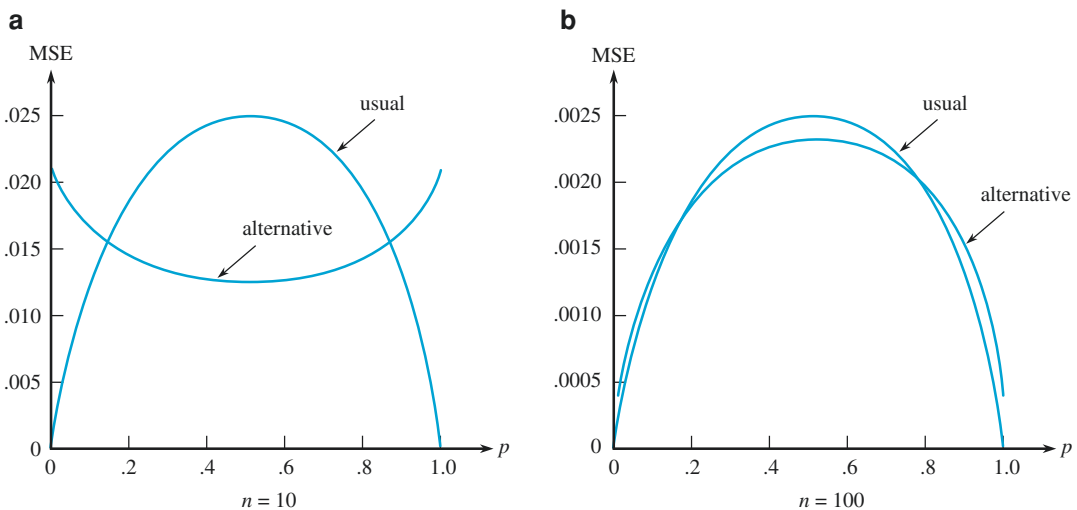
Figure 7.2 Graphs of MSE for the usual and alternative estimators of p

According to Figure 7.2, the two MSEs are quite different when n is small. In this case the alternative estimator is better for values of p near .5 (since it moves the sample proportion toward .5) but not for extreme values of p. For large n, the two MSEs are quite similar, but again neither dominates the other. ∎

Example 7.8 (Example 7.3 continued) Let's return now to the problem of estimating population variance σ^2 based on a random sample $X_1, \ldots, X_n$. First consider the sample variance estimator $S^2 = \sum(X_i - \overline{X})^2/(n-1)$. Applying the property $E(Y^2) = V(Y) + [E(Y)]^2$ to the computing formula $\sum(X_i - \overline{X})^2 = \sum X_i^2 - (1/n)(\sum X_i)^2$ from Section 1.4 gives

$$E\left[\sum (X_i - \overline{X})^2\right] = E\left[\sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2\right]$$

$$= \sum E(X_i^2) - \frac{1}{n}E\left[\left(\sum X_i\right)^2\right]$$

$$= \sum (\sigma^2 + \mu^2) - \frac{1}{n}\left[V\left(\sum X_i\right) + \left(E\left[\sum X_i\right]\right)^2\right]$$

$$= n\sigma^2 + n\mu^2 - \frac{1}{n}\left[n\sigma^2 + (n\mu)^2\right]$$

$$= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 = (n-1)\sigma^2 \Rightarrow$$

$$E[S^2] = \frac{1}{n-1}E\left[\sum (X_i - \overline{X})^2\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

Thus we have shown that *the sample variance S^2 is an unbiased estimator of σ^2 for any population distribution.*

The estimator from Example 7.3 that uses divisor n can be expressed as $(n-1)S^2/n$, and

$$E\left[\frac{(n-1)S^2}{n}\right] = \frac{n-1}{n}E(S^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

This estimator is therefore biased; in particular, its bias is $(n-1)\sigma^2/n - \sigma^2 = -\sigma^2/n$. Because the bias is negative, the estimator with divisor n tends to underestimate σ^2, and this is why the divisor $n-1$ is preferred by many statisticians (although when n is large, the bias is small and there is little difference between the two).

This is not quite the whole story, however. Let's now consider *all* estimators of the form

$$\hat{\sigma}^2 = c\sum (X_i - \overline{X})^2$$

The expected value of such an estimator is

$$E\left[c\sum (X_i - \overline{X})^2\right] = cE\left[\sum (X_i - \overline{X})^2\right] = c(n-1)\sigma^2$$

Clearly the only unbiased estimator of this type is the sample variance, with $c = 1/(n-1)$. Annoyingly, the variance of $\hat{\sigma}^2$ depends on the underlying population distribution. So suppose the random sample has come from a normal distribution. Then from Section 6.4, we know that the rv $(n-1)S^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom. The variance of a χ_{n-1}^2 rv is $2(n-1)$, so the variance of the estimator is

$$V\left[c\sum (X_i - \overline{X})^2\right] = V\left[c\sigma^2 \cdot \frac{(n-1)S^2}{\sigma^2}\right] = (c\sigma^2)^2 V\left[\frac{(n-1)S^2}{\sigma^2}\right] = c^2\sigma^4 \cdot 2(n-1)$$

Substituting these expressions into the relationship MSE = variance + (bias)^2, the value of c for which MSE is minimized turns out to be $c = 1/(n+1)$; see Exercise 65. So in this situation, minimizing the MSE yields a rather unnatural (and never used) estimator.

As a final blow, even though S^2 is unbiased for estimating σ^2, *it is not true* that the sample standard deviation S is unbiased for estimating σ. This is because the square root function is not linear, and the expected value of the square root is not the square root of the expected value. Why not find an

unbiased estimator for σ and use it, rather than S? Unfortunately there is no estimator of σ that is unbiased for all possible population distributions (although in special cases, such as the normal distribution, an unbiased estimator can be deduced). Thankfully, the bias of S is not serious unless n is quite small, so we shall generally employ it as an estimator of σ. ∎

Example 7.9 (Example 7.6 continued) Consider again the two estimators $\hat{\theta}_b$ and $\hat{\theta}_u$ for the population maximum of a Uniform $[0, \theta]$ distribution. Using the formulas presented in Example 7.6, the mean squared error of $\hat{\theta}_b$ is given by

$$\text{MSE} = \text{variance} + (\text{bias})^2 = \frac{n\theta^2}{(n+1)^2(n+2)} + \left(-\frac{\theta}{n+1}\right)^2 = \frac{2\theta^2}{(n+1)(n+2)}$$

Since $\hat{\theta}_u$ was found to be unbiased, its mean squared error is simply its variance:

$$\text{MSE} = V(\hat{\theta}_u) = V\left(\frac{n+1}{n} \cdot \hat{\theta}_b\right) = \left(\frac{n+1}{n}\right)^2 \cdot V(\hat{\theta}_b) = \left(\frac{n+1}{n}\right)^2 \cdot \frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}$$

Taken together, we find that $\hat{\theta}_u$ has less bias than $\hat{\theta}_b$ (obviously) but a larger variance. The use of mean squared error combines these two considerations, and for $n > 1$ it can be shown that $\hat{\theta}_u$ has a smaller MSE than $\hat{\theta}_b$ and is therefore the preferred estimator. ∎

Unbiased Estimation
Finding an estimator whose mean squared error is smaller than that of every other estimator for all values of the parameter is sometimes not feasible. One common approach is to restrict the class of estimators under consideration in some way, and then seek the estimator that is best in that restricted class.

Statistical practitioners who buy into the *Principle of Unbiased Estimation* would employ an unbiased estimator in preference to a biased estimator, even if the latter has a smaller MSE. On this basis, the sample proportion of successes should be preferred to the alternative estimator of p in Example 7.7, and the unbiased estimator $\hat{\theta}_u$ should be preferred to the biased estimator $\hat{\theta}_b$ in Example 7.9 (minimizing MSE would lead us to the same estimator in that instance).

In Example 7.2, we proposed several different estimators for the mean μ of a symmetric distribution. If there were a *unique* unbiased estimator for μ, the estimation dilemma could be resolved by using that estimator. Unfortunately, this is not the case.

PROPOSITION If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ, then $\overline{X}$ is an unbiased estimator of μ. If in addition the distribution is continuous and symmetric, then $\tilde{X}$ and any trimmed mean are also unbiased estimators of μ.

The fact that $E(\overline{X}) = \mu$, so $\overline{X}$ is an unbiased estimator of μ, was established previously. The unbiasedness of the other estimators is more difficult to verify; the argument requires invoking results on distributions of ordered values from Section 5.7.

According to the preceding proposition, the Principle of Unbiased Estimation by itself does not always allow us to select a single estimator. When the underlying population is normal, even the third estimator in Example 7.2 is unbiased, and there are many other unbiased estimators. If two or more estimators of a parameter are unbiased, then naturally one selects the estimator among them with the

smallest standard error (equivalently, the least variance). The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator (MVUE)** of θ.

Example 7.10 (Example 7.9 continued) We showed in Example 7.6 that when X_1, …, X_n is a random sample from a uniform distribution on [0, θ], the estimator $\hat{\theta}_u = (n+1)/n \cdot \max(X_1, \ldots, X_n)$ is unbiased for θ. However, this is not the only unbiased estimator of θ. The expected value of a uniformly distributed rv is just the midpoint of the support, so here $E(X_i) = \theta/2$. This implies that $E(\overline{X}) = \theta/2$, from which $E(2\overline{X}) = \theta$. That is, the estimator $\hat{\theta}_2 = 2\overline{X}$ is also unbiased for θ.

If X is uniformly distributed on the interval [0, θ], then from Chapter 4 we have $V(X) = \sigma^2 = (\theta - 0)^2/12 = \theta^2/12$. Hence, the variance (and MSE) of $\hat{\theta}_2$ are

$$V(\hat{\theta}_2) = V(2\overline{X}) = 4V(\overline{X}) = 4 \cdot \frac{\sigma^2}{n} = 4 \cdot \frac{\theta^2/12}{n} = \frac{\theta^2}{3n}$$

For $n > 1$, $V(\hat{\theta}_2)$ will be greater than $V(\hat{\theta}_u)$, so $\hat{\theta}_u$ is a better estimator than $\hat{\theta}_2$. More advanced methods can be used to show that $\hat{\theta}_u$ is the MVUE of θ—that is, *every* other unbiased estimator of θ has variance that exceeds the variance of $\hat{\theta}_u$. ∎

One of the triumphs of mathematical statistics has been the development of methodology for identifying the MVUE in a wide variety of situations. The most important result of this type for our purposes concerns estimating the mean μ of a normal distribution.

THEOREM   Let X_1, …, X_n be a random sample from a normal distribution with parameters μ and σ. Then the estimator $\hat{\mu} = \overline{X}$ is the MVUE for μ.

Whenever we are convinced that the population being sampled is normal, the result says that $\overline{X}$ should be used to estimate μ. For a proof in the special case that σ is known, see Exercise 55.

Again, in some situations it is possible to obtain an estimator with small bias that would be preferred to the best unbiased estimator. This is illustrated in Figure 7.3. However, MVUEs are often easier to obtain than the type of biased estimator whose distribution is pictured.

pdf of $\hat{\theta}_1$, a biased estimator

pdf of $\hat{\theta}_2$, the MVUE

θ
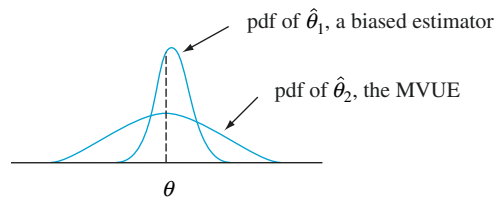
Figure 7.3   A biased estimator that is preferable to the MVUE

Consistency

As a researcher's sample size increases and thus more of the population is observed, any reasonable estimator should, in some sense, "converge to" the parameter it is estimating. For instance, the Law of Large Numbers (Section 6.2) states that the sample mean $\overline{X}$ of a random sample converges to the theoretical mean μ in a specific mathematical sense as $n \to \infty$. This intuitive notion is called *consistency*.

DEFINITION   Let $X_1, \ldots, X_n$ be a random sample from a distribution that depends on a parameter θ. Then an estimator $\hat{\theta}$ is a **consistent** estimator of θ if $\hat{\theta}$ converges to θ as $n \to \infty$, either in the sense that

1. $E[(\hat{\theta} - \theta)^2] \to 0$ as $n \to \infty$, or
2. $P(|\hat{\theta} - \theta| \geq \varepsilon) \to 0$ as $n \to \infty$ for every $\varepsilon > 0$.[1]

Statement 1 in the definition requires that the mean squared error of $\hat{\theta}$ converge to 0 as the sample size increases to infinity; this is known formally as *convergence in mean square* or *convergence in quadratic mean*. Statement 2 says that $\hat{\theta}$ *converges to θ in probability*. Intuitively, this means that the chance of an estimate differing from the value of θ by *any* small amount approaches 0 as the sample size increases. Other examples of consistent estimators include the sample proportion $\hat{P}$ as an estimator of a population proportion p and sample standard deviation S as an estimator of population standard deviation σ. All estimators introduced in subsequent chapters are consistent.

Some Complications

Although it was stated previously that $\overline{X}$ is the MVUE for a population mean when sampling from a normal distribution, that does not mean $\overline{X}$ should be used irrespective of the distribution being sampled.

Example 7.11 Suppose we wish to estimate the number of calories θ in a certain food. Using standard measurement techniques, we will obtain a random sample $X_1, \ldots, X_n$ of n calorie measurements. Imagine that the population distribution is a member of one of the following three families:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\theta)^2/(2\sigma^2)} \qquad -\infty < x < \infty \tag{7.1}$$

$$f(x) = \frac{1}{\pi[1 + (x-\theta)^2]} \qquad -\infty < x < \infty \tag{7.2}$$

$$f(x) = \frac{1}{2c} \qquad \theta - c \leq x \leq \theta + c \tag{7.3}$$

The pdf (7.1) is the normal distribution, (7.2) is called the Cauchy distribution, and (7.3) is a uniform distribution. All three distributions are symmetric about θ, which is therefore the median of each distribution. (The value θ is also the mean for the normal and uniform distributions, but the mean of the Cauchy distribution fails to exist.)

Consider the four estimators proposed in Example 7.2: $\overline{X}$, $\tilde{X}$, $\overline{X}_e$ (the average of the two extreme observations), and $\overline{X}_{\text{tr}}$ (a trimmed mean). The best estimator for θ depends crucially on which distribution is being sampled. In particular,

1. If the random sample comes from a normal distribution, then $\overline{X}$ is the best of the four estimators, since it has minimum variance among all unbiased estimators.
2. If the random sample comes from a Cauchy distribution, then $\overline{X}$ and $\overline{X}_e$ are terrible estimators for θ, whereas $\tilde{X}$ is quite good (the MVUE is not known). $\overline{X}$ and $\overline{X}_e$ are bad because they are very sensitive to outlying observations, and the heavy tails of the Cauchy distribution make a few such observations likely to appear in any sample.

---

[1]In fact, Statement 1 implies Statement 2. But there exist unusual cases for which Statement 1 fails—typically, when the variance of the estimator is infinite—and Statement 2 still holds.

3. If the underlying distribution is the particular uniform distribution in (7.3), then the best estimator is $\overline{X}_e$; in general, this estimator is greatly influenced by outlying observations, but here the lack of tails makes such observations impossible.

4. *The trimmed mean is best in none of these three situations but works reasonably well in all three.* That is, $\overline{X}_{\text{tr}}$ does not suffer too much in comparison with the best procedure in any of the three situations.

More generally, research over the past several decades has established that when estimating a point of symmetry of a continuous probability distribution, a trimmed mean with trimming proportion between 10 and 20% from each end of the sample produces reasonably behaved estimates over a very wide range of possible population models. For this reason, a trimmed mean with small trimming percentage is said to be a **robust estimator**.   ■

Example 7.12 Suppose a type of component has a lifetime distribution that is exponential with parameter λ, so that expected lifetime is $\mu = 1/\lambda$. A sample of n such components is selected, and each is put into operation. If the experiment is continued until all n lifetimes $X_1, \ldots, X_n$ have been observed, then $\overline{X}$ is an unbiased estimator of μ.

In some experiments, though, the components are left in operation only until the time of the rth failure, where $r < n$. This procedure is referred to as **censoring**. Let Y_1 denote the time of the first failure (the minimum lifetime among the n components), Y_2 denote the time at which the second failure occurs (the second smallest lifetime), and so on. Since the experiment terminates at time Y_r, the total accumulated lifetime at termination is

$$T_r = \sum_{i=1}^{r} Y_i + (n - r)Y_r$$

We now demonstrate that $\hat{\mu} = T_r/r$ is an unbiased estimator for μ. To do so, we need two properties of exponential variables:

1. The memoryless property (see Section 4.4) says that at any time point, remaining lifetime has the same exponential distribution as original lifetime.

2. If $X_1, \ldots, X_k$ are independent exponential rvs with parameter λ, then $\min(X_1, \ldots, X_k)$ is exponential with parameter $k\lambda$ and has expected value $1/(k\lambda)$. See Example 5.39.

Since all n components last until Y_1, $n - 1$ last an additional $Y_2 - Y_1$, $n - 2$ an additional $Y_3 - Y_2$ amount of time, and so on, another expression for T_r is

$$T_r = nY_1 + (n - 1)(Y_2 - Y_1) + (n - 2)(Y_3 - Y_2) + \cdots + (n - r + 1)(Y_r - Y_{r-1})$$

But Y_1 is the minimum of n exponential variables, so $E(Y_1) = 1/(n\lambda)$. Similarly, $Y_2 - Y_1$ is the smallest of the $n - 1$ remaining lifetimes, each exponential with parameter λ (by the memoryless property), so $E(Y_2 - Y_1) = 1/[(n - 1)\lambda]$. Continuing, $E(Y_{i+1} - Y_i) = 1/[(n - i)\lambda]$, so

$$E(T_r) = nE(Y_1) + (n - 1)E(Y_2 - Y_1) + \cdots + (n - r + 1)E(Y_r - Y_{r-1})$$
$$= n \cdot \frac{1}{n\lambda} + (n - 1) \cdot \frac{1}{(n - 1)\lambda} + \cdots + (n - r + 1) \cdot \frac{1}{(n - r + 1)\lambda} = \frac{r}{\lambda}$$

Therefore, $E(T_r/r) = (1/r)E(T_r) = (1/r) \cdot (r/\lambda) = 1/\lambda = \mu$ as claimed.

As an example, suppose 20 components are tested and $r = 10$. Then if the first ten failure times are 11, 15, 29, 33, 35, 40, 47, 55, 58, and 72, the point estimate of μ is

$$\hat{\mu} = \frac{11 + 15 + \cdots + 72 + (10)(72)}{10} = 111.5$$

The advantage of the experiment with censoring is that it terminates more quickly than the uncensored experiment. However, it can be shown that $V(T_r/r) = 1/(\lambda^2 r)$, which is larger than $1/(\lambda^2 n)$, the variance of $\overline{X}$ in the uncensored experiment. ∎

The form of an estimator $\hat{\theta}$ may be sufficiently complicated so that standard statistical theory cannot be applied to obtain a formula for its standard error. If we assume the population has a certain distribution $f(x; \theta)$, then we can use software to simulate repeated samples from that distribution, calculate the value of $\hat{\theta}$ for each sample, and use the standard deviation of these various $\hat{\theta}$ values to estimate $\sigma_{\hat{\theta}}$. Of course, software packages cannot perform such a simulation without the user specifying a numerical value of θ in advance, and the value of θ is unknown for our data. In many simulation studies, the researcher will therefore perform this process for a variety of θ values, each one returning a different estimated standard error of $\hat{\theta}$.

On other occasions, sample data is available from which a point estimate $\hat{\theta}$ has been obtained, so we have an estimate of θ but no measure of the uncertainty in that estimate. In that scenario, repeated values from the pdf $f(x; \hat{\theta})$—that is, the pdf specified by plugging in $\theta = \hat{\theta}$—are simulated, and the estimated standard error is obtained as before. This procedure is known as the *parametric bootstrap*; we will consider bootstrap methods in greater depth in subsequent chapters.

Exercises: Section 7.1 (1–24)

1. The accompanying data on IQ for first graders at a university laboratory school was introduced in Exercise 81 of Chapter 1.

82	96	99	102	103	103	106	107	108	108	108
108	109	110	110	111	113	113	113	113	115	115
118	118	119	121	122	122	127	132	136	140	146

   a. Calculate a point estimate of the mean value of IQ for the conceptual population of all first graders in this school, and state which estimator you used. [*Hint*: $\sum x_i = 3753$.]
   b. Calculate a point estimate of the IQ value that separates the lowest 50% of all such students from the highest 50%, and state which estimator you used.
   c. Calculate and interpret a point estimate of the population standard deviation σ. Which estimator did you use? [*Hint*: $\sum x_i^2 = 432{,}015$.]
   d. Calculate a point estimate of the proportion of all such students whose IQ exceeds

   100. [*Hint*: Think of an observation as a "success" if it exceeds 100.]
   e. Calculate a point estimate of the population *coefficient of variation* σ/μ, and state which estimator you used.

2. A sample of 20 students who had recently taken introductory statistics yielded the following information on brand of calculator owned (T = Texas Instruments, H = Hewlett-Packard, C = Casio, S = Sharp):

| T | T | H | T | C | T | T | S | C | H |
| S | S | T | H | C | T | T | T | H | T |

   a. Estimate the true proportion of all such students who own a Texas Instruments calculator.
   b. Of the ten students who owned a TI calculator, 4 had graphing calculators. Estimate the proportion of students who do not own a TI graphing calculator.

3. Consider the following sample of observations on coating thickness for low-viscosity paint ("Achieving a Target Value for a Manufacturing Process: A Case Study," *J. Qual. Technol.* 1992: 22–26):

.83	.88	.88	1.04	1.09	1.12	1.29	1.31
1.48	1.49	1.59	1.62	1.65	1.71	1.76	1.83

Assume that the distribution of coating thickness is normal (a normal probability plot strongly supports this assumption).

a. Calculate a point estimate of the mean value of coating thickness, and state which estimator you used.
b. What is the estimated standard error of the estimator that you used in part (a)?
c. Calculate a point estimate of the median of the coating thickness distribution, and state which estimator you used.
d. Calculate a point estimate of the value that separates the largest 10% of all values in the thickness distribution from the remaining 90%, and state which estimator you used. [*Hint:* Express what you are trying to estimate in terms of μ and σ.]
e. Estimate $P(X < 1.5)$, i.e., the proportion of all thickness values less than 1.5. [*Hint:* If you knew the values of μ and σ, you could calculate this probability. These values are not available, but they can be estimated.]

4. The data set mentioned in Exercise 1 also includes these third grade verbal IQ observations for males and females, respectively.

Males								
117	103	121	112	120	132	113	117	132
149	125	131	136	107	108	113	136	114
Females								
114	102	113	131	124	117	120	90	
114	109	102	114	127	127	103		

Prior to obtaining data, denote the male values by $X_1 \ldots, X_m$ and the female values by $Y_1, \ldots, Y_n$. Suppose that the X_i's constitute a random sample from a distribution with mean μ_1 and standard deviation σ_1 and that the Y_i's form a random sample (independent of the X_i's) from another distribution with mean μ_2 and standard deviation σ_2.

a. Use rules of expected value to show that $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. Calculate the estimate for the given data.
b. Use rules of variance from Chapter 5 to obtain expressions for the variance and standard deviation (standard error) of the estimator in part (a), and then compute the estimated standard error.
c. Calculate a point estimate of the ratio σ_1/σ_2 of the two standard deviations.
d. Suppose one male third grader and one female third grader are randomly selected. Calculate a point estimate of the standard deviation of the difference $X - Y$ between male and female IQ.

5. As an example of a situation in which several different statistics could reasonably be used to calculate a point estimate, consider a population of N invoices. Associated with each invoice is its "book value," the recorded amount of that invoice. Let T denote the total book value, a known amount. Some of these book values are erroneous. An audit will be carried out by randomly selecting n invoices and determining the audited (correct) value for each one. Suppose that the sample gives the following results (in dollars).

	Invoice				
	1	2	3	4	5
Book value	300	720	526	200	127
Audited value	300	520	526	200	157
Error	0	200	0	0	−30

Let $\overline{X} =$ the sample mean audited value, $\overline{Y} =$ the sample mean book value, and $\overline{D} =$ the sample mean error. Propose three different statistics for estimating the total audited (i.e., correct) value θ—one involving just N and $\overline{X}$, another involving N, T, and $\overline{D}$, and the last involving T and $\overline{X}/\overline{Y}$. Then calculate

the resulting estimates when $N = 5000$ and $T = 1{,}761{,}300$. [The article "Statistical Models and Analysis in Auditing," *Stat. Sci.* 1989: 2–33 discusses properties of these estimators.]

6. Consider the accompanying data on cycles to failure for a sample of 12 turbine blades ("Effect of Aluminized Coating on Combined Low and High Cycle Fatigue Life of Turbine Blade at Elevated Temperature," *J. Engr. Gas Turbines Power* 2019):

| 209 | 226 | 281 | 494 | 568 | 953 |
| 488 | 655 | 943 | 973 | 1193 | 1358 |

The article's authors used an appropriate probability plot to support the use of a lognormal distribution (see Section 4.5) as a model for cycles to failure.

a. Estimate the parameters of the distribution. [*Hint*: Remember that X has a lognormal distribution with parameters μ and σ if $\ln(X)$ is normally distributed with mean μ and standard deviation σ.]

b. Use the estimates of part (a) to estimate the true mean cycles to failure for this type of turbine blade. [*Hint*: What is $E(X)$ for the lognormal distribution?]

7. a. A random sample of 10 houses in a particular area, each of which is heated with natural gas, is selected and the amount of gas (therms) used during the month of January is determined for each house. The resulting observations are 103, 156, 118, 89, 125, 147, 122, 109, 138, 99. Let μ denote the average gas usage during January by all houses in this area. Compute a point estimate of μ.

b. Suppose there are 10,000 houses in this area that use natural gas for heating. Let τ denote the total amount of gas used by all of these houses during January. Estimate τ using the data of part (a). What

estimator did you use in computing your estimate?

c. Use the data in part (a) to estimate p, the proportion of all houses that used at least 100 therms.

d. Give a point estimate of the population median usage (the middle value in the population of all houses) based on the sample of part (a). What estimator did you use?

8. In a random sample of 80 components of a certain type, 12 are found to be defective.

a. Give a point estimate of the proportion of all such components that are *not* defective.

b. A system is to be constructed by randomly selecting two of these components and connecting them in series, as shown here.

The series connection implies that the system will function if and only if neither component is defective (i.e., both components work properly). Estimate the proportion of all such systems that work properly. [*Hint*: If p denotes the probability that a component works properly, express $P(\text{system works})$ in terms of p.]

c. Let $\hat{P}$ be the sample proportion of successes. Is $\hat{P}^2$ an unbiased estimator for p^2? [*Hint*: Recall that for any rv Y, $E(Y^2) = V(Y) + [E(Y)]^2$.]

9. Each of 150 newly manufactured items is examined and the number of scratches per item is recorded (the items are supposed to be free of scratches), yielding the following data:

Number of scratches per item	0	1	2	3	4	5	6	7
Observed frequency	18	37	42	30	13	7	2	1

Let X = the number of scratches on a randomly chosen item, and assume that X has a Poisson distribution with parameter μ.

a. Find an unbiased estimator of μ and compute the estimate for the data.

b. What is the standard deviation (standard error) of your estimator? Compute the estimated standard error. [Hint: $\sigma_X^2 = \mu$ when X is Poisson.]

10. Using a long rod that has length μ, you are going to lay out a square plot in which the length of each side is μ. Thus the area of the plot will be μ^2. However, you do not know the value of μ, so you decide to make n independent measurements X_1, X_2, ... X_n of the length. Assume that each X_i has mean μ (unbiased measurements) and variance σ^2.

a. Show that $\overline{X}^2$ is *not* an unbiased estimator for μ^2. [Hint: For any rv Y, $E(Y^2) = V(Y) + [E(Y)]^2$. Apply this with $Y = \overline{X}$.]

b. For what value of k is the estimator $\overline{X}^2 - kS^2$ unbiased for μ^2? [Hint: Compute $E(\overline{X}^2 - kS^2)$.]

11. Let X_1 (X_2) denote the number of male (female) teenagers in a random sample of size n_1 (n_2) who have vaped during the previous 12 months. Denote the probabilities that a randomly selected teenage male and female vaped in the last 12 months by p_1 and p_2, respectively. Define $\hat{P}_i = X_i/n_i$ for i = 1, 2.

a. Show that $\hat{P}_1 - \hat{P}_2$ is an unbiased estimator for $p_1 - p_2$. [Hint: $E(X_i) = np_i$ for i = 1, 2.]

b. What is the standard error of the estimator in part (a)?

c. How would you use the observed values x_1 and x_2 to estimate the standard error of your estimator?

d. If $n_1 = n_2 = 200$, $x_1 = 107$, and $x_2 = 62$, use the estimator of part (a) to obtain an estimate of $p_1 - p_2$.

e. Use the result of part (c) and the data of part (d) to estimate the standard error of the estimator.

12. Suppose a certain type of fertilizer has an expected yield per acre of μ_1 with variance σ^2, whereas the expected yield for a second type of fertilizer is μ_2 with the same variance σ^2. Let S_1^2 and S_2^2 denote the sample variances of yields based on sample sizes n_1 and n_2, respectively, of the two fertilizers. Show that the following pooled (combined) estimator is unbiased for estimating σ^2:

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

13. The time a customer spends in service after waiting in a queue is often modeled with an exponential distribution. Let $X_1, \ldots, X_n$ be a random sample of service times. Since the parameter λ of the exponential distribution is the reciprocal of the expected value, a reasonable estimator of λ is $\hat{\lambda} = 1/\overline{X}$.

a. Show using a moment generating function argument that $\overline{X}$ has a gamma distribution, with parameters $\alpha = n$ and $\beta = 1/(n\lambda)$.

b. Show that the mean and variance of the estimator $\hat{\lambda}$ are

$$E(\hat{\lambda}) = \frac{n\lambda}{n-1} \quad \text{and}$$

$$V(\hat{\lambda}) = \frac{n^2\lambda^2}{(n-1)^2(n-2)}$$

[Hint: Determine $E(1/Y)$ and $E(1/Y^2)$ when Y has a gamma distribution, using the gamma pdf and Expression (4.5). For the variance, use the variance shortcut formula.]

c. Propose a formula for the estimated standard error of $\hat{\lambda}$.

14. Refer back to the previous exercise. Consider the following alternative estimator of the parameter λ:

$$\hat{\lambda}_a = \frac{n-1}{\sum X_i} = \frac{n-1}{n} \cdot \frac{1}{\overline{X}} = \frac{n-1}{n}\hat{\lambda}$$

a. Determine the mean, variance, and MSE of $\hat{\lambda}_a$. [*Hint*: Use rescaling properties.]

b. Which of the two estimators, $\hat{\lambda}$ or $\hat{\lambda}_a$, is preferable? Explain your reasoning.

15. Consider a random sample $X_1, \ldots, X_n$ from the pdf

$$f(x; \theta) = .5(1 + \theta x) \quad -1 \le x \le 1$$

for some $-1 \le \theta \le 1$ (this distribution arises in particle physics). Show that $\hat{\theta} = 3\overline{X}$ is an unbiased estimator of θ. [*Hint*: First determine $\mu = E(X) = E(\overline{X})$.]

16. A sample of n captured jet fighters results in serial numbers $x_1, x_2, x_3, \ldots, x_n$. The CIA knows that the aircraft were numbered consecutively at the factory starting with α and ending with β, so that the total number of planes manufactured is $\beta - \alpha + 1$ (e.g., if $\alpha = 17$ and $\beta = 29$, then $29 - 17 + 1 = 13$ planes having serial numbers 17, 18, 19, …, 28, 29 were manufactured). However, the CIA does not know the values of α or β. A CIA statistician suggests using the estimator $\max(X_i) - \min(X_i) + 1$ to estimate the total number of planes manufactured.

a. If $n = 5$, $x_1 = 237$, $x_2 = 375$, $x_3 = 202$, $x_4 = 525$, and $x_5 = 418$, what is the corresponding estimate?

b. Under what conditions on the sample will the value of the estimate be exactly equal to the true total number of planes? Will the estimate ever be larger than the true total? Do you think the estimator is unbiased for estimating $\beta - \alpha + 1$? Explain in one or two sentences.

(A similar method was used to estimate German tank production in World War II.)

17. Let $X_1, X_2, \ldots, X_n$ represent a random sample from a *Rayleigh distribution* with pdf

$$f(x; \theta) = \frac{x}{\theta} e^{-x^2/(2\theta)} \quad x > 0$$

a. It can be shown that $E(X^2) = 2\theta$. Use this fact to construct an unbiased estimator of θ based on $\sum X_i^2$ (and use rules of expected value to show that it is unbiased).

b. Estimate θ from the following measurements of blood plasma beta concentration (in pmol/L) for $n = 10$ men, assuming the population of measurements follows a Rayleigh distribution.

16.88	10.23	4.59	6.66	13.68
14.23	19.87	9.40	6.51	10.95

18. Suppose the true average growth μ of one type of plant during a 1-year period is identical to that of a second type, but the variance of growth for the first type is σ^2, whereas for the second type, the variance is $4\sigma^2$. Let $X_1, \ldots, X_m$ be m independent growth observations on the first type [so $E(X_i) = \mu$, $V(X_i) = \sigma^2$], and let $Y_1, \ldots, Y_n$ be n independent observations on the second type [$E(Y_i) = \mu$, $V(Y_i) = 4\sigma^2$]. Let c be a numerical constant and consider the estimator $\hat{\mu} = c\overline{X} + (1 - c)\overline{Y}$; for any c between 0 and 1, this is a weighted average of the two sample means.

a. Show that for any c the estimator is unbiased.

b. For fixed m and n, what value c minimizes $V(\hat{\mu})$? [*Hint*: The estimator is a linear combination of the two sample means and these means are independent. Once you have an expression for the variance, differentiate with respect to c.]

19. In Chapter 3, we defined a negative binomial rv as the number of trials required to achieve the rth success in a sequence of independent and identical success/failure trials. The probability mass function (pmf) of X is

$$nb(x, r, p) = \binom{x - 1}{r - 1} p^r (1 - p)^{x-r}$$

$$x = r, r + 1, r + 2, \ldots$$

a. Suppose that $r \geq 2$. Show that

$$\hat{P} = (r-1)/(X-1)$$

is an unbiased estimator for p. [*Hint*: Write out $E(\hat{P})$ as a sum, then make the substitutions $y = x - 1$ and $s = r - 1$.]

b. A reporter wishing to interview five individuals who support a certain candidate begins asking people whether (S) or not (F) they support the candidate. If the sequence of responses is *SFFSFFFSSS*, estimate p = the true proportion who support the candidate.

20. Let $X_1, X_2, \ldots, X_n$ be a random sample from a pdf $f(x)$ that is symmetric about μ, so that $\tilde{X}$ is an unbiased estimator of μ. If n is large, it can be shown that $V(\tilde{X}) \approx 1/\{4n[f(\mu)]^2\}$. When the underlying pdf is Cauchy (see Example 7.11), $V(\overline{X}) = \infty$, so $\overline{X}$ is a terrible estimator. What is $V(\tilde{X})$ in this case when n is large?

21. An investigator wishes to estimate the proportion of students at a certain university who have violated the honor code. Having obtained a random sample of n students, she realizes that asking each, "Have you violated the honor code?" will probably result in some untruthful responses. Consider the following scheme, called a **randomized response** technique. The investigator makes up a deck of 100 cards, of which 50 are of type I and 50 are of type II.

Type I: Have you violated the honor code (yes or no)?
Type II: Is the last digit of your telephone number a 0, 1, or 2 (yes or no)?

Each student in the random sample is asked to mix the deck, draw a card, and answer the resulting question truthfully. Because of the irrelevant question on type II cards, a yes response no longer stigmatizes the respondent, so we assume that responses are truthful. Let p denote the proportion of honor-code violators (i.e., the probability of a randomly selected student being a violator), and let $\lambda = P(\text{yes response})$. Then λ and p are related by $\lambda = .5p + (.5)(.3)$.

a. Let Y denote the number of yes responses, so $Y \sim \text{Bin}(n, \lambda)$. Thus Y/n is an unbiased estimator of λ. Derive an estimator for p based on Y. If $n = 80$ and $y = 20$, what is your estimate? [*Hint*: Solve $\lambda = .5p + .15$ for p and then substitute Y/n for λ.]

b. Use the fact that $E(Y/n) = \lambda$ to show that your estimator $\hat{p}$ is unbiased.

c. If there were 70 type I and 30 type II cards, what would be your estimator for p?

22. Return to the problem of estimating the population proportion p and consider another adjusted estimator, namely

$$\hat{P} = \frac{X + \sqrt{n/4}}{n + \sqrt{n}}$$

(The justification for this estimator comes from the Bayesian approach to point estimation.)

a. Determine the mean squared error of this estimator. What is interesting about this MSE?

b. Compare the MSE of this estimator to the MSE of the usual estimator (the sample proportion).

23. Show that $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$ (the mean squared error proposition from earlier in this section). [*Hint*: Write μ for $E(\hat{\theta})$. Expand the two quadratic expressions, and use the variance shortcut formula to rewrite $V(\hat{\theta})$.]

24. Show that $\hat{\theta}$ is a consistent estimator of θ (in the mean-square sense) if and only if both (1) $E(\hat{\theta}) \to \theta$ and (2) $V(\hat{\theta}) \to 0$ as $n \to \infty$.

7.2 The Methods of Moments and Maximum Likelihood

The point estimators introduced in Section 7.1 were obtained via common sense and/or educated guesswork. We now introduce two "constructive" methods for obtaining point estimators: the method of moments and the method of maximum likelihood. By constructive we mean that the general definition of each type of estimator suggests explicitly how to obtain the estimator in any specific problem. Although maximum likelihood estimators are generally preferable to moment estimators because of certain efficiency properties, they often require significantly more computation than do moment estimators. (It is sometimes the case the two methods produce the same estimator.)

The Method of Moments

The basic idea of this method is to equate certain sample characteristics, such as the sample mean, to the corresponding population expected values. Then solving these equations for unknown parameter values yields the estimators.

DEFINITION   Let $X_1, \ldots, X_n$ be a random sample from some distribution. For $k = 1, 2, 3, \ldots$, the **kth population moment**, or **kth moment of the distribution**, is $E(X^k)$. The **kth sample moment** is $(1/n) \sum_{i=1}^{n} X_i^k$.

Thus the first population moment is $E(X) = \mu$ and the first sample moment is $\sum X_i / n = \overline{X}$. The second population and sample moments are $E(X^2)$ and $\sum X_i^2 / n$, respectively. The population moments will be functions of any unknown parameters $\theta_1, \theta_2, \ldots$.

DEFINITION   Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution depending on parameters $\theta_1, \ldots, \theta_m$ whose values are unknown. Then the **method of moments estimators (mmes)** $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are obtained by equating the first m sample moments to the corresponding first m population moments and solving for $\theta_1, \ldots, \theta_m$.

If, for example, $m = 2$, $E(X)$ and $E(X^2)$ will be functions of θ_1 and θ_2. Setting $E(X) = \sum X_i / n = \overline{X}$ and $E(X^2) = \sum X_i^2 / n$ gives two equations in θ_1 and θ_2. The solution then defines the estimators.

Example 7.13 Let $X_1, \ldots, X_n$ represent a random sample of service times of n customers at a certain facility, where the underlying distribution is assumed exponential with parameter λ. Since there is only one parameter to be estimated, the estimator is obtained by equating $E(X)$ to $\overline{X}$. Since $E(X) = 1/\lambda$ for an exponential distribution, this gives $1/\lambda = \overline{X}$ or $\lambda = 1/\overline{X}$. The mme of λ is then $\hat{\lambda} = 1/\overline{X}$. ∎

Example 7.14 Let $X_1, \ldots, X_n$ be a random sample from a gamma distribution with parameters α and β. From Section 4.4, $E(X) = \alpha\beta$ and $E(X^2) = \beta^2 \Gamma(\alpha + 2)/\Gamma(\alpha) = \beta^2(\alpha + 1)\alpha$. The mmes of α and β are obtained by solving

$$\overline{X} = \alpha\beta \qquad \frac{1}{n}\sum X_i^2 = \alpha(\alpha+1)\beta^2$$

A little straightforward algebra gives the estimators

$$\hat{\alpha} = \frac{\left(\overline{X}\right)^2}{\frac{1}{n}\sum X_i^2 - \left(\overline{X}\right)^2} \quad \hat{\beta} = \frac{\frac{1}{n}\sum X_i^2 - \left(\overline{X}\right)^2}{\overline{X}}$$

To illustrate, the article cited in Example 4.29 recommends using the gamma distribution to model times between attempted connections to a server from suspicious IP addresses. The article includes the following $n = 31$ observations for such interarrival times (hours) for one particular server being "hit" by a specific suspicious IP address:

2.3403	8.0347	8.4395	17.3053	2.9156	10.1836	2.1481	4.0839	2.3567	6.0122
1.0270	0.1208	12.9981	14.9370	3.7714	1.3228	0.3270	9.9028	3.4356	4.0326
3.0470	1.3922	0.3828	0.6180	4.0120	4.4803	8.6706	0.2933	2.9467	17.3828
0.9431

from which $\overline{x} = 5.157$ and $(1/31)\sum x_i^2 = 51.168$. The parameter estimates are

$$\hat{\alpha} = \frac{(5.157)^2}{51.168 - (5.157)^2} = 1.082 \quad \hat{\beta} = \frac{51.168 - (5.157)^2}{5.157} = 4.765$$

These estimates of α and β fall into the range of parameter estimates suggested by the article's authors. (We'll consider interval estimates of parameters in Chapter 8.)   ∎

Example 7.15 Let $X_1, \ldots, X_n$ be a random sample from the following discrete distribution:

$$p(x) = \binom{x + r - 1}{r - 1} p^r (1 - p)^x \quad x = 0, 1, 2, \ldots$$

This is a variant on the generalized negative binomial distribution with parameters r and p (see Chapter 3, Exercise 124). It can be shown for this distribution that $E(X) = r(1 - p)/p$ and $V(X) = r(1 - p)/p^2$, from which $E(X^2) = V(X) + [E(X)]^2 = r(1 - p)(r - rp + 1)/p^2$. Equating $E(X)$ to $\overline{X}$ and $E(X^2)$ to $(1/n)\sum X_i^2$ eventually gives

$$\hat{p} = \frac{\overline{X}}{\frac{1}{n}\sum X_i^2 - \left(\overline{X}\right)^2} \quad \hat{r} = \frac{\left(\overline{X}\right)^2}{\frac{1}{n}\sum X_i^2 - \left(\overline{X}\right)^2 - \overline{X}}$$

As an illustration, the article "Chains of Transmission and Control of Ebola Virus Disease in Conakry, Guinea, in 2014: an Observational Study" (*Lancet Infect. Dis.* 2015: 320–326) describes a study of the number of secondary Ebola cases stemming from 152 infected people (a secondary case means they give someone else the disease). The data is as follows:

| Number of cases | 0 | 1 | 2 | 3 | 4 | 5 | 8 | 9 | 14 | 17 |
| Frequency | 109 | 16 | 9 | 5 | 5 | 2 | 1 | 3 | 1 | 1 |

A follow-up letter in the same journal investigated modeling these counts with a generalized negative binomial distribution. First,

$$\bar{x} = \sum x_i/152 = [0(109) + 1(16) + \cdots + 17(1)]/152 = 0.954$$

and

$$\sum x_i^2/152 = [0^2(109) + 1^2(16) + \cdots + 17^2(1)]/152 = 6.704$$

Thus, the mmes for p and r in this case are

$$\hat{p} = \frac{0.954}{6.704 - (0.954)^2} = .165 \quad \hat{r} = \frac{(0.954)^2}{6.704 - (0.954)^2 - 0.954} = .188$$

Although r by definition must be positive, the denominator of $\hat{r}$ could potentially turn out negative, which would indicate that the generalized negative binomial distribution is not appropriate (or that the moment estimator is flawed). ∎

Maximum Likelihood Estimation

The method of maximum likelihood was first introduced by R. A. Fisher, a geneticist and statistician, in the 1920s. Most statisticians recommend this method, at least when the sample size is large, since the resulting estimators have certain desirable efficiency properties (see the proposition on large sample behavior toward the end of this section, as well as Section 7.4). The following example illustrates the key underlying concept.

Example 7.16 A May 2018 article on www.howtogeek.com discusses traditional criteria for "strong" passwords and the emerging advice to use longer *passphrases* by concatenating several everyday words. Suppose that 10 students at a certain university are randomly selected, and it is found that the first, third, and tenth students use passphrases for their email accounts, whereas the other seven students do not. Let $p = P(\text{passphrase})$; i.e., p is the proportion of all students at the university using a passphrase on their email accounts. Define Bernoulli random variables $X_1, X_2, \ldots, X_{10}$ by

$$X_i = \begin{cases} 1 & \text{if the } i\text{th student uses a passphrase} \\ 0 & \text{if not} \end{cases} \quad i = 1, 2, \ldots 10$$

For the obtained sample, $x_1 = x_3 = x_{10} = 1$ and the other seven x_i's are all zero. Students' decisions about whether to use passphrases are presumably independent of one another, so that the X_i's are independent and the probability of observing the obtained sample is

$$p \cdot (1 - p) \cdot p \cdot (1 - p) \cdot (1 - p) \cdots p = p^3(1 - p)^7 \tag{7.4}$$

We now ask, "For what value of p is the obtained sample *most likely* to have occurred?" That is, we wish to find the value of p that maximizes the joint pmf (7.4) or, equivalently, maximizes the natural

log of (7.4).[2] Figure 7.4a shows a graph of the *likelihood* (7.4) as a function of p. It appears that the graph reaches its peak above $p = .3$, which is the proportion of passphrases in the sample. Figure 7.4b shows a graph of the natural logarithm of (7.4), whose maximum will occur at the same value.

a

b
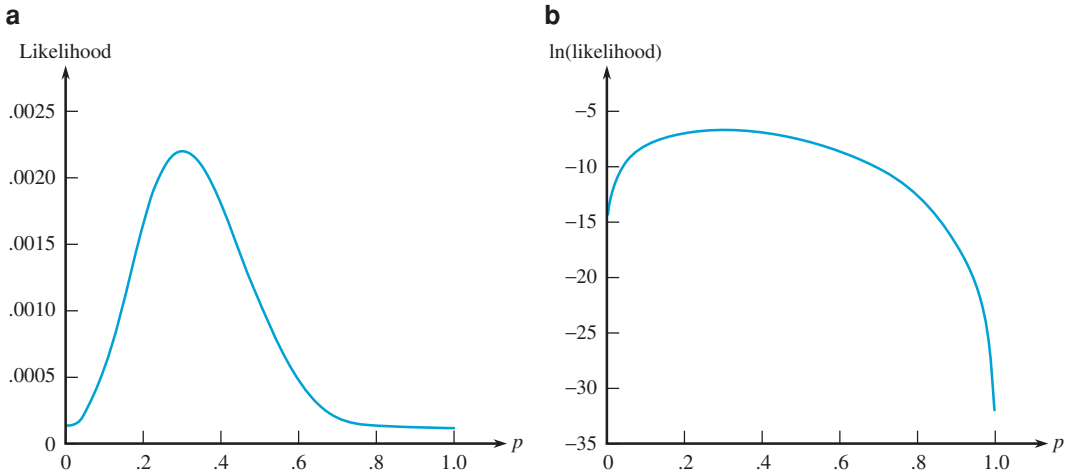
Figure 7.4 Likelihood and log likelihood plotted against p

Here,

$$\ln[p^3(1-p)^7] = 3\ln(p) + 7\ln(1-p)$$

$$\frac{d}{dp}\ln[p^3(1-p)^7] = \frac{3}{p} - \frac{7}{1-p} = 0 \Rightarrow p = \frac{3}{10}$$

So $p = 3/10 = .30$ maximizes the (log of the) probability of the specified sample, as conjectured. For that reason, the point estimate $\hat{p} = .30$ is called the *maximum likelihood estimate* of the parameter p.[3] To be clear, we could also have differentiated (7.4) directly and set that derivative equal to 0 to obtain the same result; taking the logarithm simply made the calculus easier.

Now suppose that rather than being told each individual student's decision, we had only been informed that three of the ten used passphrases. Then we would have the observed value of the binomial random variable X = the number of passphrases. The pmf of X is $\binom{10}{x}p^x(1-p)^{10-x}$; for $x = 3$, this becomes $\binom{10}{3}p^3(1-p)^7$. The binomial coefficient $\binom{10}{3}$ is irrelevant to the maximization, and so the value of p that maximizes the likelihood of observing $X = 3$ is again $\hat{p} = .30$. ∎

---

[2]Since the natural logarithm is a monotone function, finding u to maximize $\ln[g(u)]$ is equivalent to finding u to maximize $g(u)$. Taking the logarithm will frequently make differentiation easier.

[3]In general, the second derivative should be examined to make sure a maximum has been obtained, but here this is obvious from the figure.

DEFINITION	Let $X_1, \ldots, X_n$ have a joint distribution (i.e., a joint pmf or pdf) that depends on a parameter θ whose value is unknown. This joint distribution, regarded as a function of θ, is called the **likelihood function** and is denoted by $L(\theta)$. The **maximum likelihood estimate (mle)** $\hat{\theta}$ is the value of θ that maximizes the likelihood function.

Echoing the terminology from the previous section, we call $\hat{\theta}$ a maximum likelihood *estimate* if it's expressed in terms of our observed sample data and a maximum likelihood *estimator* if $\hat{\theta}$ is regarded as a function of the random variables $X_1, \ldots, X_n$.

In Example 7.16, the joint pmf of $X_1, \ldots, X_{10}$ became $p^3(1 - p)^7$ once the observed values of the X_i's were substituted. So, the likelihood function would be written $L(p) = p^3(1 - p)^7$. If we take the perspective that our data consists of a single binomial observation, then $L(p) = \binom{10}{3} p^3(1 - p)^7$. In either case, the value of p that maximizes $L(p)$ is $\hat{p} = .3$.

The likelihood function tells us how likely the observed sample is, as a function of the possible parameter value. Maximizing the likelihood gives the parameter value for which the observed sample is most likely to have been generated, that is, the parameter value that "agrees most closely" with the observed data. As in Example 7.16, maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood; the latter is typically computationally easier, since the likelihood is typically a product and so its logarithm is a sum. We will use $\ell(\theta)$ to denote the natural logarithm of the likelihood function, $\ell(\theta) = \ln[L(\theta)]$, commonly referred to as the **log-likelihood function**.

Example 7.17 Suppose $X_1, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ. Because of independence, the likelihood function is a product of the individual pdfs:

$$L(\lambda) = f(x_1, \ldots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdot \cdots \cdot (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \sum x_i}$$

Next, we determine the value of λ that maximizes the logarithm of this function:

$$\ell(\lambda) = \ln[L(\lambda)] = n \ln(\lambda) - \lambda \sum x_i$$
$$\ell'(\lambda) = \frac{n}{\lambda} - \sum x_i = 0 \Rightarrow$$
$$\lambda = \frac{n}{\sum x_i} = \frac{1}{\overline{x}}$$

Thus the mle is $\hat{\lambda} = 1/\overline{X}$. This is exactly the same as the mme that we found in Example 7.13; as noted previously, the two methods often yield the same estimator. Unfortunately, $\hat{\lambda}$ is not an unbiased estimator (see Exercise 13), since $E(1/\overline{X}) \neq 1/E(\overline{X})$. ∎

Example 7.18 In Chapter 3, we indicated that the Poisson distribution could be used for modeling the number of events of some sort that occur in a two-dimensional region (e.g., the occurrence of tornadoes in a particular Midwest county during a given time period). Assume that when the region R being sampled has area $a(R)$, the number X of events occurring in R has a Poisson distribution with mean $\lambda \cdot a(R)$, so λ represents the expected number of events per unit area, and that nonoverlapping regions yield independent X's. (This is called a *spatial Poisson process*.)

Suppose an ecologist selects n nonoverlapping regions $R_1, \ldots, R_n$ and counts the number of plants of a certain species found in each region. The joint pmf (likelihood) is then

$$L(\lambda) = p(x_1, \ldots, x_n; \lambda) = \frac{[\lambda \cdot a(R_1)]^{x_1} e^{-\lambda \cdot a(R_1)}}{x_1!} \cdot \cdots \cdot \frac{[\lambda \cdot a(R_n)]^{x_n} e^{-\lambda \cdot a(R_n)}}{x_n!}$$

$$= \frac{[a(R_1)]^{x_1} \cdot \cdots \cdot [a(R_n)]^{x_n} \cdot \lambda^{\sum x_i} \cdot e^{-\lambda \sum a(R_i)}}{x_1! \cdot \cdots \cdot x_n!} = C \cdot \lambda^{\sum x_i} \cdot e^{-\lambda \sum a(R_i)},$$

where the quantity C does not involve the parameter λ (and, hence, will not impact maximization). Then,

$$\ell(\lambda) = \ln[L(\lambda)] = \ln(C) + \ln(\lambda) \cdot \sum x_i - \lambda \sum a(R_i)$$

$$\ell'(\lambda) = 0 + \frac{\sum x_i}{\lambda} - \sum a(R_i) = 0 \Rightarrow$$

$$\lambda = \frac{\sum x_i}{\sum a(R_i)}$$

The mle is $\hat{\lambda} = \sum X_i / \sum a(R_i)$. This is intuitively reasonable because λ is the true density (plants per unit area), whereas $\hat{\lambda}$ is the sample density: $\sum X_i$ is the number of plants counted, and $\sum a(R_i)$ is just the total area sampled. Because $E(X_i) = \lambda \cdot a(R_i)$, the estimator is unbiased.

Sometimes an alternative sampling procedure is used. Instead of fixing regions to be sampled, the ecologist will select n points in the entire region of interest and let y_i = the distance from the ith point to the nearest plant. The cdf of Y = distance to the nearest plant is

$$F_Y(y) = P(Y \le y) = 1 - P(Y > y) = 1 - P\left(\begin{array}{c} \text{no plants in a} \\ \text{circle of radius } y \end{array} \right)$$

$$= 1 - \frac{e^{-\lambda \pi y^2} (\lambda \pi y^2)^0}{0!} = 1 - e^{-\lambda \pi y^2}$$

Taking the derivative of $F_Y(y)$ with respect to y yields

$$f_Y(y; \lambda) = 2\pi \lambda y e^{-\lambda \pi y^2} \quad y \ge 0$$

If we now form the likelihood $L(\lambda) = f_Y(y_1; \lambda) \cdot \cdots \cdot f_Y(y_n; \lambda)$, differentiate $\ln[L(\lambda)]$, and so on, the resulting mle is

$$\hat{\lambda} = \frac{n}{\pi \sum Y_i^2} = \frac{\text{number of plants observed}}{\text{total area sampled}}$$

which is also a sample plant density. It can be shown that in a sparse environment (small λ), the distance method is in a certain sense better, whereas in a dense environment, the first sampling method is better. ∎

The definition of mles can be extended in the natural way to distributional families that include two or more parameters. The mles of parameters $\theta_1, \ldots, \theta_m$ are those values $\hat{\theta}_1, \ldots, \hat{\theta}_m$ that maximize the likelihood function $L(\theta_1, \ldots, \theta_m)$ or, equivalently, the logarithm of the likelihood function.

Example 7.19 Let $X_1, \ldots, X_n$ be a random sample from a normal distribution, which includes the two parameters μ and σ. The likelihood function is

$$L(\mu, \sigma) = f(x_1, \ldots, x_n; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)}$$

$$= \left(2\pi\sigma^2\right)^{-n/2} e^{-\sum (x_i-\mu)^2/(2\sigma^2)}$$

so

$$\ell(\mu, \sigma) = \ln[L(\mu, \sigma)] = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum (x_i - \mu)^2$$

To find the maximizing values of μ and σ, we must take the *partial* derivatives of $\ell(\mu, \sigma)$ with respect to both μ and σ, equate them to zero, and solve the resulting two equations. Omitting the details, the resulting mles are

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma} = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n}}$$

Notice that the mle of σ is *not* the sample standard deviation, S, since the denominator in the mle is n and not $n-1$. ∎

Example 7.20 Let $X_1, \ldots, X_n$ be a random sample from a Weibull pdf

$$f(x; \alpha, \beta) = \frac{\alpha}{\beta^\alpha} \cdot x^{\alpha-1} \cdot e^{-(x/\beta)^\alpha} \quad x \geq 0$$

Writing the likelihood $L(\alpha, \beta)$ and log-likelihood $\ell(\alpha, \beta)$, then setting both $\partial\ell/\partial\alpha = 0$ and $\partial\ell/\partial\beta = 0$ yields the equations

$$\alpha = \left[\frac{\sum [x_i^\alpha \cdot \ln(x_i)]}{\sum x_i^\alpha} - \frac{\sum \ln(x_i)}{n}\right]^{-1} \quad \beta = \left(\frac{\sum x_i^\alpha}{n}\right)^{1/\alpha}$$

These two equations cannot be solved explicitly to give general formulas for the mles $\hat{\alpha}$ and $\hat{\beta}$. Instead, for any sample $x_1, \ldots, x_n$, the equations must be solved using an iterative numerical procedure.

The iterative mle computations can be done using statistical software. In R, the command `fitdistr(x, ``weibull'')` will return $\hat{\alpha}$ and $\hat{\beta}$ assuming the data is stored in the vector `x` (the MASS package must be installed first). As an example, consider the following data on the survival time (weeks) of male mice subjected to 240 rads of gamma radiation (from A. J. Gross and V. Clark, *Survival Distributions: Reliability Applications in the Biomedical Sciences*):

152	115	109	94	88	137	152	77	160	165
125	40	128	123	136	101	62	153	83	69

A Weibull probability plot supports the plausibility of assuming that survival time has a Weibull distribution. With the aid of software, maximum likelihood estimates of the Weibull parameters are

$\hat{\alpha} = 3.799$ and $\hat{\beta} = 125.88$. Figure 7.5 shows the Weibull log likelihood as a function of both α and β. The surface near the top has a rounded shape, allowing the maximum to be found easily, but for some distributions the surface can be much more irregular, and the maximum may be hard to find.
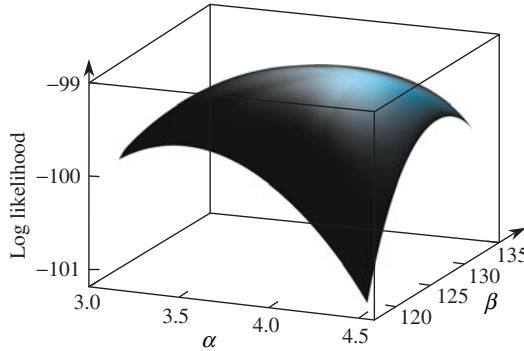
Figure 7.5   Weibull log likelihood for Example 7.20   ∎

Some Properties of MLEs

In Example 7.19, we obtained the mle of σ when the underlying distribution is normal. The mle of σ^2, as well as many other mles, can be easily derived using the following proposition.

MLE INVARIANCE PRINCIPLE	Let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ be the mles of the parameters $\theta_1, \theta_2, \ldots, \theta_m$. Then the mle of any function $h(\theta_1, \theta_2, \ldots, \theta_m)$ of these parameters is the function $h(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ of the mles.

For an intuitive idea of the proof, consider the special case $m = 1$, with $\theta_1 = \theta$, and assume that $h(\cdot)$ is a one-to-one function. On the graph of the likelihood as a function of the parameter θ, the highest point occurs where $\theta = \hat{\theta}$. Now consider the graph of the likelihood as a function of $h(\theta)$. In the new graph the same heights occur, but the height that was previously plotted at $\theta = a$ is now plotted at $h(\theta) = h(a)$, and the highest point is now plotted at $h(\theta) = h(\hat{\theta})$. Thus, the maximum remains the same, but it now occurs at $h(\hat{\theta})$.

Example 7.21   (Example 7.19 continued) In the case of a random sample from a normal distribution, the mles of μ and σ are $\hat{\mu} = \overline{X}$ and $\hat{\sigma} = \sqrt{\sum (X_i - \overline{X})^2 / n}$. To obtain the mle of the function $h(\mu, \sigma) = \sigma^2$, substitute the mles into the function:

$$\widehat{\sigma^2} = \hat{\sigma}^2 = \frac{1}{n} \sum (X_i - \overline{X})^2$$

The mle of σ^2 is not the unbiased estimator (the sample variance S^2 is), although they are close when n is large. Similarly, the mle of the *population coefficient of variation*, defined by $h(\mu, \sigma) = 100\mu/\sigma$, is simply $100\hat{\mu}/\hat{\sigma}$.   ∎

Example 7.22 (Example 7.20 continued) From Section 4.5, the mean value of a Weibull rv X is

$$\mu = \beta \cdot \Gamma(1 + 1/\alpha)$$

The mle of μ is therefore $\hat{\mu} = \hat{\beta} \cdot \Gamma(1 + 1/\hat{\alpha})$, where $\hat{\alpha}$ and $\hat{\beta}$ are the mles of α and β. In particular, the mle of μ in this case is not the mme $\overline{X}$, although the latter is an unbiased estimator. At least for large n, $\hat{\mu}$ is a better estimator than $\overline{X}$ because the mle has lower mean squared error. ∎

The method of maximum likelihood estimation has considerable intuitive appeal. The following proposition provides additional rationale for the use of mles; see Section 7.4 for more details.

THEOREM   Under very general conditions on the joint distribution of the sample, when the sample size is large, the maximum likelihood estimator of any parameter θ (1) is close to θ (consistency), (2) is approximately unbiased, and (3) has variance that is nearly as small as can be achieved by any unbiased estimator. Stated another way, the mle $\hat{\theta}$ is at least approximately the MVUE of θ.

Because of this result and the fact that calculus-based techniques can usually be used to derive the mles (although numerical methods, such as Newton–Raphson, are sometimes necessary), maximum likelihood estimation is the most widely used estimation technique among statisticians. Many of the estimators used in the rest of this book are mles.

One consequence of the preceding theorem is that when the mle and the moments estimator differ for a given distribution, the mle will nearly always have smaller variance. Thus, although formulas for mmes are often easier to determine, the extra computation required for mles is typically worth the price.

Some Complications
Sometimes calculus cannot be used to obtain mles.

Example 7.23 Suppose the waiting time for a bus is uniformly distributed on $[0, \theta]$ and the results $x_1, \ldots, x_n$ of a random sample from this distribution have been observed. Since $f(x; \theta) = 1/\theta$ for $0 \le x \le \theta$ and 0 otherwise,

$$L(\theta) = f(x_1, \ldots, x_n; \theta) = \begin{cases} 1/\theta^n & 0 \le x_1 \le \theta, \ldots, 0 \le x_n \le \theta \\ 0 & \text{otherwise} \end{cases}$$

As long as $\theta \ge \max(x_i)$, $L(\theta) = 1/\theta^n > 0$, but for $\theta < \max(x_i)$, the likelihood drops to 0. This is illustrated in Figure 7.6. Calculus will not work because the maximum of the likelihood occurs at a point of discontinuity, but the figure shows that the mle is $\hat{\theta} = \max(x_i)$. Thus if my waiting times are 2.3, 3.7, 1.5, 0.4, and 3.2, then the mle is $\hat{\theta} = 3.7$. Note that this mle is biased (see Example 7.6).
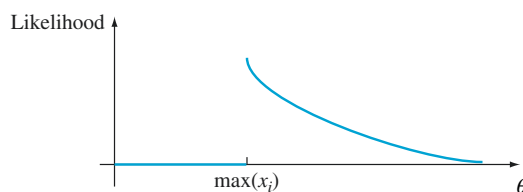
Figure 7.6 The likelihood function for Example 7.23 ∎

Example 7.24 A method often used to estimate the size of a wildlife population involves performing a *capture/recapture* experiment. In this experiment, an initial sample of M animals is captured, each of these animals is tagged, and the animals are then returned to the population. After allowing enough time for the tagged individuals to mix into the population, another sample of size n is captured. With X = the number of tagged animals in the second sample, the objective is to use the observed x to estimate the population size N.

The parameter of interest is $\theta = N$, which can assume only integer values, so even after determining the likelihood function (the pmf of X here), using calculus to obtain N would present difficulties. If we think of a "success" as a previously tagged animal being recaptured, then the sampling is without replacement from a population containing M successes and $N - M$ failures, so that X is a hypergeometric rv and the likelihood function is

$$L(N) = h(x; n, M, N) = \frac{\binom{M}{x} \cdot \binom{N - M}{n - x}}{\binom{N}{n}}$$

The integer-valued nature of N notwithstanding, it would be difficult to take the derivative of $L(N)$. However, let's consider the ratio of $L(N)$ to $L(N - 1)$:

$$\frac{L(N)}{L(N - 1)} = \cdots = \frac{(N - M) \cdot (N - n)}{N(N - M - n + x)}$$

This ratio is larger than 1 if and only if $N < Mn/x$. The value of N for which $L(N)$ is maximized is therefore the largest integer less than Mn/x. If we use standard mathematical notation $[r]$ for the greatest integer less than or equal to r, the mle of N is $\hat{N} = [Mn/x]$. As an illustration, if $M = 200$ fish are taken from a lake and tagged, subsequently $n = 100$ fish are recaptured, and among the 100 there are $x = 11$ tagged fish, then $\hat{N} = [(200)(100)/11] = [1818.18] = 1818$.

The estimate is actually rather intuitive; x/n is the proportion of the recaptured sample that is tagged, whereas M/N is the proportion of the entire population that is tagged. The estimate is obtained by equating these two proportions (estimating a population proportion by a sample proportion). ∎

Obtaining an mle requires that the underlying distribution be specified. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from *some* pdf $f(x; \theta)$ that is symmetric about θ, but the investigator is unsure of the form of the f function. It is then desirable to use an estimator that is *robust*, that is, one that performs well for a wide variety of underlying pdfs. One such estimator, called an *M-estimator*, is based on a generalization of maximum likelihood estimation. Instead of maximizing the log-likelihood $\sum \ln[f(x; \theta)]$ for a specified f, one maximizes $\sum \psi[f(x; \theta)]$, where the "objective function" ψ is selected to yield an estimator with good robustness properties. The book by David Hoaglin et al. (see the bibliography) contains a good exposition on this subject.

Exercises: Section 7.2 (25–37)

25. A random sample of n bike helmets manufactured by a company is selected. Let X = the number among the n that are flawed, and let $p = P(\text{flawed})$. Assume that only X is observed, rather than the sequence of S's and F's.

  a. Derive the maximum likelihood estimator of p. If $n = 20$ and $x = 3$, what is the estimate?
  b. Is the estimator of part (a) unbiased?
  c. If $n = 20$ and $x = 3$, what is the mle of the probability $\theta = (1 - p)^5$ that none of the next five helmets examined is flawed?

26. Let X have a Weibull distribution with parameters α and β, so

$$E(X) = \beta \cdot \Gamma(1 + 1/\alpha)$$
$$V(X) = \beta^2 \{\Gamma(1 + 2/\alpha) - [\Gamma(1 + 1/\alpha)]^2\}$$

   a. Based on a random sample $X_1, \ldots, X_n$, write equations for the method of moments estimators of β and α. Show that, once the estimate of α has been obtained, the estimate of β can be found using the gamma function and that the estimate of α is the solution to a complicated equation involving the gamma function.

   b. If $n = 20$, $\bar{x} = 28.0$, and $\sum x_i^2 = 16{,}500$, compute the estimates. [*Hint*: $[\Gamma(1.2)]^2/\Gamma(1.4) = .95$.]

27. Let X denote the proportion of allotted time that a randomly selected student spends working on a certain aptitude test. Suppose the pdf of X is

$$f(x; \theta) = (\theta + 1)x^\theta \quad 0 \leq x \leq 1$$

for some $\theta > -1$. A random sample of ten students yields data $x_1 = .92$, $x_2 = .79$, $x_3 = .90$, $x_4 = .65$, $x_5 = .86$, $x_6 = .47$, $x_7 = .73$, $x_8 = .97$, $x_9 = .94$, $x_{10} = .77$.

   a. Use the method of moments to obtain an estimator of θ, and then compute the estimate for this data.

   b. Obtain the maximum likelihood estimator of θ, and then compute the estimate for the given data.

28. Two different computer systems are monitored for a total of n weeks. Let X_i denote the number of breakdowns of the first system during the ith week, and suppose the X_i's are independent and drawn from a Poisson distribution with parameter μ_1. Similarly, let Y_i denote the number of breakdowns of the second system during the ith week, and assume independence with each Y_i Poisson with parameter μ_2. Derive the mles of μ_1, μ_2, and $\mu_1 - \mu_2$. [*Hint*: Using independence, write the joint pmf (likelihood) of the X_i's and Y_i's together.]

29. Refer to Exercise 25. Instead of selecting $n = 20$ helmets to examine, suppose we examine helmets in succession until we have found $r = 3$ flawed ones. If the 20th helmet is the third flawed one, what is the mle of p? Is this the same as the estimate in Exercise 25? Why or why not? Is it the same as the estimate computed from the unbiased estimator of Exercise 19?

30. Six Pepperidge Farm bagels were weighed, yielding the following data (grams):

   117.6   109.5   111.6   109.2   119.1   110.8

   a. Assuming that the six bagels are a random sample and the weight is normally distributed, estimate the true average weight and standard deviation of the weight using maximum likelihood.

   b. Again assuming a normal distribution, estimate the weight below which 95% of all bagels will have their weights. [*Hint*: What is the 95th percentile in terms of μ and σ? Now use the invariance principle.]

   c. Suppose we choose another bagel and weigh it. Let $X =$ weight of the bagel. Use the given data to obtain the mle of the probability $P(X \leq 113.4)$. [*Hint*: $P(X \leq 113.4) = \Phi[(113.4 - \mu)/\sigma)]$.]

31. Suppose a measurement is made on some physical characteristic whose value is known, and let X denote the resulting measurement error. It is often reasonable to assume that $E(X) = 0$ and that X has a normal distribution. Thus, the pdf of any particular measurement error is

$$f(x; \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-x^2/2\theta}$$

where θ denotes the population variance. Let $X_1, \ldots X_n$ be a random sample of such measurement errors.

   a. Determine the likelihood function of θ.

b. Obtain and simplify the log-likelihood function.

c. Differentiate the log-likelihood function to determine the mle of θ.

d. The *precision* of a normal distribution is defined as $\tau = 1/\theta$. Find the mle of τ.

32. Let $X_1, \ldots, X_n$ be a random sample from a gamma distribution with parameters α and β.

a. Derive the equations whose solution yields the maximum likelihood estimators of α and β. Does it appear that they can be solved explicitly?

b. Show that the mle of $\mu = \alpha\beta$ is $\hat{\mu} = \overline{X}$.

33. Let $X_1, X_2, \ldots, X_n$ represent a random sample from the Rayleigh distribution with density function given in Exercise 17.

a. Determine the maximum likelihood estimator of θ and then calculate the estimate for the blood plasma beta concentration data given in that exercise. Is this estimator the same as the unbiased estimator suggested in Exercise 17?

b. Determine the mle of the *median* of the blood plasma beta concentration distribution. [*Hint*: First express the median of the Rayleigh distribution in terms of θ.]

34. Consider a random sample $X_1, X_2, \ldots, X_n$ from the shifted exponential pdf

$$f(x; \lambda, \theta) = \lambda e^{-\lambda(x-\theta)} \quad x \geq \theta$$

Taking $\theta = 0$ gives the pdf of the exponential distribution considered previously (with positive density to the right of zero). An example of the shifted exponential distribution appeared in Example 4.5, in which the variable of interest was hazardous flood rate and θ was the lowest water flow rate considered hazardous.

a. Obtain the maximum likelihood estimators of θ and λ.

b. If $n = 10$ hazardous flood rate observations are made, resulting in the values 13.11, 10.64, 12.55, 12.20, 15.44,

13.42, 20.39, 18.93, 27.82, and 11.30, calculate the maximum likelihood estimates of θ and λ.

35. Twenty identical components are put on test. The lifetime distribution of each is exponential with parameter λ. The experimenter then leaves the test facility unmonitored. On her return 24 h later, the experimenter immediately terminates the test after noticing that $y = 15$ of the 20 components are still in operation (so 5 have failed). Derive the mle of λ. [*Hint*: Let $Y =$ the number that survive 24 h. Then $Y \sim \text{Bin}(n, p)$. What is the mle of p? Now notice that $p = P(X_i \geq 24)$, where X_i is exponentially distributed. This relates λ to p, so the former can be estimated once the latter has been.]

36. The article "A Model of Pedestrians' Waiting Times for Street Crossings at Signalized Intersections" (*Transp. Res.* 2013: 17–28) suggested that under some circumstances the distribution of waiting time X could be modeled with the following pdf:

$$f(x; \theta, \tau) = \frac{\theta}{\tau}(1 - x/\tau)^{\theta-1} \quad 0 \leq x < \tau$$

where $\theta > 0$ and $\tau > 0$.

a. Suppose we observe a random sample of waiting times $X_1, \ldots, X_n$, and suppose that the value of the parameter τ is known. Find the mle of θ.

b. Suppose instead that θ is known but τ is unknown. Determine an equation whose solution is the mle of τ.

37. Let $X_1, \ldots, X_n$ be a random sample from the Laplace distribution (also called the double exponential distribution) with pdf $f(x; \theta) = e^{-|x-\theta|}$ for $-\infty < x < \infty$.

a. Determine the method of moments estimator for θ.

b. Determine the maximum likelihood estimator for θ. [*Hint*: It can be shown that the expression $\sum |x_i - c|$ is minimized by $c = \tilde{x}$, the median of the x_i's.]

7.3   Sufficiency

An investigator who wishes to make an inference about some parameter θ will base conclusions on the value of one or more statistics—the sample mean $\overline{X}$, the sample standard deviation S, the sample range $Y_n - Y_1$, and so on. Intuitively, some statistics will contain more information about θ than will others. *Sufficiency*, the topic of this section, will help us decide which functions of the data are most informative for making inferences.

As a first point, we note that a statistic $T = t(X_1, \ldots, X_n)$ will not be useful for drawing conclusions about θ unless the distribution of T depends on θ. Consider, for example, a random sample of size $n = 2$ from a normal distribution with mean μ and variance σ^2, and let $T = X_1 - X_2$. Then T has a normal distribution with mean 0 and variance $2\sigma^2$, which does not depend on μ. Thus this statistic cannot be used as a basis for drawing any conclusions about μ, although it certainly does carry information about the variance σ^2.

The relevance of this observation to sufficiency is as follows. Suppose an investigator is given the value of some statistic T, and then examines the *conditional* distribution of the sample $X_1, \ldots, X_n$ given the value of the statistic—for example, the conditional distribution given that $T = \overline{X} = 28.7$. If this conditional distribution does not depend upon θ, then it can be concluded that there is no *additional* information about θ in the sample over and above what is provided by T. In this sense, for purposes of making inferences about θ, it is *sufficient* to know the value of T, which contains all information in the data relevant to θ.

Example 7.25   An investigation of major defects on new vehicles of a certain type involved selecting an initial random sample of $n = 3$ vehicles and determining for each one the value of X = the number of major defects. This resulted in observations $x_1 = 1$, $x_2 = 0$, and $x_3 = 3$. You, as a consulting statistician, have been provided with a description of the experiment, from which it is reasonable to assume that X has a Poisson distribution, but you have been told only that the total number of defects T for the three sampled vehicles was 4.

Knowing that $T = \sum X_i = 4$, would there be any additional advantage in having the observed values of the individual X_i's when making an inference about the Poisson parameter μ? Or, is it instead the case that the statistic T contains all relevant information about μ in the data? To address this issue, consider the conditional distribution of (X_1, X_2, X_3) given that $\sum X_i = 4$. First of all, there are only a few possible (x_1, x_2, x_3) triples for which $x_1 + x_2 + x_3 = 4$. For example, $(0, 4, 0)$ is a possibility, as are $(2, 2, 0)$ and $(1, 0, 3)$, but not $(1, 2, 3)$ or $(5, 0, 2)$. That is,

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | T = 4) = 0 \quad \text{unless } x_1 + x_2 + x_3 = 4$$

Now consider the triple $(2, 1, 1)$, which is consistent with $T = 4$. A moment generating function argument shows that T has a Poisson distribution with parameter 3μ. From this we calculate the conditional probability that $(X_1, X_2, X_3) = (2, 1, 1)$, given $T = \sum X_i = 4$, as follows:

$$
\begin{aligned}
P(X_1 = 2, X_2 = 1, X_3 = 1 | T = 4) &= \frac{P(X_1 = 2, X_2 = 1, X_3 = 1 \cap T = 4)}{P(T = 4)} \\
&= \frac{P(X_1 = 2, X_2 = 1, X_3 = 1)}{P(T = 4)} \\
&= \frac{\frac{e^{-\mu}\mu^2}{2!} \cdot \frac{e^{-\mu}\mu^1}{1!} \cdot \frac{e^{-\mu}\mu^1}{1!}}{\frac{e^{-3\mu}\mu^4}{4!}} = \frac{4}{27}
\end{aligned}
$$

The particular probability isn't important; what's critical is that this conditional probability *does not depend on the unknown parameter* μ. The same holds true for every other triple that sums to 4, indicating that the conditional distribution of (X_1, X_2, X_3) given T does not involve μ. Thus once the value of the statistic $T = \sum X_i$ has been provided, there is no additional information about μ in the individual observations.

To put this another way, think of obtaining the data from the experiment in two stages:

1. Observe the value of $T = X_1 + X_2 + X_3$ from a Poisson distribution with parameter 3μ.
2. Having observed $T = 4$, now obtain the individual x_i's from the conditional distribution

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | T = 4)$$

Since the conditional distribution in step 2 does not involve μ, there is no additional information about μ resulting from the second stage of the data generation process. This argument holds more generally for any sample size n and values of t other than 4 (e.g., the total number of defects among 10 randomly selected vehicles might be $\sum X_i = 16$). Once the value of $\sum X_i$ is known, there is no further information in the data about the Poisson parameter; it is "sufficient" to be told the total. ∎

DEFINITION   A statistic $T = t(X_1, \ldots, X_n)$ is said to be **sufficient** for making inferences about a parameter θ if the joint distribution of $X_1, \ldots, X_n$ given that $T = t$ does not depend upon θ, for every possible value t of the statistic T.

The notion of sufficiency formalizes the idea that a statistic T contains all relevant information about θ. Once the value of T for the given data is available, it is of no benefit to know anything else about the sample.

The Factorization Theorem

How can a sufficient statistic be identified? It may seem as though one would have to select a statistic, determine the conditional distribution of the X_i's given any particular value of the statistic (no easy task—look at the last example!), and keep doing this until hitting paydirt by finding one that satisfies the defining condition. This would be terribly time-consuming, and when the X_i's are continuous there are additional technical difficulties in obtaining the relevant conditional distribution. Fortunately, the next result provides a relatively straightforward way of proceeding.

THE NEYMAN FACTORIZATION THEOREM   Let $f(x_1, \ldots, x_n; \theta)$ denote the joint pmf or pdf of $X_1, \ldots, X_n$. Then $T = t(X_1, \ldots, X_n)$ is a sufficient statistic for θ if and only if there exist functions g and h such that

$$f(x_1, \ldots, x_n; \theta) = g(t(x_1, \ldots, x_n); \theta) \cdot h(x_1, \ldots, x_n)$$

That is, the joint pmf or pdf can be represented as a product of two factors, in which one factor includes θ and involves the data only through $t(x_1, \ldots, x_n)$ while the other factor does not depend on θ.

Before sketching a proof of this theorem, we consider several examples.

Example 7.26 Let's generalize the previous example by considering a random sample $X_1, \ldots, X_n$ from a Poisson distribution with parameter μ, for example, the numbers of blemishes on n independently selected iPhone cases or the numbers of errors in n batches of tax returns where each batch consists of many returns. The joint pmf of these variables is

$$f(x_1, \ldots, x_n; \mu) = \frac{e^{-\mu}\mu^{x_1}}{x_1!} \cdot \cdots \cdot \frac{e^{-\mu}\mu^{x_n}}{x_n!} = \left(e^{-n\mu}\mu^{\Sigma x_i}\right)\left(\frac{1}{x_1! \cdot \cdots \cdot x_n!}\right)$$

The factor inside the first set of parentheses includes the parameter μ and involves the data only through $\sum X_i$, whereas the factor inside the second set of parentheses does not depend on μ. So we have the desired factorization, and by the factorization theorem the sufficient statistic for μ is $T = \sum X_i$, as we ascertained in Example 7.25 directly from the definition of sufficiency.   ∎

A sufficient statistic is *not* unique: any one-to-one function of a sufficient statistic is itself sufficient. In the Poisson example, the sample mean $\overline{X} = (1/n)\sum X_i$ is a one-to-one function of $\sum X_i$ (knowing the value of the sum of the n observations is equivalent to knowing their mean), so the sample mean is also a sufficient statistic.

Example 7.27 Suppose that the waiting time for a bus on a weekday morning is uniformly distributed on the interval from 0 to θ, and consider a random sample $X_1, \ldots, X_n$ of waiting times (i.e., times on n independently selected mornings). The joint pdf of these times is

$$f(x_1, \ldots, x_n; \theta) = \frac{1}{\theta} \cdot \frac{1}{\theta} \cdot \cdots \cdot \frac{1}{\theta} = \frac{1}{\theta^n} \quad 0 \le x_1 \le \theta, \ldots, 0 \le x_n \le \theta$$

To obtain the desired factorization, we introduce notation for an **indicator function**: $I(A) = 1$ if the statement A is true, and $I(A) = 0$ otherwise. For instance, we may write the joint pdf of the wait times more formally as

$$f(x_1, \ldots, x_n; \theta) = \frac{1}{\theta^n}I(0 \le x_1 \le \theta, \ldots, 0 \le x_n \le \theta)$$

The statement A is that all x_i's are between 0 and θ. But the x_i's will all be between 0 and θ if and only if (1) the smallest of the x_i's is at least 0 and (2) the largest is at most θ. Thus, the joint pdf can be reexpressed as

$$f(x_1, \ldots, x_n; \theta) = \frac{1}{\theta^n}I(0 \le \min(x_1, \ldots, x_n) \text{ and } \max(x_1, \ldots, x_n) \le \theta)$$

$$= \left[\frac{1}{\theta^n}I(\max(x_1, \ldots, x_n) \le \theta)\right] \cdot I(0 \le \min(x_1, \ldots, x_n))$$

The factor inside the square brackets includes θ and involves the x_i's only through the function $t(x_1, \ldots, x_n) = \max(x_1, \ldots, x_n)$. Voilà, we have our desired factorization, and the sufficient statistic for the uniform parameter θ is $T = \max(X_1, \ldots, X_n)$. All the information about θ in this uniform random sample is contained in the largest of the n observations; knowing the values of the other $n-1$ observations provides no further information toward estimating θ.   ∎

Proof of the Factorization Theorem A general proof when the X_i's constitute a random sample from a continuous distribution is fraught with technical details that are beyond the level of our text. So

we content ourselves with a proof in the discrete case. For the sake of concise notation, denote X_1, X_2, ..., X_n by $\boldsymbol{X}$ and x_1, x_2, ..., x_n by $\boldsymbol{x}$.

Suppose first that $T = t(\boldsymbol{x})$ is sufficient, so that $P(\boldsymbol{X} = \boldsymbol{x} \mid T = t)$ does not depend upon θ. Focus on a value t for which $t(\boldsymbol{x}) = t$ (e.g., $\boldsymbol{x} = (3, 0, 1)$ and $t(\boldsymbol{x}) = \sum x_i$, so $t = 4$). The event that $\boldsymbol{X} = \boldsymbol{x}$ is then identical to the event that both $\boldsymbol{X} = \boldsymbol{x}$ and $T = t$ because the first equality implies the second one. Thus

$$
\begin{aligned}
f(\boldsymbol{x}; \theta) = P(\boldsymbol{X} = \boldsymbol{x}; \theta) &= P(\boldsymbol{X} = \boldsymbol{x} \cap T = t; \theta) \\
&= P(\boldsymbol{X} = \boldsymbol{x} | T = t; \theta) \cdot P(T = t; \theta) = P(\boldsymbol{X} = \boldsymbol{x} | T = t) \cdot P(T = t; \theta)
\end{aligned}
$$

Since the first factor in the last product does not involve θ and the other involves the data only through t, we have our desired factorization.

Now let's go the other way: assume a factorization, and show that T is sufficient, i.e., that the conditional probability that $\boldsymbol{X} = \boldsymbol{x}$ given that $T = t$ does not involve θ.

$$
\begin{aligned}
P(\boldsymbol{X} = \boldsymbol{x} | T = t; \theta) &= \frac{P(\boldsymbol{X} = \boldsymbol{x} \cap T = t; \theta)}{P(T = t; \theta)} = \frac{P(\boldsymbol{X} = \boldsymbol{x}; \theta)}{P(T = t; \theta)} = \frac{g(t; \theta) h(\boldsymbol{x})}{\sum_{\boldsymbol{u}:t(\boldsymbol{u})=t} P(\boldsymbol{X} = \boldsymbol{u}; \theta)} \\
&= \frac{g(t; \theta) h(\boldsymbol{x})}{\sum_{\boldsymbol{u}:t(\boldsymbol{u})=t} g(t(\boldsymbol{u}); \theta) \cdot h(\boldsymbol{u})} = \frac{g(t; \theta) h(\boldsymbol{x})}{\sum_{\boldsymbol{u}:t(\boldsymbol{u})=t} g(t; \theta) \cdot h(\boldsymbol{u})} = \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{u}:t(\boldsymbol{u})=t} h(\boldsymbol{u})}
\end{aligned}
$$

Sure enough, this final ratio does not involve θ. ∎

Jointly Sufficient Statistics

When the joint pmf or pdf of the data involves a single unknown parameter θ, there is frequently a single statistic (single function of the data) that is sufficient. However, when there are several unknown parameters—for example, the mean μ and standard deviation σ of a normal distribution, or the shape parameter α and scale parameter β of a gamma distribution—we must expand our notion of sufficiency.

DEFINITION
Suppose the joint distribution of $X_1, \ldots, X_n$ involves m unknown parameters θ_1, θ_2, ..., θ_m. The k statistics $T_1 = t_1(X_1, \ldots, X_n), \ldots, T_k = t_k(X_1, \ldots, X_n)$ are said to be **jointly sufficient** for the parameters if the conditional distribution of the X_i's given that $T_1 = t_1, \ldots, T_k = t_k$ does not depend on any of the unknown parameters, and this is true for all possible values $t_1, t_2, \ldots, t_k$ of the statistics.

Example 7.28 Consider a random sample X_1, X_2, X_3 of size $n = 3$ from any continuous distribution, and let $T_1 < T_2 < T_3$ be their ordered values (these were denoted $Y_1 < Y_2 < Y_3$ in Section 5.7.) Then given, for example, that the three ordered values are $21.4 < 23.8 < 26.0$, the original X_i's are equally likely to be any one of the $3! = 6$ permutations of these numbers: $(23.8, 21.4, 26.0)$, $(26.0, 23.8, 21.4)$ and so on. More formally, for any values t_1, t_2, and t_3 satisfying $t_1 < t_2 < t_3$,

$$
\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | T_1 = t_1, T_2 = t_2, T_3 = t_3) \\
&= \begin{cases} 1/3! & (x_1, x_2, x_3) = (t_1, t_2, t_3), \ldots, (t_3, t_2, t_1) \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

This conditional distribution clearly does not involve any unknown parameters. Generalizing this argument to a sample of size n, we see that for a random sample from a continuous distribution, the n ordered values are jointly sufficient for $\theta_1, \theta_2, \ldots, \theta_m$ regardless of whether $m = 1$ (e.g., the exponential distribution has a single parameter) or 2 (the normal distribution) or even $m > 2$. ∎

The factorization theorem extends to the case of jointly sufficient statistics: $T_1, \ldots, T_k$ are jointly sufficient for $\theta_1, \ldots, \theta_m$ if and only if the joint pmf or pdf of the X_i's can be represented as a product of two factors, where the first includes the θ_i's and involves the data only through $t_1, \ldots, t_k$ and the second does not involve the θ_i's.

Example 7.29 Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma)$ distribution. The joint pdf is

$$f(x_1, \ldots, x_n; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-(x_i - \mu)^2/(2\sigma^2)} = \left[\frac{1}{\sigma^n} \cdot e^{-\left(\Sigma x_i^2 - 2\mu\Sigma x_i + n\mu^2\right)/(2\sigma^2)}\right] \cdot \left(\frac{1}{2\pi}\right)^{n/2}$$

This factorization shows that the two statistics $\sum X_i$ and $\sum X_i^2$ are jointly sufficient for the two parameters μ and σ. Since $\sum (X_i - \overline{X})^2 = \sum X_i^2 - n(\overline{X})^2$, there is a one-to-one correspondence between the two sufficient statistics and the statistics $\overline{X}$ and $\sum (X_i - \overline{X})^2$; that is, values of the two original sufficient statistics uniquely determine values of the latter two statistics, and vice versa. This implies that the latter two statistics are also jointly sufficient, which in turn implies that the sample mean and sample standard deviation are jointly sufficient statistics. The sample mean and sample standard deviation (or sample variance) encapsulate all the information about μ and σ that is contained in the sample data. ∎

Minimal Sufficiency

When $X_1, \ldots, X_n$ constitute a random sample from a normal distribution, the n ordered values are jointly sufficient for μ and σ (see Example 7.28), and the sample mean and sample sd are also jointly sufficient (as shown in Example 7.29). Both the ordered values and the pair $(\overline{X}, S)$ reduce the data without any information loss, but the sample mean and variance represent a greater reduction. In general, we would like the greatest possible reduction without information loss. A **minimal** (possibly jointly) **sufficient statistic** is a function of every other sufficient statistic. That is, given the value(s) of any other sufficient statistic(s), the value(s) of the minimal sufficient statistic(s) can be calculated. A minimal sufficient statistic is the sufficient statistic having the smallest dimensionality, and thus represents the greatest possible reduction of the data without any information loss.

A general discussion of minimal sufficiency is beyond the scope of our text. In the case of a normal distribution with values of both μ and σ unknown, it can be shown that the sample mean and sample sd are jointly minimal sufficient (so the same is true of $\sum X_i$ and $\sum X_i^2$). It is intuitively reasonable that because there are two unknown parameters, there should be a pair of sufficient statistics. It is indeed often the case that the number of minimal sufficient statistic(s) matches the number of unknown parameters. But this is not always true. Consider a random sample $X_1, \ldots, X_n$ from the pdf $f(x; \theta) = 1/\{\pi[1 + (x - \theta)]^2\}$, i.e., from a Cauchy distribution with location parameter θ. Because the Cauchy distribution is continuous, the n ordered values are jointly sufficient for θ. It would seem, though, that a single sufficient statistic (one-dimensional) could be found for the single parameter θ. Unfortunately this is not the case: it can be shown that the ordered values are *minimal* sufficient! So going beyond the ordered values to any single function of the X_i's as a point estimator of θ entails a loss of information from the original data.

Improving an Estimator

Because a sufficient statistic contains all the information the data has to offer about the value of θ, it is reasonable that an estimator of θ, or any function of θ, should depend on the data only through the sufficient statistic. A general result due to C. R. Rao and David Blackwell shows how to start with an unbiased statistic that is not a function of sufficient statistics and create an improved estimator that is both unbiased and sufficient.

RAO–BLACKWELL THEOREM

Suppose that the joint distribution of $X_1, \ldots, X_n$ depends on some unknown parameter θ and that T is sufficient for θ. Consider estimating $h(\theta)$, a specified function of θ. If U is any unbiased estimator for estimating $h(\theta)$, then the estimator $U^* = E(U|T)$ is also unbiased for $h(\theta)$ and has variance no greater than the original unbiased estimator U.

Proof First of all, we must show that U^* is indeed an estimator—i.e., that it is a function of the X_i's and not of θ. This follows because, given that T is sufficient, the distribution of U conditional on T does not involve θ, so the expected value $E(U|T)$ will of course not involve θ. Second, the fact that U and U^* have the same expected value (i.e., they are both unbiased estimators of $h(\theta)$) follows from the Law of Total Expectation introduced in Section 5.4:

$$E(U^*) = E[E(U|T)] = E(U) = h(\theta)$$

Finally, the fact that U^* has smaller variance than U is a consequence of Law of Total Variance:

$$V(U) = V[E(U|T)] + E[V(U|T)] = V(U^*) + E[V(U|T)]$$

Because $V(U|T)$, being a variance, is positive, it follows that $V(U) \geq V(U^*)$ as desired. ∎

Example 7.30 Suppose again that the number of major defects on a randomly selected new vehicle of a certain type has a Poisson distribution with parameter μ. Now consider estimating $e^{-\mu}$, the probability that a vehicle has no such defects, based on a random sample of n vehicles. Let's start with the very simple estimator

$$U = \begin{cases} 1 & \text{if } X_1 = 0 \\ 0 & \text{if } X_1 > 0 \end{cases}$$

Using indicator function notation, this could be abbreviated $U = I(X_1 = 0)$. Then

$$E(U) = 1 \cdot P(X_1 = 0) + 0 \cdot P(X_1 > 0) = P(X_1 = 0) = \frac{e^{-\mu} \mu^0}{0!} = e^{-\mu}$$

Our estimator is therefore unbiased for estimating the probability of no defects. But the sufficient statistic here is $T = \sum X_i$, and of course the estimator U is not a function of T. The improved estimator is $U^* = E(U|T) = P(X_1 = 0 | \sum X_i)$. The event that $X_1 = 0$ and $T = t$ is identical to the event that the first vehicle has no defects and the total number of defects on the last $n - 1$ vehicles is t. Also, an mgf argument shows that T has a Poisson($n\mu$) distribution and the sum of the last $n - 1$ X_i's has a Poisson distribution with parameter $(n - 1)\mu$. Thus

$$P(X_1 = 0 | T = t) = \frac{P(X_1 = 0 \cap T = t)}{P(T = t)} = \frac{P\left(X_1 = 0 \cap \sum_{i=2}^{n} X_i = t\right)}{P(T = t)}$$

$$= \frac{P(X_1 = 0)P\left(\sum_{i=2}^{n} X_i = t\right)}{P(T = t)} = \frac{\frac{e^{-\mu}\mu^0}{0!} \cdot \frac{e^{-(n-1)\mu}[(n-1)\mu]^t}{t!}}{\frac{e^{-n\mu}(n\mu)^t}{t!}} = \left(1 - \frac{1}{n}\right)^t$$

That is, the improved unbiased estimator is $U^* = (1 - 1/n)^T$. Though the variance of U^* is difficult to derive, the Rao–Blackwell Theorem guarantees that its variance is no larger than that of U.

If, for example, there are a total of $t = 15$ defects among $n = 10$ randomly selected vehicles, then the estimate is $u^* = (1 - 1/10)^{15} = 206$. For this same sample, $\hat{\mu} = \bar{x} = 1.5$, so the maximum likelihood estimate of $e^{-\mu}$ is $e^{-1.5} = .223$. Here, as in some other situations, the principles of unbiased estimation and maximum likelihood are in conflict. However, if n is large, the improved estimate is $(1 - 1/n)^t = [(1 - 1/n)^n]^{\bar{x}} \approx e^{-\bar{x}}$, which is the mle of $e^{-\mu}$. That is, the unbiased and maximum likelihood estimators are "asymptotically equivalent." ∎

Further Comments

The Rao–Blackwell Theorem also helps us limit the scope of possible estimators to consider for a given distribution. If the statistic U is purely a function of the sufficient statistic T (and doesn't otherwise rely on the X_i's), then U and U^* are the same—in a sense, there was nothing to improve. If U is *not* purely a function of T, then the term $E[V(U|T)]$ in the proof will be strictly positive, and so U^* has strictly smaller variance than U. Said another way, for any statistic not based solely on a sufficient statistic, there exists some other estimator that is superior.

For example, in Section 7.1 we looked at several potential estimators for the parameter θ of a Uniform[0, θ] distribution, including $\max(X_1, \ldots, X_n)$ and $2\bar{X}$. In fact, one could concoct an endless set of candidates—when asked, students often propose estimators of the form $\bar{X} + cS$ for some judicious choice $c > 0$. But we saw in Example 7.27 that $\max(X_1, \ldots, X_n)$ is sufficient for θ, so any statistic that is not purely a function of the sample maximum is necessarily inferior to some other estimator. Since neither $\bar{X}$ nor S can be completely determined by the sample maximum, any estimator relying on one or both of these should be rejected out of hand.

We have emphasized that in general there will not be a unique sufficient statistic. Suppose there are two different sufficient statistics T_1 and T_2 such that the first one is not a one-to-one function of the second (e.g., we are not considering $T_1 = \sum X_i$ and $T_2 = \bar{X}$). Then it would be distressing if we started with an unbiased estimator U and found that $E(U \mid T_1) \neq E(U \mid T_2)$, so our improved estimator depended on which sufficient statistic we used. Fortunately there are general conditions under which, starting with a minimal sufficient statistic T, the improved estimator is the *unique* MVUE (minimum variance unbiased estimator).

Maximum likelihood is by far the most popular method for obtaining point estimates, so it would be disappointing if maximum likelihood estimators did not make full use of sample information. Fortunately the mles do not suffer from this defect. If $T_1, \ldots, T_k$ are jointly sufficient statistics for parameters $\theta_1, \ldots, \theta_m$, then the joint pmf or pdf factors as

$$f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m) = g(t_1, \ldots, t_k; \theta_1, \ldots, \theta_m) \cdot h(x_1, \ldots, x_n),$$

and the mles result from maximizing $f(\cdot)$ with respect to the θ_i's. Because the $h(\cdot)$ factor does not involve the parameters, this is equivalent to maximizing the $g(\cdot)$ factor with respect to the θ_i's. The resulting $\hat{\theta}_i$'s will involve the data only through the t_i's. Thus it is always possible to find a maximum

likelihood estimator that is a function of just the sufficient statistic(s). There are contrived examples of situations where the mle is not unique, in which case an mle that is not a function of the sufficient statistics can be constructed—but there is also one that *is* a function of the sufficient statistics.

The concept of sufficiency is very compelling when an investigator is sure the underlying distribution that generated the data is a member of some particular family (normal, exponential, etc.). However, two different families of distributions might each furnish plausible models for the data in a particular application, and yet the sufficient statistics for these two families might be different (an analogous comment applies to maximum likelihood estimation). For example, there are data sets for which a gamma probability plot suggests that a member of the gamma family would give a reasonable model and also a lognormal probability plot (normal probability plot of the logs of the observations) indicates that lognormality is plausible. Yet the jointly sufficient statistics for the parameters of the gamma family are not the same as those for the parameters of the lognormal family. When estimating some parameter θ in such situations (e.g., the mean μ or median $\tilde{\mu}$), one would look for a *robust estimator* that performs well for a wide variety of underlying distributions, as discussed in Section 7.1.

Exercises: Section 7.3 (38–50)

38. The long run proportion of vehicles that pass a certain emissions test is p. Suppose that three vehicles are independently selected for testing. Let $X_i = 1$ if the ith vehicle passes the test and $X_i = 0$ otherwise ($i = 1, 2, 3$), and let $T = X_1 + X_2 + X_3$. Use the definition of sufficiency to show that T is sufficient for p by obtaining the conditional distribution of the X_i's given that $T = t$ for each possible value t. Then generalize by giving an analogous argument for the case of n vehicles.

39. Components of a certain type are shipped in batches of size k. Suppose that whether or not any particular component is satisfactory is independent of the condition of any other component, and that the long run proportion of satisfactory components is p. Consider n batches, and let X_i denote the number of satisfactory components in the ith batch ($i = 1, 2, \ldots, n$). Statistician A is provided with the values of all the X_i's, whereas statistician B is given only the value of $T = \sum X_i$. Use a conditional probability argument to decide whether statistician A has more information about p than does statistician B.

40. Let $X_1, \ldots, X_n$ be a random sample of component lifetimes from an exponential distribution with parameter λ. Use the

factorization theorem to show that $\sum X_i$ is a sufficient statistic for λ.

41. Identify a pair of jointly sufficient statistics for the two parameters of a gamma distribution based on a random sample of size n from that distribution.

42. Identify a pair of jointly sufficient statistics for the two parameters of a beta distribution based on a random sample of size n from that distribution.

43. Messages are sent repeatedly across a noisy communication system until r arrive successfully. Let $X =$ the number of transmission required, so that X has a negative binomial distribution with parameters r (known) and p (unknown). Determine a sufficient statistic for p based on a random sample $X_1, \ldots, X_n$ from this negative binomial distribution.

44. Suppose waiting time for delivery of an item is uniform on the interval from θ_1 to θ_2. Consider a random sample of n waiting times, and use the factorization theorem to show that the sample minimum and maximum are a pair of jointly sufficient statistics for θ_1 and θ_2. [*Hint*: Introduce an appropriate indicator function as we did in Example 7.27.]

45. For $\theta > 0$ consider a random sample from a uniform distribution on the interval from θ

to 2θ, and use the factorization theorem to determine a sufficient statistic for θ.

46. Suppose that survival time X has a log-normal distribution with parameters μ and σ (which are the mean and standard deviation of $\ln(X)$, not of X itself). Are $\sum X_i$ and $\sum X_i^2$ jointly sufficient for the two parameters? If not, what is a pair of jointly sufficient statistics?

47. The probability that any particular component of a certain type works in a satisfactory manner is p. If n of these components are independently selected, then the statistic X, the number among the selected components that perform in a satisfactory manner, is sufficient for p. You must purchase two of these n components for a particular system. Obtain an unbiased statistic for the probability that exactly one of your purchased components will perform in a satisfactory manner. [*Hint*: Start with the statistic U, the indicator function of the event that exactly one of the first two components in the sample of size n performs as desired, and improve on it by conditioning on the sufficient statistic.]

48. In Example 7.30, we started with $U = I(X_1 = 0)$ and used a conditional expectation argument to obtain an unbiased estimator of the zero-defect probability based on the sufficient statistic. Consider now starting with a different statistic: $U = \sum I(X_i = 0)/n$. Show that the improved estimator based on the sufficient statistic is identical to the one obtained in the cited example. [*Hint*: Use the general property $E(Y + Z|T) = E(Y|T) + E(Z|T)$.]

49. In this section, it was established that $\sum X_i$ and $\overline{X}$ are both sufficient statistics for estimating the parameter μ of a Poisson distribution. We know that $E(\overline{X}) = \mu$, but it is also true that $E(S^2) = \sigma^2 = \mu$ for Poisson data. So, another unbiased estimator for μ is $\hat{\mu} = (\overline{X} + S^2)/2$. Which of these three estimators—$\overline{X}, S^2$, or $\hat{\mu}$—is the best choice for estimating μ? Why?

50. A particular quality characteristic of items produced using a certain process is known to be normally distributed with mean μ and standard deviation 1. Let X denote the value of the characteristic for a randomly selected item. An unbiased estimator for the parameter $\theta = P(X \leq c)$, where c is a critical threshold, is desired. The estimator will be based on a random sample $X_1, \ldots, X_n$.

    a. Obtain a sufficient statistic for μ.

    b. Consider the estimator $\hat{\theta} = I(X_1 \leq c)$. Obtain an improved unbiased estimator based on the sufficient statistic (it is actually the minimum variance unbiased estimator). [*Hint*: You may use the following facts: (1) The joint distribution of X_1 and $\overline{X}$ is bivariate normal with means μ and μ, variances 1 and $1/n$, respectively, and correlation ρ (which you should determine). (2) If Y_1 and Y_2 have a bivariate normal distribution, then the conditional distribution of Y_1 given that $Y_2 = y_2$ is normal with mean $\mu_1 + (\rho\sigma_1/\sigma_2)(y_2 - \mu_2)$ and variance $\sigma_1^2(1 - \rho)^2$.]

7.4   Information and Efficiency

In this section we introduce the idea of *Fisher information* and two of its applications. The first application is to find the minimum possible variance for an unbiased estimator. The second application is to show that the maximum likelihood estimator is asymptotically unbiased and normal (that is, for large n it has expected value approximately θ and it has approximately a normal distribution) with the minimum possible variance.

To motivate Fisher information, consider a rv $Y \sim \text{Bin}(n, p)$ with p unknown, and imagine determining the mle of p from the log-likelihood function

$$\ell(p) = \ln\left[\binom{n}{y} p^y (1-p)^{n-y}\right] = \ln\binom{n}{y} + y\ln(p) + (n-y)\ln(1-p)$$

Figure 7.7 presents a graph of $\ell(p)$ for two cases: ($n = 25$, $y = 19$) and ($n = 100$, $y = 76$). By definition, the mle maximizes $\ell(p)$; both functions graphed in Figure 7.7 achieve a maximum at .76, because the mle for the binomial model is $\hat{p} = y/n$ and here $19/25 = .76 = 76/100$. But the curves are not identical; in particular, the graph for $n = 100$ is much more concave than the one for $n = 25$. From calculus, this means that the *second derivative* of $\ell(p)$ has greater magnitude when $n = 100$ than when $n = 25$. Notice that in the vicinity of the local maximum, $\ell(p)$ is concave down and so its second derivative is negative; the preceding observation can thus be restated as $-\ell''(p)$ is larger when n is larger.
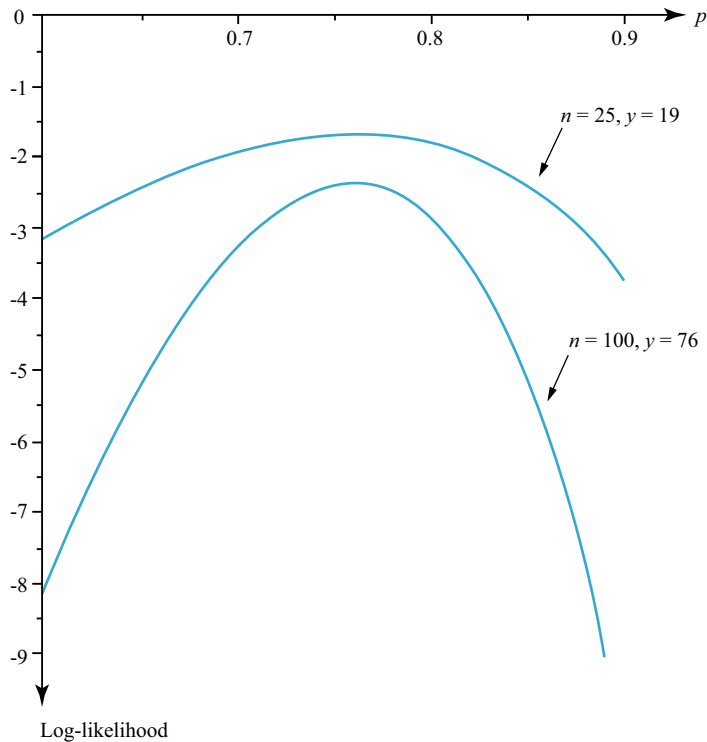
Figure 7.7 Binomial log-likelihood functions for $n = 25$ and $n = 100$

What does any of this have to do with "information"? Intuitively, a sample of size $n = 100$ contains more information than does a sample of size $n = 25$. Statistician R. A. Fisher was one of the first to notice the connection between sample size and the concavity of the log-likelihood function, leading to the following definition.

DEFINITION   Let $f(x; \theta)$ denote a pmf or pdf. The **Fisher information $I(\theta)$** in a single observation X from $f(x; \theta)$ is defined by

$$I(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2}\ln(f(X;\theta))\right] \tag{7.5}$$

Partial derivative notation is used in (7.5) to emphasize that the log-likelihood function depends on both X and θ. Since X is a random variable in the definition $\ell(\theta) = \ln f(X; \theta)$, $\ell(\theta)$ and its derivatives with respect to θ are also random variables. Thus (7.5) can be reexpressed as $I(\theta) = E[-\ell''(\theta)]$.

Example 7.31   Let X be a Bernoulli rv, so $f(x; p) = p^x(1-p)^{1-x}$, $x = 0, 1$. Then the second derivative of the log-likelihood function is

$$\ell''(p) = \frac{\partial^2}{\partial p^2}\ln[p^x(1-p)^{1-x}] = \frac{\partial^2}{\partial p^2}[x\ln(p)] + \frac{\partial^2}{\partial p^2}[(1-x)\ln(1-p)] = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

To calculate Fisher information, multiply by –1, replace x with X, and calculate the expected value of the resulting expression:

$$I(p) = E\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] = \frac{E(X)}{p^2} + \frac{1-E(X)}{(1-p)^2} = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)}$$

The denominator of this expression is maximized when $p = .5$, so the Fisher information in a single Bernoulli trial is smallest when $p = .5$ and increases as p approaches 0 or 1.   ∎

Example 7.32   Suppose X has the pdf $f(x; \theta) = \theta x^{\theta-1}$ for $0 \leq x \leq 1$. Then

$$\ell''(\theta) = \frac{\partial^2}{\partial\theta^2}\ln(\theta x^{\theta-1}) = \frac{\partial^2}{\partial\theta^2}\ln(\theta) + \frac{\partial^2}{\partial\theta^2}[(\theta-1)\ln(x)] = -\frac{1}{\theta^2} + 0 = -\frac{1}{\theta^2}$$

Since x does not appear in the second derivative, Fisher information here is easily determined to be $I(\theta) = E[-(-1/\theta^2)] = 1/\theta^2$.   ∎

Although Expression (7.5) is often the computationally simplest method for determining Fisher information, it is useful to have an alternative expression for $I(\theta)$.

PROPOSITION
$$I(\theta) = V\left[\frac{\partial}{\partial\theta}\ln f(X;\theta)\right] \tag{7.6}$$

provided that the order of the partial derivative and expectation operations in the definition of Fisher information can be interchanged. Critically, for this interchange to be valid, the support of the distribution (i.e., the set of possible x values) cannot depend on θ.

The quantity $\frac{\partial}{\partial\theta}\ln f(X;\theta)$ that appears in (7.6) is referred to as the **score function** and will shortly play an important role. The score function is simply $\ell'(\theta)$, the first derivative of the log-likelihood function, treated as a random variable. Under that perspective, the foregoing proposition can be restated as $I(\theta) = V(\ell'(\theta))$.

Proof The proof presented here assumes a discrete distribution; for the continuous case, replace the summations below with integrals. We first establish that, under the assumptions of the proposition, *the score function has expected value equal to zero*; this fact will prove useful in its own right. By the law of the unconscious statistician,

$$E[\ell'(\theta)] = E\left[\frac{\partial}{\partial\theta}\ln(f(X;\theta))\right] = \sum_x \frac{\partial}{\partial\theta}\ln(f(x;\theta)) \cdot f(x;\theta)$$

$$= \sum_x \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)} \cdot f(x;\theta) = \sum_x \frac{\partial}{\partial\theta}f(x;\theta) = \frac{d}{d\theta}\sum_x f(x;\theta)$$

$$= \frac{d}{d\theta}[1] \quad \text{because every pmf sums to 1}$$

$$= 0$$

To establish the equivalency of (7.5) and (7.6), take another derivative, which must also be 0 since $E[\ell'(\theta)] = 0$:

$$0 = \frac{d}{d\theta}E[\ell'(\theta)] = \frac{d}{d\theta}\sum_x \frac{\partial}{\partial\theta}\ln(f(x;\theta)) \cdot f(x;\theta) = \sum_x \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\ln(f(x;\theta)) \cdot f(x;\theta)\right]$$

$$= \sum_x \left[\frac{\partial^2}{\partial\theta^2}\ln(f(x;\theta)) \cdot f(x;\theta) + \frac{\partial}{\partial\theta}\ln(f(x;\theta)) \cdot \frac{\partial}{\partial\theta}f(x;\theta)\right]$$

$$= \sum_x \left[\frac{\partial^2}{\partial\theta^2}\ln(f(x;\theta)) \cdot f(x;\theta)\right] + \sum_x \left[\frac{\partial}{\partial\theta}\ln(f(x;\theta)) \cdot \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)} \cdot f(x;\theta)\right]$$

$$= E\left[\frac{\partial^2}{\partial\theta^2}\ln(f(X;\theta))\right] + E\left[\frac{\partial}{\partial\theta}\ln(f(X;\theta)) \cdot \frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\right]$$

$$= -I(\theta) + E\left[\frac{\partial}{\partial\theta}\ln(f(X;\theta)) \cdot \frac{\partial}{\partial\theta}\ln(f(X;\theta))\right] = -I(\theta) + E[\{\ell'(\theta)\}^2]$$

Therefore, $I(\theta) = E[\{\ell'(\theta)\}^2] = V(\ell'(\theta)) + \{E[\ell'(\theta)]\}^2 = V(\ell'(\theta)) + 0^2 = V(\ell'(\theta))$, completing the proof.   ∎

Example 7.33 Reconsider the single Bernoulli observation X from Example 7.31. The score function is

$$\ell'(p) = \frac{\partial}{\partial p}\ln(f(X;p)) = \frac{\partial}{\partial p}[X\ln p + (1-X)\ln(1-p)] = \frac{X}{p} - \frac{1-X}{1-p} = \frac{X-p}{p(1-p)}$$

[This expression indeed has mean zero, as indicated in the foregoing proof.] Apply (7.6):

$$I(p) = V\left[\frac{X - p}{p(1 - p)}\right] = \frac{V(X - p)}{[p(1 - p)]^2} = \frac{V(X)}{[p(1 - p)]^2} = \frac{p(1 - p)}{[p(1 - p)]^2} = \frac{1}{p(1 - p)},$$

which agrees with the result in Example 7.31.   ∎

In principle, the same method could be applied to the pdf from Example 7.32, for which the score function is

$$\ell'(\theta) = \frac{\partial}{\partial \theta}[\ln(\theta) + (\theta - 1)\ln(X)] = \frac{1}{\theta} + \ln(X)$$

Then Fisher information *could* be calculated via (7.6):

$$I(\theta) = V\left(\frac{1}{\theta} + \ln(X)\right) = V(\ln(X))$$

While the calculus to determine the variance of $\ln(X)$ is not insurmountable, the method using (7.5) shown in Example 7.32 is computationally much easier.

Information in a Random Sample

The definition of Fisher information extends to n rvs $X_1, \ldots, X_n$; simply replace $f(X; \theta)$ in (7.5) or (7.6) with the joint pmf/pdf of the X_i's. When $X_1, \ldots, X_n$ represent a random sample from some distribution $f(x; \theta)$, the Fisher information in the sample can be easily computed from the information in a single observation.

ADDITIVE PRINCIPLE OF INFORMATION

Let $X_1, \ldots, X_n$ be a random sample from a distribution with pmf or pdf $f(x; \theta)$. Then the Fisher information in $X_1, \ldots, X_n$ is simply n times the Fisher information in a single observation. That is, if $I_n(\theta)$ denotes the Fisher information in the sample, then

$$I_n(\theta) = n \cdot I(\theta),$$

where $I(\theta)$ denotes the Fisher information in a single observation from $f(x; \theta)$.

Proof Since the X_i's form a random sample, $f(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \cdot \cdots \cdot f(x_n; \theta)$. The result then follows from simple linearity properties:

$$I_n(\theta) = E\left[-\frac{\partial^2}{\partial \theta^2}\ln(f(X_1, \ldots, X_n; \theta))\right] = E\left[-\frac{\partial^2}{\partial \theta^2}\ln(f(X_1; \theta) \cdot \cdots \cdot f(X_n; \theta))\right]$$

$$= E\left[-\frac{\partial^2}{\partial \theta^2}\sum_{i=1}^{n}\ln(f(X_i; \theta))\right] = E\left[-\sum_{i=1}^{n}\frac{\partial^2}{\partial \theta^2}\ln(f(X_i; \theta))\right]$$

$$= \sum_{i=1}^{n}E\left[-\frac{\partial^2}{\partial \theta^2}\ln(f(X_i; \theta))\right] = \sum_{i=1}^{n}I(\theta) = n \cdot I(\theta)$$

                                                                                                                                                            ∎

The Additive Principle of Information makes sense intuitively, because it says that twice as many observations yield twice as much information. This property also saves us the hassle of constructing large joint pmf/pdf expressions. The aforementioned connections between Fisher information and the log-likelihood function still apply: with the log-likelihood $\ell(\theta) = \ln f(X_1, \ldots, X_n; \theta)$ regarded as a random variable,

$$I_n(\theta) = E[-\ell''(\theta)] = V(\ell'(\theta))$$

Example 7.34 Continuing with Example 7.31, let X_1, X_2, …, X_n be a random sample from a Bernoulli distribution. We saw that the information in a single observation is $I(p) = 1/[p(1-p)]$, and therefore the Fisher information in the random sample is $I_n(p) = nI(p) = n/[p(1-p)]$.

Astute readers will notice that Fisher information here is exactly the reciprocal of the variance of $\hat{P}$, which is the mle of p for Bernoulli data. As we'll see later in this section, this is not a coincidence. ∎

The Cramér–Rao Inequality

We now use the concept of Fisher information to show that if a statistic is an unbiased estimator of θ, then its minimum possible variance is the reciprocal of $I_n(\theta)$. Harald Cramér in Sweden and C. R. Rao in India independently derived this inequality during World War II, but R. A. Fisher had some notion of it 20 years previously.

CRAMÉR–RAO INEQUALITY	Let $X_1, \ldots, X_n$ be a random sample from the distribution with pmf or pdf $f(x; \theta)$ whose support does not depend on θ. If the statistic $T = t(X_1, \ldots, X_n)$ is an unbiased estimator of the parameter θ, then $$V(T) \geq \frac{1}{I_n(\theta)} = \frac{1}{nI(\theta)}$$

Proof The clever idea here is to consider the correlation ρ between T and the score function and exploit the fact that $-1 \leq \rho \leq 1$. We will need the fact from an earlier proof in this section that the mean of the score function is zero: $E[\ell'(\theta)] = 0$. Using this fact and the covariance expression $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$, the covariance of T and the score function $\ell'(\theta)$ is

$$\text{Cov}(T, \ell'(\theta)) = E(T \cdot \ell'(\theta)) - 0 = E\left[T \cdot \frac{\partial}{\partial \theta} \ln f(X_1, \ldots, X_n; \theta)\right] = E\left[T \cdot \frac{\frac{\partial}{\partial \theta} f(X_1, \ldots, X_n; \theta)}{f(X_1, \ldots, X_n; \theta)}\right]$$

$$= \sum_{x_1, \ldots, x_n} t(x_1, \ldots, x_n) \cdot \frac{\frac{\partial}{\partial \theta} f(x_1, \ldots, x_n; \theta)}{f(x_1, \ldots, x_n; \theta)} \cdot f(x_1, \ldots, x_n; \theta)$$

$$= \sum_{x_1, \ldots, x_n} t(x_1, \ldots, x_n) \cdot \frac{\partial}{\partial \theta} f(x_1, \ldots, x_n; \theta)$$

$$= \frac{d}{d\theta} \sum_{x_1, \ldots, x_n} t(x_1, \ldots, x_n) f(x_1, \ldots, x_n; \theta) = \frac{d}{d\theta} E(T)$$

If $T = t(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of θ, then $E(T) = \theta$, so the derivative in the last expression is just 1, from which we deduce that $\text{Cov}(T, \ell'(\theta)) = 1$.

Now recall from Section 5.2 that the correlation between two rvs X and Y is $\rho_{X,Y} = \text{Cov}(X, Y)/(\sigma_X \sigma_Y)$. Therefore,

$$\text{Cov}(X, Y)^2 = \rho_{X,Y}^2 \sigma_X^2 \sigma_Y^2 \leq 1 \sigma_X^2 \sigma_Y^2 = V(X)V(Y)$$

Apply this to T and the score function $\ell'(\theta)$:

$$1 = \text{Cov}(T, \ell'(\theta))^2 \leq V(T)V(\ell'(\theta)) = V(T) \cdot I_n(\theta),$$

and the desired inequality follows.   ∎

Because the variance of T must be at least $1/I_n(\theta)$, it is natural to call T an efficient estimator of θ if $V(T) = 1/I_n(\theta)$.

DEFINITION   Let T be an unbiased estimator of θ. The **efficiency** of T is the ratio of the Cramér–Rao lower bound to the variance of T:

$$\text{efficiency of } T = \frac{1/I_n(\theta)}{V(T)} = \frac{1}{V(T) \cdot I_n(\theta)}$$

T is said to be an **efficient** estimator of θ if T achieves the lower bound (so its efficiency is 1; otherwise, efficiency will be less than 1). An efficient estimator is a minimum variance unbiased (MVUE) estimator, as discussed in Section 7.1.

Example 7.35 (Example 7.34 continued) Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution. We saw that the Fisher information in the sample is $I_n(p) = n/[p(1-p)]$, and therefore the Cramér–Rao lower bound on the variance of *any* unbiased estimator of p is $1/I_n(p) = p(1-p)/n$. Let $T = \hat{P} = \sum X_i/n$, the sample proportion of successes. It was established in Example 7.4 that $\hat{P}$ is an unbiased estimator of p and that $V(\hat{P}) = p(1-p)/n$. Because T is unbiased and $V(T)$ is equal to the lower bound, T has efficiency 1 and therefore it is an efficient estimator.   ∎

The Cramér–Rao inequality can be generalized to an estimator whose expected value is not θ itself but rather some function $h(\theta)$. Using a similar proof, it can be shown (under the same requirements about the pmf/pdf) that the lower bound on the variance of any statistic T with mean $h(\theta)$ is

$$V(T) \geq \frac{[h'(\theta)]^2}{I_n(\theta)}$$

In the special case that T is unbiased for θ, then $h(\theta) = \theta$, $h'(\theta) = 1$, and we have the original Cramér–Rao inequality. (See Exercises 59–60 for applications of this more general result.)

Large-Sample Properties of the MLE

As mentioned briefly in Section 7.2, the maximum likelihood estimator $\hat{\theta}$ has some nice properties. First of all it is *consistent*, which means that it converges in probability to the parameter θ as the sample size increases. A verification of this is beyond the level of this book, but we can use it as a basis for showing that the mle is asymptotically normal with mean θ (asymptotic unbiasedness) and variance equal to the Cramér–Rao lower bound.

THEOREM   Let $X_1, \ldots, X_n$ be a random sample from a distribution whose support does not depend on θ. Then for large n the maximum likelihood estimator $\hat{\theta}$ has approximately a normal distribution with mean θ and variance $1/[nI(\theta)]$.

A proof of this result appears in the appendix to this chapter.

Example 7.36 It was established in Example 7.16 that the mle of p when sampling from a Bernoulli distribution is the sample proportion $\hat{P} = \sum X_i / n$. Recall from Example 7.35 that $\hat{P}$ is unbiased and efficient with the minimum variance of the Cramér–Rao inequality. Finally, $\hat{P}$ is asymptotically normal by the Central Limit Theorem. These properties are in accord with the asymptotic distribution given by the theorem, $\hat{P} \sim N(p, 1/[nI(p)])$. ∎

Example 7.37 (Example 7.32 continued) Consider a random sample $X_1, \ldots, X_n$ from the distribution with pdf $f(x; \theta) = \theta x^{\theta - 1}$ for $0 \leq x \leq 1$. The Fisher information in a single observation was found to be $I(\theta) = 1/\theta^2$. The maximum likelihood estimator of θ (see Exercise 27 for a similar example) is

$$\hat{\theta} = \frac{-1}{\sum \ln(X_i)/n} \tag{7.7}$$

The expected value of $\ln(X)$ for this distribution is $-1/\theta$, so the denominator of (7.7) converges in probability to $-1/\theta$ by the Law of Large Numbers. Therefore $\hat{\theta}$ converges in probability to θ, which means that $\hat{\theta}$ is consistent. (We knew this because the mle is always consistent, but it is also nice to show it directly.) Determining the exact distribution of $\hat{\theta}$ is quite difficult. However, by the preceding theorem, for large n the distribution of $\hat{\theta}$ is approximately normal, with mean θ and variance $1/[nI(\theta)] = \theta^2/n$. ∎

Sufficiency and Efficiency

As we discussed in Section 7.3, the Rao–Blackwell Theorem implies that any estimator not based purely on sufficient statistics is necessarily inferior (has greater variance than) some other statistic. So, we cannot expect a statistic to be an *efficient* estimator without first being *sufficient*.

   The proof of the Cramér–Rao inequality considered the correlation between two random variables: a statistic T and the score function $\ell'(\theta)$. The inequality followed from the fact that $|\rho| \leq 1$, but we know from Chapter 5 that equality can only occur—that is, $|\rho| = 1$—if the two rvs are linear functions of each other. So, suppose a statistic $T = t(X_1, \ldots, X_n)$ is *sufficient* for θ. By the Factorization Theorem, the score function may then be written as

$$\ell'(\theta) = \frac{\partial}{\partial \theta} \ln f(x_1, \ldots, x_n; \theta) = \frac{\partial}{\partial \theta} \ln[g(t; \theta) \cdot h(x_1, \ldots, x_n)]$$
$$= \frac{\partial}{\partial \theta} \ln[g(t; \theta)] + \frac{\partial}{\partial \theta} \ln[h(x_1, \ldots, x_n)] = \frac{\partial}{\partial \theta} \ln[g(t(x_1, \ldots, x_n); \theta)]$$

Therefore, an estimator can only be efficient if it is a linear function of $\frac{\partial}{\partial \theta} \ln[g(t(X_1, \ldots, X_n); \theta)]$. In particular, an *efficient* estimator can only depend on $X_1, \ldots, X_n$ through the *sufficient* statistic $T = t(X_1, \ldots, X_n)$. This result is consistent with the Rao–Blackwell Theorem.

Exercises: Section 7.4 (51–60)

51. The number of attempts required to successfully transmit a message across a noisy channel can be modeled by a geometric distribution, whose pmf is $(1 - p)^{x-1} p$ for $x = 1, 2, 3, \ldots$. To estimate the unknown parameter p we obtain the random sample $X_1, X_2, \ldots, X_n$ from this geometric distribution.

    a. Find the Fisher information in a single observation X using both (7.5) and (7.6).

    b. What is the Fisher information in the random sample?

    c. Determine the Cramér–Rao lower bound for the variance of an unbiased estimator of p.

52. Assume that the number of alpha particles emitted in one second by a particular radioactive source has a Poisson distribution with parameter μ. Consider estimating μ based on a random sample $X_1, X_2, \ldots, X_n$.

    a. Find the Fisher information in a single observation using both (7.5) and (7.6).

    b. Find the Cramér–Rao lower bound for the variance of an unbiased estimator of μ.

    c. Determine the mle of μ and show that the mle is an efficient estimator.

    d. Is the asymptotic distribution of the mle in accord with the last theorem of this section? Explain.

53. Let $X_1, \ldots, X_n$ be a random sample from the Uniform[0, θ] distribution.

    a. Use the expression $I(\theta) = E[(\ell'(\theta))^2]$ to determine the Fisher information in a single observation from this distribution.

    b. Find the Cramér–Rao lower bound for the variance of an unbiased estimator of θ.

    c. In Examples 7.9 and 7.10, two unbiased estimators for θ were proposed, one with variance $\theta^2/[n(n+2)]$ and another with variance $\theta^2/(3n)$. Compare these variances to part (b) and explain why they seem to contradict the Cramér–Rao inequality. What assumption is violated, causing the inequality not to apply here?

54. Survival times have the exponential distribution with pdf $f(x; \lambda) = \lambda e^{-\lambda x}$ for $x \geq 0$, where $\lambda > 0$ is unknown. However, we wish to estimate the mean $\mu = 1/\lambda$ based on the random sample $X_1, X_2, \ldots, X_n$, so let's re-express the pdf in the form $(1/\mu)e^{-x/\mu}$.

    a. Find the information in a single observation and the Cramér–Rao lower bound.

    b. Determine the mle of μ.

    c. Find the mean and variance of the mle.

    d. Is the mle an efficient estimator? Explain.

55. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution with known standard deviation σ.

    a. Find the mle of μ.

    b. Find the distribution of the mle.

    c. Is the mle an efficient estimator? Explain.

    d. How does the answer to part (b) compare with the asymptotic distribution given by the second theorem?

56. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution with known mean μ but with the variance σ^2 as the unknown parameter.

a. Find the Fisher information for σ^2 in a single observation and the Cramér–Rao lower bound.

b. Find the mle of σ^2.

c. Find the distribution of the mle.

d. Is the mle an efficient estimator? Explain.

e. Is the answer to part (c) in conflict with the asymptotic distribution of the mle given by the second theorem? Explain.

57. Let $X_1, X_2, \ldots, X_n$ be a random sample from the normal distribution with known mean μ but with the standard deviation σ as the unknown parameter.

a. Find the Fisher information for σ in a single observation.

b. Compare the answer in part (a) to the answer in Exercise 56(a). Does the information depend on the parameterization?

58. Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution with pdf $f(x; \theta)$. For large n, the variance of the sample median is approximately $1/\{4n[f(\tilde{\mu}; \theta)]^2\}$. If $X_1, X_2, \ldots, X_n$ is a random sample from the normal distribution with known standard deviation σ and unknown μ, determine the efficiency of the sample median.

59. Return to the geometric distribution from Exercise 51. Let $X_1, \ldots, X_n$ be a random sample from this distribution, and let $\overline{X}$ denote the sample mean.

a. Determine the expected value and variance of $\overline{X}$ as functions of p.

b. Using the generalization of the Cramér–Rao inequality presented in this section, determine the lower bound for the variance of any estimator whose expectation is equal to $E(\overline{X})$ from part (a).

c. Is $\overline{X}$ an efficient estimator of its expectation?

60. Return to the exponential distribution from Exercise 54. Let $X_1, \ldots, X_n$ be a random sample from this distribution, and let $\overline{X}$ denote the sample mean.

a. Find the Fisher information for λ in a single observation from this distribution.

b. Determine the expected value and variance of $\overline{X}$ as functions of λ.

c. Using the generalization of the Cramér–Rao inequality presented in this section, determine the lower bound for the variance of any estimator whose expectation is equal to $E(\overline{X})$ from part (b). Is $\overline{X}$ an efficient estimator of its expectation?

d. Does it follow that $1/\overline{X}$ is the MVUE of λ? Why or why not?

Supplementary Exercises: (61–78)

61. At time $t = 0$, there is one individual alive in a certain population. A **pure birth process** then unfolds as follows. The time until the first birth is exponentially distributed with parameter λ. After the first birth, there are two individuals alive. The time until the first gives birth again is exponential with parameter λ, and similarly for the second individual. Therefore, the time until the next birth is the minimum of two exponential (λ) variables, which is exponential with parameter 2λ. Similarly, once the second birth has occurred, there are three individuals alive, so the time until the next birth is an exponential rv with parameter 3λ, and so on (the memoryless property of the exponential distribution is being used here). Suppose the process is observed until the sixth birth has occurred and the successive birth times are 25.2, 41.7, 51.2, 55.5, 59.5, 61.8 (from which you should calculate the times between successive births). Derive the mle of λ. [*Hint*: The likelihood is a product of exponential terms.]

62. Let $X_1, \ldots, X_n$ be a random sample from a Uniform[0, θ] distribution, and let Y_n denote the largest observation: $Y_n = \max(X_1, \ldots, X_n)$. [This is the rv denoted $\hat{\theta}_b$ in several examples in Section 7.1.]

a. Show that the pdf of Y_n is

$$f(y) = \frac{ny^{n-1}}{\theta^n} \quad 0 \leq y \leq \theta$$

[*Hint*: Use the methods of Section 5.7, or use the relationship $F(y) = P(Y \leq y) = P(X_1 \leq y \cap \cdots \cap X_n \leq y)$.]

b. Use part (a) to determine the mean and variance of Y_n.

63. The proportion of iron in rock specimens from a certain quarry is assumed to follow a standard beta distribution with unknown parameters α and β. Suppose the following observations are made on a sample of $n = 6$ specimens: .873, .437, .249, .712, .501, .618. Calculate the method of moments estimates for α and β. [*Hint*: Be careful in determining the formula for $E(X^2)$.]

64. Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $[-\theta, \theta]$.

a. Determine the mle of θ. [*Hint*: Look back at what we did in Example 7.23.]

b. Give an intuitive argument for why the mle is either biased or unbiased.

c. Determine a sufficient statistic for θ. [*Hint*: See Example 7.27.]

d. Use the results of Section 5.7 to determine the joint pdf of the smallest order statistic Y_1 and the largest order statistic Y_n. Then use it to obtain the expected value of the mle. [*Hint*: Draw the region of joint positive density for Y_1 and Y_n, and identify what the mle is for each part of this region.]

e. What is an unbiased estimator for θ?

65. Carry out the details for minimizing MSE in Example 7.8: show that $c = 1/(n + 1)$ minimizes the MSE of $\hat{\sigma}^2 = c \sum (X_i - \overline{X})^2$ when the population distribution is normal.

66. Let $X_1, \ldots, X_n$ be a random sample from a pdf that is symmetric about μ. An estimator for μ that has been found to perform well for a variety of underlying distributions is the *Hodges–Lehmann estimator*. To define

it, first compute for each $i \leq j$ and each $j = 1, 2, \ldots, n$ the pairwise average $\overline{X}_{i,j} = (X_i + X_j)/2$. Then the estimator is $\hat{\mu} =$ the median of the $\overline{X}_{i,j}$'s. Compute the value of this estimate using the data of Exercise 53 of Chapter 1. [*Hint*: Construct a square table with the x_i's listed on the left margin and on top. Then compute averages on and above the diagonal.]

67. For a normal population distribution, the statistic

$$\hat{\sigma} = \text{median}(|X_1 - \tilde{X}|, \ldots, |X_n - \tilde{X}|)/.6745$$

can be used to estimate σ. This estimator is more resistant to the effects of outliers than is the sample standard deviation. Compute both the corresponding point estimate and s for the data of Example 7.2.

68. When the sample standard deviation S is based on a random sample from a normal population distribution, it can be shown that

$$E(S) = \sqrt{2/(n-1)}\,\Gamma(n/2)\sigma/\Gamma[(n-1)/2]$$

Use this to obtain an unbiased estimator for σ of the form cS. What is c when $n = 20$?

69. Each of n specimens is to be weighed twice on the same scale. Let X_i and Y_i denote the two observed weights for the ith specimen. Suppose X_i and Y_i are independent of each other, each normally distributed with mean value μ_i (the true weight of specimen i) and variance σ^2.

a. Show that the mle of σ^2 is $\hat{\sigma}^2 = \sum (X_i - Y_i)^2/(4n)$. [*Hint*: If $\overline{z} = (z_1 + z_2)/2$, then $\sum (z_i - \overline{z})^2 = (z_1 - z_2)^2/2$.]

b. Is the mle $\hat{\sigma}^2$ an unbiased estimator of σ^2? Find an unbiased estimator of σ^2. [*Hint*: For any rv Z, $E(Z^2) = V(Z) + [E(Z)]^2$. Apply this to $Z = X_i - Y_i$.]

70. For $0 < \theta < 1$ consider a random sample from a uniform distribution on the interval from θ to $1/\theta$. Identify a sufficient statistic for θ.

71. Let p denote the proportion of all individuals who are allergic to a particular medication. An investigator tests individual after individual to obtain a group of r individuals who have the allergy. Let $X_i = 1$ if the ith individual tested has the allergy and $X_i = 0$ otherwise ($i = 1, 2, 3,...$). Recall that in this situation, Y = the number of individuals tested to obtain the desired group has a negative binomial distribution. Use the *definition* of sufficiency to show that Y is a sufficient statistic for p.

72. The fraction of a bottle that is filled with a particular liquid is a continuous random variable X with pdf $f(x; \theta) = \theta x^{\theta-1}$ for $0 < x < 1$ (where $\theta > 0$).

    a. Obtain the method of moments estimator for θ.
    b. Is the estimator of (a) a sufficient statistic? If not, what is a sufficient statistic, and what is an estimator of θ (not necessarily unbiased) based on a sufficient statistic?

73. Let $X_1, ..., X_n$ be a random sample from a normal distribution with both μ and σ unknown. An unbiased estimator of $\theta = P(X \leq c)$ based on the jointly sufficient statistics is desired. Let $k = \sqrt{n/(n-1)}$ and $w = (c - \hat{\mu})/\hat{\sigma}$. Then it can be shown that the minimum variance unbiased estimator for θ is

$$\hat{\theta} = \begin{cases} 0 & kw \leq -1 \\ P\left(T < \frac{kw\sqrt{n-2}}{\sqrt{1-k^2w^2}}\right) & -1 < kw < 1 \\ 1 & kw \geq 1 \end{cases}$$

where T has a t distribution with $n-2$ df. The article "Big and Bad: How the S.U.V. Ran over Automobile Safety" (*The New Yorker*, Jan. 24, 2004) reported that when an engineer with Consumers Union (the product testing and rating organization that publishes *Consumer Reports*) performed three different trials in which a Chevrolet Blazer was accelerated to 60 mph and then suddenly braked, the stopping distances (ft) were 146.2, 151.6, and 153.4, respectively. Assuming that braking distance is normally distributed, obtain the minimum variance unbiased estimate for the probability that distance is at most 150 ft, and compare to the maximum likelihood estimate of this probability.

74. Here is a result that allows for easy identification of a *minimal* sufficient statistic: Suppose there is a function $t(x_1, ..., x_n)$ such that for any two sets of observations $x_1, ..., x_n$ and $y_1, ..., y_n$, the likelihood ratio $f(x_1, ..., x_n; \theta)/f(y_1, ..., y_n; \theta)$ doesn't depend on θ if and only if $t(x_1, ..., x_n) = t(y_1, ..., y_n)$. Then $T = t(X_1, ..., X_n)$ is a minimal sufficient statistic. The result is also valid if θ is replaced by $\theta_1, ..., \theta_m$, in which case there will typically be several jointly minimal sufficient statistics. For example, if the underlying pdf is exponential with parameter λ, then the likelihood ratio is $\lambda^{\Sigma x_i - \Sigma y_i}$, which will not depend on λ if and only if $\sum x_i = \sum y_i$, so $T = \sum x_i$ is a minimal sufficient statistic for λ (and so is the sample mean).

    a. Identify a minimal sufficient statistic when the X_i's are a random sample from a Poisson distribution.
    b. Identify a minimal sufficient statistic or jointly minimal sufficient statistics when the X_i's are a random sample from a normal distribution with mean θ and variance θ.
    c. Identify a minimal sufficient statistic or jointly minimal sufficient statistics when the X_i's are a random sample from a normal distribution with mean θ and standard deviation θ.

75. The principle of unbiased estimation has been criticized on the grounds that in some situations the only unbiased estimator is patently ridiculous. Here is one such example. Suppose that the number of blemishes X on a randomly selected piece

of fruit has a Poisson distribution with parameter μ. You are going to purchase two such pieces of fruit and wish to estimate $\theta = e^{-2\mu}$, the probability that neither of these has any blemishes. But your estimate is based on observing the value of X for a single piece. Obtain an estimator $\hat{\theta} = d(X)$ that is unbiased for θ; i.e., such that $E[d(X)] = e^{-2\mu}$. [*Hint*: Set the summation for $E[d(X)]$ equal to $e^{-2\mu}$, cancel $e^{-\mu}$ from both sides, then expand what remains on the right-hand side in a Taylor series and compare the two sides to determine $d(X)$.] If $X = 200$, what is the estimate? Does this seem reasonable? What is the estimate if $X = 199$? Is this reasonable?

76. Let X, the payoff from playing a certain game, have pmf

$$p(x; \theta) = \begin{cases} \theta & x = -1 \\ (1 - \theta)^2 \theta^x & x = 0, 1, 2, \dots \end{cases}$$

a. Verify that $p(x; \theta)$ is a legitimate pmf, and determine the expected payoff. [*Hint*: Look back at how the properties of a geometric random variable were developed in Chapter 3.]

b. Let $X_1, \dots, X_n$ be the payoffs from n independent games of this type. Determine the mle of θ. [*Hint*: Let Y denote the number of observations among the n that equal -1; that is, $Y = \sum I(X_i = -1)$, where $I(A) = 1$ if A occurs and 0 otherwise. Then, write the likelihood as a single expression in terms of $\sum x_i$ and y.]

c. What is the approximate variance of the mle when n is large?

77. *Regression through the origin.* Let x denote the number of items in an order and y denote time (min) necessary to process the order. Processing time may be determined by various factors other than order size. So for any particular value of x, we now regard

the value of total production time as a random variable Y. Consider the following data obtained by specifying various values of x and determining total production time for each one.

x	10	15	18	20	25
y	301	455	533	599	750
x	27	30	35	36	40
y	810	903	1054	1088	1196

a. Plot the observed (x, y) pairs on a two-dimensional coordinate system. Do all points fall *exactly* on a line passing through $(0, 0)$? Do the points tend to fall *close to* such a line?

b. Consider the following probability model for the data. Values $x_1, x_2, \dots, x_n$ are specified, and at each x_i we will observe a value of the dependent variable Y_i. Assume that the $Y_1, \dots, Y_n$ are independent and normally distributed, with Y_i having mean value βx_i and variance σ^2. That is, rather than assume that $y = \beta x$, a linear function of x passing through the origin, we are assuming that the *mean value* of Y is a linear function of x and that the variance of Y is the same for any particular x value. Obtain formulas for the maximum likelihood estimates of β and σ^2, and then calculate the estimates for the given data. How would you interpret the estimate of β? What value of processing time would you predict when $x = 25$? [*Hint*: The likelihood is a product of individual normal pdfs with different mean values and the same variance. Proceed as in the estimation via maximum likelihood of the parameters μ and σ^2 based on a random sample from a normal population distribution.]

78. Reconsider the "regression through the origin" situation presented in the previous exercise. Consider the following three estimators for the slope parameter β (one of

which is the mle obtained in the previous exercise):

$$\hat{\beta}_1 = \frac{\sum Y_i}{\sum x_i} \quad \hat{\beta}_2 = \frac{1}{n}\sum \frac{Y_i}{x_i} \quad \hat{\beta}_3 = \frac{\sum x_i Y_i}{\sum x_i^2}$$

a. Show that all three of these estimators are unbiased for β.
b. Determine the variance of all three estimators, and comment on what you find.

Proof of the Asymptotic Distribution of the MLE

Let $\hat{\theta}$ denote the mle of θ, and consider again the score function $\ell'(\theta) = \frac{\partial}{\partial \theta}\ln f(X_1, X_2, \ldots, X_n; \theta)$. Its derivative $\ell''(\theta)$ at the true parameter value θ is approximately equal to the following difference quotient:

$$\ell''(\theta) \approx \frac{\ell'(\hat{\theta}) - \ell'(\theta)}{\hat{\theta} - \theta} \tag{7.8}$$

Moreover, the error in Equation (7.8)—i.e., the difference between the two sides of the $\approx$ sign—approaches zero as $n \to \infty$ because $\hat{\theta}$ approaches θ (consistency). Now, because $\hat{\theta}$ is the mle, by definition $\ell'(\hat{\theta}) = 0$, and (7.8) can be re-arranged to write

$$\hat{\theta} - \theta \approx \frac{\ell'(\theta)}{-\ell''(\theta)} \Rightarrow \sqrt{n}\left(\hat{\theta} - \theta\right) \approx \frac{\sqrt{n}\ell'(\theta)}{-\ell''(\theta)} = \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{\frac{1}{n}[-\ell''(\theta)]} \tag{7.9}$$

Similar to the proof of the additive principle of information, the denominator may be written as

$$\frac{1}{n}[-\ell''(\theta)] = \frac{1}{n}\left[\left(-\frac{\partial^2}{\partial \theta^2}\ln f(X_1; \theta)\right) + \cdots + \left(-\frac{\partial^2}{\partial \theta^2}\ln f(X_n; \theta)\right)\right],$$

the average of n iid random variables each with mean $I(\theta)$. Therefore, by the Law of Large Numbers, the denominator converges to $I(\theta)$. At the same time, the numerator of (7.9) is

$$\frac{1}{\sqrt{n}}\ell'(\theta) = \frac{1}{\sqrt{n}}\left[\left(\frac{\partial}{\partial \theta}\ln f(X_1; \theta)\right) + \cdots + \left(\frac{\partial}{\partial \theta}\ln f(X_n; \theta)\right)\right]$$

The terms in parentheses are also iid, each with mean 0 (the mean of the score function is zero) and variance $I(\theta)$ by (7.6). It follows from the Central Limit Theorem that the numerator converges to a normal rv with mean 0 and standard deviation $\sqrt{I(\theta)}$.

Combining these two results, the ratio on the right-hand side of (7.9) is approximately normal with mean 0 and standard deviation $\sqrt{I(\theta)}/I(\theta) = 1/\sqrt{I(\theta)}$. That is, $\sqrt{n}(\hat{\theta} - \theta)$ is approximately $N(0, 1/\sqrt{I(\theta)})$, and it follows that $\hat{\theta}$ is approximately normal with mean θ and variance $1/[nI(\theta)]$, the Cramér–Rao lower bound. ∎

Statistical Intervals Based on a Single Sample

<div style="text-align:right">**8**</div>

Introduction

A point estimate, because it is a single number, by itself provides no information about the precision and reliability of estimation. Consider, for example, using the statistic $\overline{X}$ to calculate a point estimate for the true average breaking strength of a certain brand of paper towels, and suppose that $\overline{x} = 9322.7$ grams. Because of sampling variability, it is virtually never the case that $\overline{x} = \mu$. The point estimate says nothing about how close it might be to μ. An alternative to reporting a single sensible value for the parameter being estimated is to calculate and report an entire interval of plausible values—an *interval estimate* or *confidence interval* (CI).

A confidence interval is calculated by first selecting a *confidence level*, which is a measure of the degree of reliability of the interval. A confidence interval with a 95% confidence level for the true average breaking strength might have a lower limit of 9162.5 and an upper limit of 9482.9. Then at the 95% confidence level, any value of μ between 9162.5 and 9482.9 g is plausible. The higher the confidence level, the more strongly we believe that the value of the parameter being estimated lies within the interval (an interpretation of any particular confidence level will be given shortly).

Information about the precision of an interval estimate is conveyed by the width of the interval. If the confidence level is high and the resulting interval is quite narrow, our knowledge of the value of the parameter is reasonably precise. A very wide confidence interval, however, gives the message that there is a great deal of uncertainty concerning the value of what we are estimating. Figure 8.1 shows 95% confidence intervals for true average breaking strengths of two different brands of paper towels. One of these intervals suggests precise knowledge about μ, whereas the other suggests a very wide range of plausible values.
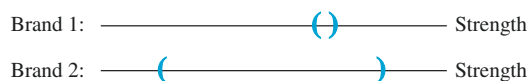
Figure 8.1 Confidence intervals indicating precise (Brand 1) and imprecise (Brand 2) information about μ

8.1   Basic Properties of Confidence Intervals

The basic concepts and properties of confidence intervals (CIs) are most easily introduced by first focusing on a simple, albeit somewhat unrealistic, problem situation. Suppose that the parameter of interest is a population mean μ and that

1. The population distribution is normal.
2. The value of the population standard deviation σ is known.

Population normality is often a reasonable assumption and can be checked by examining a normal probability plot of the sample data. However, if the value of μ is unknown, it is unlikely that the value of σ would be available (knowledge of a population's center typically precedes information concerning spread). In later sections, we will develop methods based on less restrictive assumptions.

Example 8.1 Titanium alloys are used in everything from offshore oil operations to toys (remember the fidget spinner?). The article "Statistical Analysis of Tensile Strength and Elongation of Pulse TIG Welded Titanium Alloy Joints Using Weibull Distribution" (*Cogent Engr.* 2016) described an experiment designed to study various characteristics of a certain type of weld. A total of $n = 31$ experimental runs resulted in a sample mean tensile strength of $\bar{x} = 1064$ MPa, and the data suggests that tensile strength measurements can be modeled with a normal distribution (despite the Weibull reference in the article's title!). Assuming the population standard deviation for tensile strength of these welds is $\sigma = 55$ MPa (a value suggested by data in the article), we will see shortly how to obtain an interval of plausible values for μ, the true average tensile strength of all such titanium alloy welds. ∎

The actual sample observations x_1, x_2, ..., x_n are assumed to be the result of a random sample X_1, ..., X_n from a $N(\mu,\sigma)$ distribution. The results of Chapter 6 then imply that the sample mean $\overline{X}$ is normally distributed, with expected value μ and standard deviation $\sigma/\sqrt{n}$. Standardizing $\overline{X}$ by first subtracting its expected value and then dividing by its standard deviation yields the variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{8.1}$$

Then Z has a standard normal distribution. Because the area under the standard normal curve between -1.96 and 1.96 is .95,

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95 \tag{8.2}$$

The next step in the development of our CI is to manipulate the inequalities inside the parentheses in (8.2) so that they appear in the equivalent form $l < \mu < u$, where the endpoints l and u involve $\overline{X}$ and $\sigma/\sqrt{n}$. Multiplying all terms in the inequalities by $\sigma/\sqrt{n}$, subtracting $\overline{X}$ from each term, and then multiplying through by -1 (to eliminate the negative sign in front of μ) gives

$$\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

These endpoints also result from replacing each < by = in (8.2) and solving for μ.

Because this last set of inequalities is equivalent to those inside (8.2), it follows that

$$P\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = .95 \tag{8.3}$$

The event inside the parentheses in (8.3) has a somewhat unfamiliar appearance. Previously, the random quantity has appeared in the middle with constants on both ends, as in $a \leq Y \leq b$. But in (8.3) the random quantity appears on the two ends and the unknown constant μ appears in the middle. To interpret (8.3), think of a *random* interval having left endpoint $\overline{X} - 1.96 \cdot \sigma/\sqrt{n}$ and right endpoint $\overline{X} + 1.96 \cdot \sigma/\sqrt{n}$, which in interval notation is

$$\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \ \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{8.4}$$

The interval (8.4) is random because the two endpoints of the interval involve a random variable. Note that the interval is centered at the sample mean $\overline{X}$ and extends $1.96 \cdot \sigma/\sqrt{n}$ to each side of $\overline{X}$. Thus the interval's width is $2 \cdot 1.96 \cdot \sigma/\sqrt{n}$, which is not random; only the location of the interval, its midpoint $\overline{X}$, is random (see Figure 8.2). Now (8.3) can be paraphrased as "*the probability is .95 that the random interval* (8.4) *includes or covers the true value of* μ." Before any experiment is performed and any data is gathered, it is quite likely (probability .95) that μ will lie inside the interval in Expression (8.4).
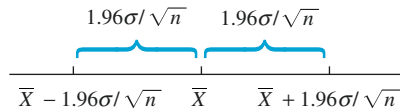
Figure 8.2 The random interval (8.4) centered at $\overline{X}$

DEFINITION   If after observing $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n,$ we compute the observed sample mean $\overline{x}$ and then substitute $\overline{x}$ into (8.4) in place of $\overline{X}$, the resulting fixed interval is called a **95% confidence interval for μ**. This CI can be expressed either as

$$\left(\overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \text{ is a 95\% confidence interval for } \mu$$

or as

$$\overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \text{ with 95\% confidence}$$

A concise expression for the interval is $\overline{x} \pm 1.96 \cdot \sigma/\sqrt{n}$, where $-$ gives the left endpoint (lower limit) and $+$ gives the right endpoint (upper limit).

Example 8.2 (Example 8.1 Continued) The quantities needed for computation of the 95% CI for true average tensile strength are $\sigma = 55$, $n = 31$, and $\bar{x} = 1064$. The resulting interval is

$$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} = 1064 \pm 1.96 \cdot \frac{55}{\sqrt{31}} = 1064 \pm 19.4 = (1044.6, 1083.4)$$

We infer at the 95% confidence level that $1044.6 < \mu < 1083.4$. That is, with a high degree of certainty, the data indicates that the true mean tensile strength for this type of titanium alloy weld is between 1044.6 and 1083.4 MPa.   ■

Interpreting a Confidence Level
The confidence level 95% for the interval just defined was inherited from the probability .95 for the random interval (8.4). Intervals having other levels of confidence will be introduced shortly. For now, though, consider how 95% confidence can be interpreted.

We started with an event whose probability was .95—that the random interval (8.4) would capture the true value of μ—and then used the data in Example 8.1 to compute the CI (1044.6, 1083.4). It's therefore tempting to conclude that μ is between 1044.6 and 1083.4 with probability .95. But by substituting $\bar{x} = 1064$ for $\overline{X}$, all randomness disappears; the interval (1044.6, 1083.4) is not random, and neither is μ (while its value is unfortunately unknown to us, μ is still a constant). Thus it is *incorrect* to write $P(\mu$ lies in (1044.6, 1083.4)) = .95.

A correct interpretation of "95% confidence" relies on the long-run relative frequency interpretation of probability. To say that an event A has probability .95 is to say that if the experiment on which A is defined is performed over and over again, in the long run A will occur 95% of the time. Suppose we obtain another sample of tensile strength values and compute another 95% interval. Then we consider repeating this for a third sample, a fourth sample, and so on. Let A be the event that $\overline{X} - 1.96 \cdot \sigma/\sqrt{n} < \mu < \overline{X} + 1.96 \cdot \sigma/\sqrt{n}$. Since $P(A) = .95$, in the long run 95% of our computed CIs will contain μ. This is illustrated in Figure 8.3, where the vertical line cuts the measurement axis at the true (but unknown) value of μ. Notice that of the 11 intervals pictured, only intervals 3 and 11 fail to contain μ. In the long run, only 5% of all intervals so constructed would fail to contain μ.
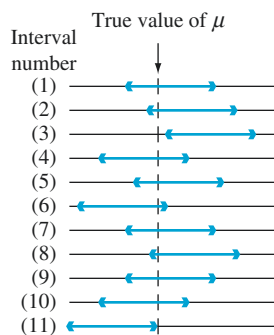
Figure 8.3 Repeated construction of 95% CIs

According to this interpretation, the confidence level 95% is not so much a statement about any particular interval such as (1044.6, 1083.4), but pertains to what would happen if a very large number of intervals were constructed using the same formula. Although this may seem unsatisfactory, the root of the difficulty lies with our interpretation of probability—it applies to a long sequence of

replications of an experiment, rather than just a single replication. There is another approach to the construction and interpretation of CIs that uses the notion of subjective probability and Bayes' theorem, as discussed in Chapter 15. The interval presented here (as well as each interval presented subsequently) is called a "classical" CI because its interpretation rests on the classical notion of probability (although the main ideas were developed as recently as the 1930s).

Other Levels of Confidence

The confidence level of 95% was inherited from the probability .95 for the initial inequalities in (8.2). If a confidence level of 99% is desired, the initial probability of .95 must be replaced by .99, which necessitates changing the z critical value in (8.2) from 1.96 to 2.576. A 99% CI then results from using 2.576 in place of 1.96 in the formula for the 95% CI.

This suggests that any desired level of confidence can be achieved by replacing 1.96 or 2.576 with the appropriate standard normal critical value. As Figure 8.4 shows, a probability of $1 - \alpha$ is achieved by using $z_{\alpha/2}$, which captures upper-tail area $\alpha/2$, in place of 1.96.
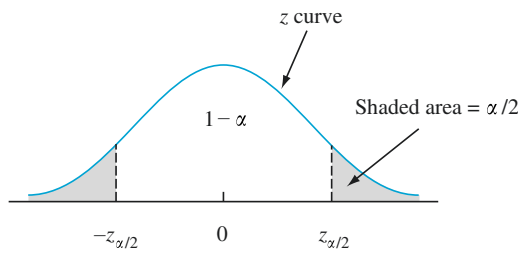
Figure 8.4 $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$

DEFINITION   A **$100(1 - \alpha)\%$ confidence interval** for the mean μ of a normal population when the value of σ is known is given by

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right) \tag{8.5}$$

or, equivalently, by $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

The z critical values for the most commonly used confidence levels are displayed in Table 8.1.

Table 8.1   Values of $z_{\alpha/2}$ for 90, 95, and 99% confidence

Confidence level (%)	α	$\alpha/2$	$z_{\alpha/2}$
90	.10	.05	1.645
95	.05	.025	1.960
99	.01	.005	2.576

Example 8.3 An introductory course has recently been changed, and the homework is now done online through a course management system instead of from the textbook exercises. How can we see if there has been improvement in student performance? Past experience suggests that the distribution of final exam scores under the old system was normally distributed with mean 65 and standard deviation 13. It is believed that the distribution is still normal with standard deviation 13, but the mean has potentially changed. A random sample of 40 students has a mean final exam score of 70.7. Let's calculate a confidence interval for the new population mean using a confidence level of 90%. From Table 8.1, the z critical value is $z_{\alpha/2} = z_{.05} = 1.645$. The desired interval is then

$$70.7 \pm 1.645 \cdot \frac{13}{\sqrt{40}} = 70.7 \pm 3.4 = (67.3, 74.1)$$

With 90% confidence, we can say that $67.3 < \mu < 74.1$, i.e., the true mean final exam score of all students using the new homework system will be between 67.3 and 74.1. In particular, at a confidence level of 90%, 65 is not a plausible value of μ. Thus we can be confident that the population mean has improved over the previous value of 65.   ∎

Confidence Level, Precision, and Choice of Sample Size

Why settle for a confidence level of 95% when a level of 99% is achievable? Because the price paid for the higher confidence level is a wider interval. The 95% interval extends $1.96 \cdot \sigma/\sqrt{n}$ to each side of $\bar{x}$, so the width of the interval is $2(1.96) \cdot \sigma/\sqrt{n} = 3.92 \cdot \sigma/\sqrt{n}$. Similarly, the width of the 99% interval is $2(2.576) \cdot \sigma/\sqrt{n} = 5.152 \cdot \sigma/\sqrt{n}$. That is, we have more confidence in the 99% interval precisely because it is wider. The higher the desired degree of confidence, the wider the resulting interval. In fact, the only 100% CI for μ is $(-\infty, \infty)$, which is not terribly informative because, even before sampling, we knew that this interval covers μ.

If we think of the width of the interval as specifying its precision (with narrower intervals being more precise), then the confidence level (or reliability) of the interval is inversely related to its precision. A highly reliable interval estimate may be imprecise in that the endpoints of the interval may be far apart, whereas a precise interval may possess relatively low reliability. Thus it cannot be said unequivocally that a 99% interval is to be preferred to a 95% interval; the gain in reliability entails a loss in precision.

An appealing strategy is to specify both the desired confidence level and interval width and then determine the necessary sample size.

Example 8.4 Extensive monitoring of a certain operating system has suggested that response time to a particular editing command is normally distributed with standard deviation 25 ms. A new operating system has been installed, and an estimate of the true average response time μ for the new environment is desired. Assuming that response times are still normally distributed with $\sigma = 25$, what sample size is necessary to ensure that the resulting 95% CI has a width of (at most) 10? The sample size n must satisfy

$$10 = 2 \cdot (1.96) \cdot (25/\sqrt{n})$$

Re-arranging this equation gives

$$\sqrt{n} = 2 \cdot (1.96) \cdot (25)/10 = 9.80$$

so

$$n = 9.80^2 = 96.04$$

Since n must be an integer, a sample size of 97 is required.   ∎

The general formula for the sample size n necessary to ensure an interval width w is obtained from $w = 2 \cdot z_{\alpha/2} \cdot \sigma/\sqrt{n}$ as

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{w/2} \right)^2 \tag{8.6}$$

The smaller the desired width w, the larger n must be. In addition, n is an increasing function of σ (more population variability necessitates a larger sample size) and also of the confidence level $100(1 - \alpha)\%$ (as α decreases, $z_{\alpha/2}$ increases).

The half-width $1.96 \cdot \sigma/\sqrt{n}$ of the 95% CI is sometimes called the **margin of error** associated with a 95% confidence level; that is, with 95% confidence, the point estimate $\bar{x}$ will be no farther than this from μ. Before obtaining data, an investigator may wish to determine a sample size for which a particular value of the margin of error is achieved. For example, with μ representing the average fuel efficiency (mpg) for all cars of a certain type, the objective of an investigation may be to estimate μ to within 1 mpg with 95% confidence. More generally, if we wish to estimate μ to within an amount b (the specified bound on the margin of error) with $100(1 - \alpha)\%$ confidence, the necessary sample size results from replacing $w/2$ by b in (8.6).

Deriving a General Confidence Interval

Let $X_1, X_2, \ldots, X_n$ denote the sample on which the CI for a parameter θ is to be based. The general strategy for deriving a CI relies on finding what's known as a *pivotal quantity*.

DEFINITION   Suppose a random variable satisfying the following two properties can be found:

1. The variable is a function of both $X_1, \ldots, X_n$ and θ.
2. The probability distribution of the variable does *not* depend on θ or on any other unknown parameters.

Such a random variable is called a **pivotal quantity**.

For example, if the population distribution is normal with σ known and $\theta = \mu$ unknown, the variable $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ in (8.1) satisfies both properties: (1) Z clearly depends functionally on the X_i's and μ, yet (2) Z has a $N(0, 1)$ distribution, which does not depend on μ. Hence Z is a pivotal quantity. In general, the form of a pivotal quantity is usually suggested by examining the distribution of an appropriate estimator $\hat{\theta}$.

Let $h(X_1, \ldots, X_n, \theta)$ denote a general pivotal quantity. For any α between 0 and 1, constants a and b can be found to satisfy

$$P(a < h(X_1, \ldots, X_n, \theta) < b) = 1 - \alpha \tag{8.7}$$

Critically, because of the second property of a pivotal quantity, a and b do not depend on θ. In the normal example, $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$. Now suppose the inequalities in (8.7) can be manipulated to isolate θ (typically, replace $<$ by $=$ and solve for θ), giving the equivalent statement

$$P(l(X_1, \ldots, X_n) < \theta < u(X_1, \ldots, X_n)) = 1 - \alpha$$

Then $l(x_1, \ldots, x_n)$ and $u(x_1, \ldots, x_n)$ are the lower and upper confidence limits, respectively, for a $100(1 - \alpha)\%$ CI. In the normal example, we saw that $l(X_1, \ldots, X_n) = \overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n}$ and $u(X_1, \ldots, X_n) = \overline{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

Example 8.5 A theoretical model suggests that the time-to-breakdown of an insulating fluid between electrodes at a particular voltage has an exponential distribution with unknown parameter λ (see Section 4.4). A random sample of $n = 10$ breakdown times yields the following sample data (in min): $x_1 = 41.53$, $x_2 = 18.73$, $x_3 = 2.99$, $x_4 = 30.34$, $x_5 = 12.33$, $x_6 = 117.52$, $x_7 = 73.02$, $x_8 = 223.63$, $x_9 = 4.00$, $x_{10} = 26.78$. A 95% CI for both λ and for the true average breakdown time are desired.

Let $h(X_1, X_2, \ldots, X_n, \lambda) = 2\lambda \Sigma X_i$. Using a moment generating function argument, it can be shown that this random variable has a chi-squared distribution (see Section 6.3) with $2n$ degrees of freedom. Since h is a function of both the X_i's and λ, yet its distribution χ_{2n}^2 does not depend on λ, it is a pivotal quantity.

Appendix Table A.5 pictures a typical chi-squared density curve and tabulates critical values that capture specified tail areas. The $\nu = 2n = 2(10) = 20$ row of the table shows that the .025 and .975 quantiles are 9.591 and 34.170, respectively. Thus for $n = 10$,

$$P(9.591 < 2\lambda \sum X_i < 34.170) = .95$$

Division by $2 \sum X_i$ isolates λ, yielding

$$P\left(9.591 / \left(2 \sum X_i\right) < \lambda < 34.170 / \left(2 \sum X_i\right)\right) = .95$$

The lower limit of the 95% CI for λ is $l = 9.591/(2\Sigma x_i)$, and the upper limit is $u = 34.170/(2\sum x_i)$. For the given data, $\sum x_i = 550.87$, giving the interval (.00871, .03101). Based on the data, we are 95% confident that the true value of the parameter λ is between .00871 and .03101.

The mean of an exponential rv is $\mu = 1/\lambda$. Since

$$P\left(2 \sum X_i / 34.170 < 1/\lambda < 2 \sum X_i / 9.591\right) = .95$$

the 95% CI for true average breakdown time is $(2\sum x_i/34.170, 2\sum x_i/9.591) = (32.24, 114.87)$. With 95% confidence, true mean breakdown time under these experimental conditions is between 32.24 and 114.87 min. This interval is obviously quite wide, reflecting substantial variability in breakdown times and a small-sample size. Notice also that the two endpoints are not equidistant from the point estimate; unlike in the normal case, here the CI for μ is not of the form $\overline{x} \pm c$. ∎

A General Large-Sample Confidence Interval

Let $X_1, X_2, \ldots, X_n$ be a random sample from any population having a mean μ and standard deviation σ. Provided that n is large, the Central Limit Theorem (CLT) implies that $\overline{X}$ has approximately a

normal distribution whatever the nature of the population distribution. It then follows that $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ has approximately a standard normal distribution, so that

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

An argument parallel with that given earlier in this section yields $\overline{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$ as a large-sample CI for μ with a confidence level of *approximately* $100(1 - \alpha)\%$. That is, when n is large, the CI (8.5) for μ remains valid whatever the population distribution (provided that the qualifier "approximately" is inserted in front of the confidence level).

The foregoing example is a special case of a general large-sample CI for a parameter θ. Suppose that $\hat{\theta}$ is an estimator satisfying the following properties:

1. $\hat{\theta}$ has approximately a normal distribution;
2. $\hat{\theta}$ is (at least approximately) unbiased for θ; and
3. an expression for $\sigma_{\hat{\theta}}$, the standard deviation of $\hat{\theta}$, is available.

For example, in the above discussion $\theta = \mu$, $\hat{\mu} = \overline{X}$ is an unbiased estimator whose distribution is approximately normal when n is large, and $\sigma_{\hat{\mu}} = \sigma_{\overline{X}} = \sigma/\sqrt{n}$. In Section 7.4, we saw that under very general conditions a maximum likelihood estimator $\hat{\theta}$ satisfies the first two properties when n is large, so what follows can be applied to many mles.

Standardizing $\hat{\theta}$ yields the rv $Z = (\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$, which has approximately a standard normal distribution, making Z an approximate pivotal quantity. This justifies the probability statement

$$P\left(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) \approx 1 - \alpha \tag{8.8}$$

from which a $100(1 - \alpha)\%$ CI for θ may potentially be obtained. How we proceed then depends on the formula for $\sigma_{\hat{\theta}}$.

Suppose first that $\sigma_{\hat{\theta}}$ does not involve any unknown parameters. Then replacing each $<$ by $=$ in (8.8) and solving for θ results in confidence limits $\hat{\theta} \pm z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$ for θ.

Next, suppose that $\sigma_{\hat{\theta}}$ doesn't involve θ itself but does involve at least one *other* unknown parameter. Let $s_{\hat{\theta}}$ be the estimate of $\sigma_{\hat{\theta}}$ obtained by using estimates in place of the unknown parameters, e.g., $s/\sqrt{n}$ estimates $\sigma/\sqrt{n}$. Under general conditions (essentially that $s_{\hat{\theta}}$ be close to $\sigma_{\hat{\theta}}$ for most samples), a valid CI for θ is then $\hat{\theta} \pm z_{\alpha/2} \cdot s_{\hat{\theta}}$. The interval $\overline{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$ is an example; we will encounter this interval in the next section.

Finally, suppose that $\sigma_{\hat{\theta}}$ involves the unknown θ itself. This is the case, for example, when $\theta = p$, a population proportion, as we'll see in Section 8.3. Then $(\hat{\theta} - \theta)/\sigma_{\hat{\theta}} = z_{\alpha/2}$ can be difficult to solve. An approximate solution can often be obtained by replacing θ in $\sigma_{\hat{\theta}}$ by its estimate $\hat{\theta}$. This results in an estimated standard deviation $s_{\hat{\theta}}$, and the corresponding interval is again $\hat{\theta} \pm z_{\alpha/2} \cdot s_{\hat{\theta}}$.

Example 8.6   A shipping company offers a flat fee for packages weighing up to 1 lb. Let $X_1, \ldots, X_n$ represent the weights (lb) of a random sample of packages ready for shipment, modeled by the pdf $f(x; \theta) = \theta x^{\theta-1}$ for $0 \le x \le 1$. The goal is to obtain a CI for the parameter θ.

From Example 7.37, the maximum likelihood estimator of θ is $\hat{\theta} = -n/\sum \ln(X_i)$, which for large n has approximately a normal distribution with mean θ and standard deviation $\theta/\sqrt{n}$. The preceding discussion then suggests a CI of the form $\hat{\theta} \pm z_{\alpha/2}\theta/\sqrt{n}$, but this is impractical—the standard error is a function of the unknown parameter θ itself. One solution is to solve the system of inequalities

$$-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\theta/\sqrt{n}} < z_{\alpha/2}$$

suggested by (8.8) for θ. Alternatively, we could substitute $\hat{\theta}$ for θ in the standard deviation formula, resulting in a CI with endpoints $\hat{\theta} \pm z_{\alpha/2}\hat{\theta}/\sqrt{n}$. That is, once the data is obtained and the value $\hat{\theta} = -n/\sum \ln(x_i)$ is calculated, that value of $\hat{\theta}$ is used twice to compute the CI.   ∎

One-Sided Confidence Intervals (Confidence Bounds)

The confidence intervals discussed thus far give both a lower confidence bound *and* an upper confidence bound for the parameter being estimated. In some circumstances, an investigator will want only one of these two types of bounds. For example, a psychologist may wish to calculate a 95% *upper* confidence bound for true average reaction time to a particular stimulus, or a surgeon may want only a *lower* confidence bound for true average remission time after colon cancer surgery.

In general, an **upper confidence bound** for a parameter θ with confidence level $100(1 - \alpha)\%$ based on a random sample $X_1, \ldots, X_n$ is a quantity $u(X_1, \ldots, X_n)$ such that

$$P(\theta < u(X_1, \ldots, X_n)) = 1 - \alpha$$

Similarly, a **lower confidence bound** $l(X_1, \ldots, X_n)$ satisfies $P(l(X_1, \ldots, X_n) < \theta) = 1 - \alpha$. As with two-sided confidence intervals, such bounds are evaluated by substituting the observed values $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. One-sided confidence bounds are often obtained by identifying a pivotal quantity and manipulating an appropriate inequality statement to isolate the parameter θ.

Example 8.7 Consider again the scenario of a random sample $X_1, \ldots, X_n$ from a normal distribution for which σ is known. Because the cumulative area under the standard normal curve to the left of 1.645 is .95,

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.645\right) = .95$$

Manipulating the inequality inside the parentheses to isolate μ on one side gives the inequality $\overline{x} - 1.645\sigma/\sqrt{n} < \mu$; the expression on the left is a lower confidence bound for μ. Applied to the data from Example 8.1, we obtain a 95% lower confidence bound of $1064 - 1.645 \cdot 55/\sqrt{31} = 1047.75$ MPa for the true average tensile strength.

Starting with $P(-1.645 < Z) = .95$ and manipulating the inequality results in an upper confidence bound. A similar argument gives a one-sided bound associated with any other confidence level.   ∎

Exercises: Section 8.1 (1–12)

1. Consider a normal population distribution with the value of σ known.

    a. What is the confidence level for the interval $\bar{x} \pm 2.81\sigma/\sqrt{n}$?

    b. What is the confidence level for the interval $\bar{x} \pm 1.44\sigma/\sqrt{n}$?

    c. What value of $z_{\alpha/2}$ in the CI Formula (8.5) results in a confidence level of 99.7%?

    d. Answer the question posed in part (c) for a confidence level of 75%.

2. Each of the following is a confidence interval computed from (8.5) for μ = true average (i.e., population mean) resonance frequency (Hz) for all tennis rackets of a certain type:
    (114.4, 115.6)   (114.1, 115.9)

    a. What is the value of the sample mean resonance frequency?

    b. Both intervals were calculated from the same sample data. The confidence level for one of these intervals is 90% and for the other is 99%. Which of the intervals has the 90% confidence level, and why?

3. Suppose that a random sample of 50 bottles of a particular brand of cough syrup is selected and the alcohol content of each bottle is determined. Let μ denote the average alcohol content for the population of all bottles of the brand under study. Suppose that the resulting 95% confidence interval is (7.8, 9.4).

    a. Would a 90% confidence interval calculated from this same sample have been narrower or wider than the given interval? Explain your reasoning.

    b. Consider the following statement: There is a 95% chance that μ is between 7.8 and 9.4. Is this statement correct? Why or why not?

    c. Consider the following statement: We can be highly confident that 95% of all bottles of this type of cough syrup have an alcohol content that is between 7.8 and

9.4. Is this statement correct? Why or why not?

    d. Consider the following statement: If the process of selecting a sample of size 50 and then computing the corresponding 95% interval is repeated 100 times, 95 of the resulting intervals will include μ. Is this statement correct? Why or why not?

4. A CI is desired for the true average stray-load loss μ (watts) for a certain type of induction motor when the line current is held at 10 amps for a speed of 1500 rpm. Assume that stray-load loss is normally distributed with $\sigma = 3.0$.

    a. Compute a 95% CI for μ when $n = 25$ and $\bar{x} = 58.3$.

    b. Compute a 95% CI for μ when $n = 100$ and $\bar{x} = 58.3$.

    c. Compute a 99% CI for μ when $n = 100$ and $\bar{x} = 58.3$.

    d. Compute an 82% CI for μ when $n = 100$ and $\bar{x} = 58.3$.

    e. How large must n be if the width of the 99% interval for μ is to be 1.0?

5. Assume that the helium porosity (in percentage) of coal samples taken from any particular seam is normally distributed with true standard deviation .75.

    a. Compute a 95% CI for the true average porosity of a certain seam if the average porosity for 20 specimens from the seam was 4.85.

    b. Compute a 98% CI for true average porosity of another seam based on 16 specimens with a sample average porosity of 4.56.

    c. How large a sample size is necessary if the width of the 95% interval is to be .40?

    d. What sample size is necessary to estimate true average porosity to within .2 with 99% confidence?

6. On the basis of extensive tests, the yield point of a particular type of mild steel reinforcing bar is known to be normally distributed with

$\sigma = 100$. The composition of the bar has been slightly modified, but the modification is not believed to have affected either the normality or the value of σ.

   a. Assuming this to be the case, if a sample of 25 modified bars resulted in a sample average yield point of 8439 lb, compute a 90% CI for the true average yield point of the modified bar.

   b. How would you modify the interval in part (a) to obtain a confidence level of 92%?

7. By how much must the sample size n be increased if the width of the CI (8.5) is to be halved? If the sample size is increased by a factor of 25, what effect will this have on the width of the interval? Justify your assertions.

8. Let $\alpha_1 > 0$, $\alpha_2 > 0$, with $\alpha_1 + \alpha_2 = \alpha$. Then

$$P\left(-z_{\alpha_1} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha_2}\right) = 1 - \alpha$$

   a. Use this equation to derive a more general expression for a $100(1 - \alpha)$% CI for μ of which the interval (8.5) is a special case.

   b. Let $\alpha = .05$ and $\alpha_1 = \alpha/4$, $\alpha_2 = 3\alpha/4$. Does this result in a narrower or wider interval than the interval (8.5)?

9. a. Generalize the method of Example 8.7 to obtain a lower bound for μ with a confidence level of $100(1 - \alpha)$%.

   b. Use part (a) to calculate a 99.5% confidence lower bound for the data in Exercise 5a.

   c. What is the analogous formula for a $100(1 - \alpha)$% confidence upper bound on μ? Compute this 99% upper bound for the data of Exercise 4a.

10. A random sample of $n = 15$ heat pumps of a certain type yielded the following observations on lifetime (in years):

2.0	1.3	6.0	1.9	5.1	.4	1.0	5.3
15.7	.7	4.8	.9	12.2	5.3	.6	

   a. Assume that the lifetime distribution is exponential and use an argument parallel to that of Example 8.5 to obtain a 95% CI for expected (true average) lifetime.

   b. How should the interval of part (a) be altered to achieve a confidence level of 99%?

   c. What is a 95% CI for the standard deviation of the lifetime distribution? [*Hint*: What is the standard deviation of an exponential random variable?]

11. Consider the next 1000 95% CIs for μ that a statistical consultant will obtain for various clients. Suppose the data sets on which the intervals are based are selected independently of one another. How many of these 1000 intervals do you expect to capture the corresponding value of μ? What is the probability that between 940 and 960 of these intervals contain the corresponding value of μ? [*Hint*: Let $Y =$ the number among the 1000 intervals that contain μ. What kind of random variable is Y?]

12. The superintendent of a large school district, having once had a course in probability and statistics, believes that the number of teachers absent on any given day has a Poisson distribution with parameter μ. Use the accompanying data on absences for 50 days to derive a large-sample CI for μ. [*Hint*: The mean and variance of a Poisson variable both equal μ, so

$$Z = \frac{\overline{X} - \mu}{\sqrt{\mu/n}}$$

has approximately a standard normal distribution and is thus a pivotal quantity. Now proceed as in Example 8.6.]

Number of absences	0	1	2	3	4	5	6	7	8	9	10
Frequency	1	4	8	10	8	7	5	3	2	1	1

8.2   The One-Sample *t* Interval and Its Relatives

The CI for μ given in the previous section assumed that the population distribution is normal with the value of σ known. The derivation of the interval relied on the pivotal quantity $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ in (8.1), which has a standard normal distribution under these assumptions. In this section, we will construct a CI for μ for the more realistic situation when σ is unknown; this is the interval estimate used in practice.

Consider the variable obtained by replacing σ in Z by the *sample* standard deviation S. Define a new random variable T by

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{8.9}$$

It is important to contrast the behavior of Z in repeated sampling with that of T (this is really a refresher from Section 6.4). The only variability in Z from one sample to another is because the value of $\overline{X}$ in the numerator varies in value. However, there are two sources of sample-to-sample variability in T: both $\overline{X}$ in the numerator and S in the denominator. Because of this extra variation in T, it stands to reason that the distribution of T should be more spread out than that of Z. That is, the density curve for T should be more spread out than the standard normal curve.

The One-Sample *t* Confidence Interval

Suppose that $X_1, \ldots, X_n$ is a random sample from a *normal* population distribution. Then Gosset's Theorem from Section 6.4 states that the rv T in (8.9) follows a *t* distribution with $n - 1$ degrees of freedom (df). Properties of the *t* family of distributions were detailed in Section 6.3; for now, it suffices to recall that the *t* distribution with v df has a symmetric, bell-shaped density curve centered at 0 that is wider than a standard normal curve but converges to the standard normal curve as $v \to \infty$ (so the z curve may be thought of as the *t* curve with df $= \infty$). See Figure 6.16 for an illustration. Recall also the notation for values that capture particular upper-tail *t* curve areas.

NOTATION   Let $t_{\alpha,v}$ = the number on the measurement axis for which the area under the *t* curve with v df to the right of $t_{\alpha,v}$ is α; $t_{\alpha,v}$ is called a **t critical value**.

This notation is illustrated in Figure 8.5. Appendix Table A.6 gives $t_{\alpha,v}$ for selected values of α and v. The columns of the table correspond to different values of α. To obtain $t_{.05,15}$, go to the $\alpha = .05$ column, look down to the $v = 15$ row, and read $t_{.05,15} = 1.753$. Similarly, $t_{.05,22} = 1.717$ (.05 column, $v = 22$ row), and $t_{.01,22} = 2.508$. Statistical software packages can provide *t* critical values for any specified tail area and df; for example, $t_{\alpha,v}$ can be obtained in R with the command $\mathtt{qt\,(1 - \alpha,\ v)}$.
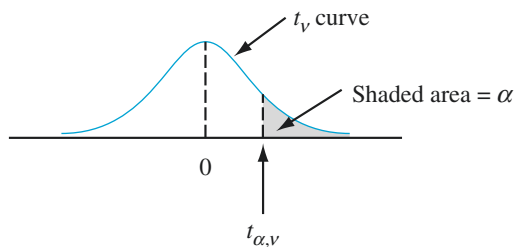
Figure 8.5   A pictorial definition of $t_{\alpha,v}$

The values of $t_{\alpha,v}$ exhibit regular behavior as we move across a row or down a column. For fixed v, $t_{\alpha,v}$ increases as α decreases, since we must move farther to the right of zero to capture area α in the tail. For fixed α, as v is increased (i.e., as we look down any particular column of the t table) the value of $t_{\alpha,v}$ decreases. This is because a larger value of v implies a t distribution with smaller spread, so it is not necessary to go so far from zero to capture tail area α. Furthermore, $t_{\alpha,v}$ decreases more slowly as v increases. Consequently, the table values are shown in increments of 2 between 30 and 40 df and then jump to $v = 50$, 60, 120, and finally ∞. Because t_∞ is the standard normal curve, the familiar z_α values appear in the last row of the table.

Now let's obtain the desired confidence interval. The pivotal quantity T in (8.9) has a t_{n-1} distribution, and the area under the corresponding t density curve between $-t_{\alpha/2,n-1}$ and $t_{\alpha/2,n-1}$ is $1 - \alpha$ (area $\alpha/2$ lies in each tail), so

$$P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1 - \alpha \tag{8.10}$$

Expression (8.10) differs from similar expressions in Section 8.1 in that T and $t_{\alpha/2,n-1}$ are used in place of Z and $z_{\alpha/2}$, but it can be manipulated in the same manner to obtain a confidence interval for μ.

PROPOSITION   Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from the results of a random sample from a *normal* population with mean μ. Then a **100(1 − α)% confidence interval for μ**, also called the **one-sample t CI**, is

$$\left(\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \right) \tag{8.11}$$

or, more compactly, $\bar{x} \pm t_{\alpha/2,n-1} \cdot s/\sqrt{n}$.

An **upper confidence bound for μ** is

$$\bar{x} + t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}$$

and replacing + by − in this latter expression gives a **lower confidence bound for μ**; both have confidence level $100(1 - \alpha)\%$.

Example 8.8 Have you ever dreamed of owning a Porsche? Even though academic salaries leave little room for luxuries, the authors thought maybe the purchase of a used Boxster, the least expensive Porsche model, might be feasible. So on July 15, 2019 we went to www.cars.com to peruse prices. The news was discouraging, so we instead selected a random sample of 16 such vehicles and obtained the following odometer readings (miles):

80,000	30,100	97,500	58,551	73,787	51,800	69,267	44,530
42,192	104,920	41,442	27,418	43,436	77,219	5991	14,362

Figure 8.6 shows a normal probability plot of the data; this version includes a superimposed line which makes it easier to judge whether the pattern in the plot is reasonably linear. Very clearly that is the case. It is therefore quite plausible that the distribution of odometer readings is (at least approximately) normal, which validates the use of the one-sample t confidence interval to estimate the population mean odometer reading, μ.

Figure 8.6 Normal probability plot of the Boxster odometer reading data
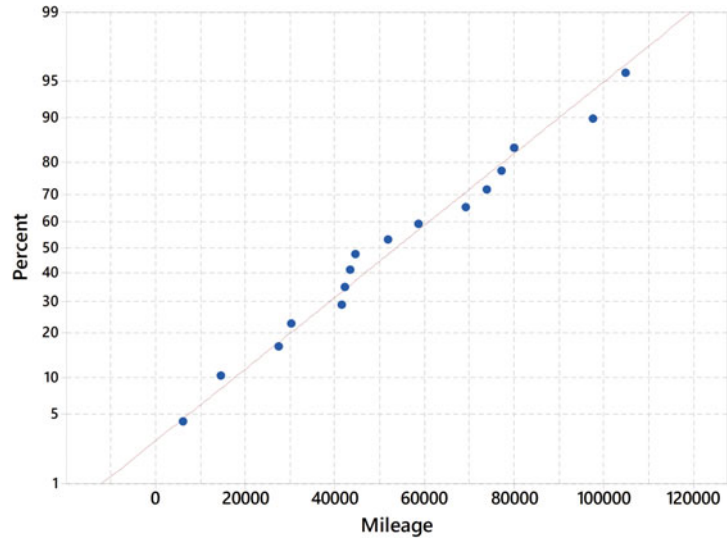
The sample mean and standard deviation are 53,907.2 and 28,287.2, respectively, and the (estimated) standard error of the mean is $s/\sqrt{n} = 7071.8$. Table A.6 shows that the t critical value for a confidence level of 95% when df $= 16 - 1 = 15$ is $t_{.025,15} = 2.131$. The confidence interval is then

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} = 53{,}907.2 \pm (2.131)(7071.8) = 53{,}907.2 \pm 15{,}070.0$$
$$= (38{,}837.2,\ 68{,}977.2)$$

That is, we can say with a confidence level of 95% that $38{,}837.2 < \mu < 68{,}977.2$. This CI is quite wide, indicating that our knowledge of μ is imprecise.

Remember that it is *not correct* at this point to write $P(38{,}837.2 < \mu < 68{,}977.2) = .95$, because nothing inside the parentheses is random. The interval we have calculated may or may not include the actual value of μ. If we were to obtain sample after sample of size 16 from this population and for each one use (8.11) with $t = 2.131$, in the long run 95% of the calculated CIs would include μ whereas 5% would not. Without knowing the value of μ, we can't know whether the *particular* interval we have calculated is one of the "good" 95% or the "bad" 5%. ∎

Gosset's Theorem and the resulting one-sample t CI (8.11) assume a normal population distribution, which can be validated using a normal probability plot. Thankfully, the one-sample t CI for μ is robust to small or even moderate departures from normality unless n is quite small. By "robust," we mean that if a t critical value for 95% confidence is used in calculating the interval, the actual confidence will be reasonably close to the nominal 95% level, and similarly for other confidence levels. As a result, many practitioners use (8.11) if *either* the population distribution is plausibly normal *or* the sample size is "large"—$n \geq 40$ is a popular criterion.

It's worth noting that if the sample size is large, whether we use a t or z critical value does not make much practical difference. Thanks to the Central Limit Theorem, the random variable $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ has an approximately standard normal distribution when n is large; simultaneously, S is highly likely to be close to σ, suggesting that T in (8.9) is also approximately normal. Thus, for large n, one may apply the CI formula

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \right) \tag{8.12}$$

in lieu of (8.11). A large-sample upper confidence bound for μ results from replacing $z_{\alpha/2}$ with z_α in the upper limit of the interval (8.12); an analogous lower bound is obtained from the same replacement made to the lower limit of (8.12).

Example 8.9 A survey published by Gallup (Nov. 1, 2019) of 1526 adults asked how much each person planned to "personally spend on Christmas gifts" in 2019. The mean response was \$942 with a standard deviation of \$1116. Clearly the distribution of planned expenses is strongly positively skewed (the standard deviation exceeds the mean); nevertheless, let's calculate an (approximate) 99% CI for μ, the mean amount all US adults planned to spend on Christmas presents in 2019.

Because $n = 1526$ is very large, either Expression (8.11) or (8.12) is appropriate, even though the population distribution is nonnormal. The z and t critical values are $z_{.005} = 2.576$ and $t_{.005,1525} = 2.579$, so the resulting CIs will be essentially identical. Using the latter, the resulting CI is

$$942 \pm 2.579 \cdot \frac{1116}{\sqrt{1526}} = 942 \pm 73.7 = (868.3, 1015.7)$$

At the 99% confidence level, we conclude that the average amount US adults planned to personally spend on Christmas gifts in 2019 was between \$868.30 and \$1015.70.   ∎

Sample Size Determination
In Section 8.1, we considered the problem of determining the sample size required to achieve a certain level of precision at a prescribed confidence level. Under the assumptions of that section, we derived the formula

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{w/2} \right)^2$$

for the minimum sample size necessary to place an upper bound w on the width of the interval. Given the discussion in this section, it might seem like the natural update to this formula is

$$n = \left(\frac{t_{\alpha/2,n-1} \cdot s}{w/2} \right)^2 \tag{8.13}$$

where s is the sample standard deviation. However, this formula presents two practical problems.

First, sample size determination typically occurs *before* a study is carried out, in which case the researcher doesn't yet have a value for s. This can be addressed by using a sample standard deviation from a previous, similar study, though that assumes variability has not changed significantly. Another, more conservative method is to use range/4 as a crude estimate of the standard deviation; this is "conservative" in the sense that for many distributions range/4 exceeds s, so use of this estimate in (8.13) typically returns a somewhat larger n than needed. The range/4 formula has the advantage that even in the absence of reliable data from which to calculate s, the range of potential values from a particular process is typically easier to "guess."

Second, n now appears on both sides of the equation: we need to know n before finding the t critical value, which then determines the sample size n on the left-hand side of (8.13). Technology can solve this problem: many statistical software packages will search iteratively to find the smallest n which satisfies (8.13). Alternatively, you can simply replace $t_{\alpha/2,n-1}$ with $z_{\alpha/2}$ in (8.13) and solve for n, but the result will be slightly smaller than the sample size actually required. Notice, though, that even software still requires a value (or an estimate) for s.

Example 8.10 A supplier of carbon-ceramic brake disks for high-performance cars has recently redesigned its manufacturing process. The company needs to estimate, among other things, the true mean density μ of their new ceramic material. Data from the previous process suggests that the standard deviation for ceramic density is around 0.2 g/cm^3. Assuming this value still approximately holds for the new process, how large a sample will the company require to obtain a 99% CI for μ no wider than 0.1 g/cm^3? Apply (8.13) with $s = 0.2$, $w = 0.1$, and $t_{.005,n-1} \approx z_{.005} = 2.576$:

$$n \approx \left(\frac{2.576(0.2)}{0.1/2} \right)^2 = 106.17$$

Since n must be an integer, a minimum of 107 ceramic specimens is required. As noted previously, this will be a slight underestimate, because we used 2.576 in place of the unknown t critical value. Statistical software, which does not use this approximation, indicates that at least 110 specimens are required. ■

A Prediction Interval for a Single Future Value

In many applications, an investigator wishes to predict a *single* value of a variable to be observed at some future time, rather than to estimate the *mean* value of that variable.

Example 8.11 Scientists worldwide routinely monitor the general health of forests, and engineers investigate mechanical properties of various wood types. Consider the following core wood density measurements (g/mm^3) from a sample of 25 canopy trees in western Thailand ("Radial Variation of Wood Functional Traits Reflect Size-Related Adaptations of Tree Mechanics and Hydraulics," *Functional Ecology* 2017):

391.2	431.0	447.1	375.3	470.7	543.7	592.7	546.7	601.8	598.8
492.3	454.4	548.7	494.9	585.6	647.8	639.2	700.4	640.1	620.5
755.2	668.7	644.6	717.7	663.0					

Figure 8.7 shows a normal probability plot from R software. The straightness of the pattern provides support for assuming that core wood density measurements in this population are at least approximately normal.

The sample mean and standard deviation are $\bar{x} = 570.9$ g/mm^3 and $s = 103.9$ g/mm^3, respectively. A 95% CI for μ = the population mean core wood density is

$$\bar{x} \pm t_{.025,24} \cdot \frac{s}{\sqrt{n}} = 570.9 \pm 2.064 \cdot \frac{103.9}{\sqrt{25}} = 570.9 \pm 42.9$$
$$= (528.0, 613.8)$$
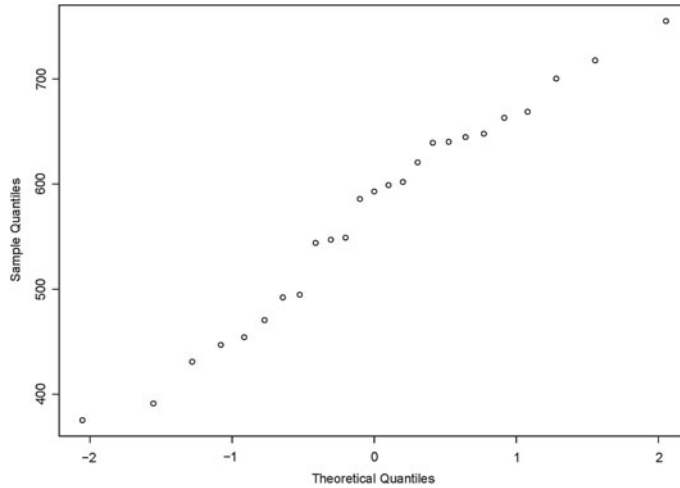
Figure 8.7 Normal probability plot for the wood density data of Example 8.11

This is fine if researchers are interested in *average* properties of these trees. But what about a single tree from this forest—what should we *predict* for its core wood density? A *point prediction*, analogous to a point estimate, is just $\bar{x} = 570.9$ g/mm^3. But this prediction unfortunately gives no information about reliability or precision. A different type of interval is required to make inferences about the density of an individual wood specimen. ∎

The general scenario is as follows: We have available a random sample X_1, X_2, …, X_n from a normal population distribution, and we wish to predict the value of X_{n+1}, a single future observation. A point predictor is $\overline{X}$, and the resulting prediction error is $\overline{X} - X_{n+1}$. The expected value of the prediction error is

$$E(\overline{X} - X_{n+1}) = E(\overline{X}) - E(X_{n+1}) = \mu - \mu = 0$$

Since X_{n+1} is independent of X_1, …, X_n, it is independent of $\overline{X}$, so the variance of the prediction error is

$$V(\overline{X} - X_{n+1}) = V(\overline{X}) + V(X_{n+1}) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2 \left(1 + \frac{1}{n}\right)$$

The prediction error is a linear combination of independent normally distributed rvs, so itself is normally distributed. Thus

$$Z = \frac{(\overline{X} - X_{n+1}) - 0}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}} = \frac{\overline{X} - X_{n+1}}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}}$$

has a standard normal distribution. As in the derivation of the distribution of $(\overline{X} - \mu)/(S/\sqrt{n})$ in Section 6.4, it can be shown (Exercise 42) that replacing σ by the sample standard deviation S (of $X_1, \ldots, X_n$) results in

$$T = \frac{\overline{X} - X_{n+1}}{S\sqrt{1 + \frac{1}{n}}} \sim t \text{ distribution with } n - 1 \text{ df}$$

Manipulating this T variable using Expression (8.10) to isolate X_{n+1} gives the following result.

PROPOSITION   A **prediction interval (PI)** for a single observation to be selected from a normal population distribution is

$$\overline{x} \pm t_{\alpha/2, n-1} \cdot s \sqrt{1 + \frac{1}{n}} \tag{8.14}$$

The *prediction level* is $100(1 - \alpha)\%$.

The interpretation of a 95% prediction level is similar to that of a 95% confidence level: if the interval (8.14) is calculated for sample after sample, in the long run 95% of these intervals will include the corresponding future values of X.

Example 8.12 (Example 8.11 continued) With $n = 25$, $\overline{x} = 570.9$, $s = 103.9$, and $t_{.025,24} = 2.064$, a 95% PI for the core wood density of a single tree in this Thai forest is

$$\overline{x} \pm t_{.025,24} \cdot s \sqrt{1 + \frac{1}{n}} = 570.9 \pm (2.064)(103.9)\sqrt{1 + \frac{1}{25}} = 570.9 \pm 218.7$$
$$= (352.2, 789.6)$$

So, with 95% certainty, we predict that the core wood density of an as-yet-unmeasured tree in this forest will be between 352.2 and 789.6 g/mm^3. This PI is quite wide—more than five times as wide as the previous CI—a reflection of substantial variability in wood density in this forest. ∎

It's worth contrasting the behavior of the one-sample *t* CI (8.11) with the PI (8.14). The PI is wider than the CI because there is more variability in the prediction error (due to X_{n+1}) than in the estimation error. In fact, as n gets arbitrarily large, the CI shrinks to the single value μ, while the PI approaches $\mu \pm z_{\alpha/2} \cdot \sigma$, an interval that covers the middle $100(1 - \alpha)\%$ of a normal distribution. That's as it should be: there is uncertainty about a single future X value even when there is no need to estimate any parameters.

Tolerance Intervals

In addition to confidence intervals and prediction intervals, statisticians are sometimes called upon to obtain a third type of interval called a *tolerance interval* (TI). A TI is an interval that with a high degree of reliability captures a specified percentage of the population distribution. For example, if the population distribution of women's heights is normal, then the interval from $\mu - 1.645\sigma$ to

$\mu + 1.645\sigma$ captures 90% of the height values in the population of women. It can then be shown that if μ and σ are replaced by their natural estimates $\bar{x}$ and s based on a sample of size $n = 20$ and the z critical value 1.645 is replaced by a *tolerance critical value* 2.310, the resulting interval contains at least 90% of the population values with a confidence level of 95%.

Please consult one of the references for more information on TIs. And before you calculate a particular statistical interval, be sure that it is the correct type of interval to fulfill your objective!

Intervals Based on Nonnormal Population Distributions

As mentioned previously, the one-sample t CI for μ is robust to small or even moderate departures from normality unless n is quite small. If, however, n is small and the population distribution is highly nonnormal, then your *actual* confidence level may be considerably different from the one you *think* you get from using a particular t critical value. It would certainly be distressing to believe that your confidence level is about 95% when in fact it was really more like 88% (or worse)! The *bootstrap* technique, discussed in the last section of this chapter, has been found to be quite successful at estimating parameters in a wide variety of nonnormal situations.

In contrast to the confidence interval, the validity of the prediction interval described in this section is closely tied to the normality assumption. The prediction interval (8.14) should not be used in the absence of compelling evidence for normality. The excellent reference *Statistical Intervals* by Meeker et al., cited in the bibliography, discusses alternative procedures of this sort for various other situations.

Exercises: Section 8.2 (13–42)

13. Determine the values of the following quantities:
    a. $t_{.1,15}$ b. $t_{.05,15}$ c. $t_{.05,25}$ d. $t_{.05,40}$ e. $t_{.005,40}$

14. Determine the t critical value that will capture the desired t curve area in each of the following cases:
    a. Central area = .95, df = 10
    b. Central area = .95, df = 20
    c. Central area = .99, df = 20
    d. Central area = .99, df = 50
    e. Upper-tail area = .01, df = 25
    f. Lower-tail area = .025, df = 5

15. Determine the t critical value for a two-sided confidence interval in each of the following situations:
    a. Confidence level = 95%, df = 10
    b. Confidence level = 95%, df = 15
    c. Confidence level = 99%, df = 15
    d. Confidence level = 99%, $n = 5$
    e. Confidence level = 98%, df = 24
    f. Confidence level = 99%, $n = 38$

16. Determine the t critical value for a lower or an upper confidence bound for each of the situations described in the previous exercise.

17. Here are the alcohol percentages for a random sample of 16 beers (light beers excluded):

4.68	4.13	4.80	4.63	5.08	5.79	6.29	6.79
4.93	4.25	5.70	4.74	5.88	6.77	6.04	4.95

    a. Construct a normal probability plot of the data. Is it plausible that these values represent a sample from a normal distribution?
    b. Calculate a 95% CI for the mean alcohol percentage of all nonlight beers.
    c. Calculate a 95% CI for the mean amount of alcohol, in ounces, in a 12-oz. serving of (again, nonlight) beer.

18. A random sample of 50 patients who had been seen at an outpatient clinic was selected, and the waiting time to see a

physician was determined for each one, resulting in a sample mean time of 40.3 min and a sample standard deviation of 28.0 min (suggested by the article "An Example of Good but Partially Successful OR Engagement: Improving Outpatient Clinic Operations," *Interfaces* 28, #5).

a. Calculate and interpret a 95% upper confidence bound for true average waiting time.

b. Based on the sample mean and standard deviation, why is it doubtful that the population of waiting times is normally distributed? Does that invalidate the confidence bound you calculated in part (a)?

19. Exercise 16 of Chapter 1 presented data on the noise level (dBA) experienced by a sample of 77 individuals working at a particular office.

a. Construct a 95% confidence interval for the true average noise level experienced by people working in this office.

b. Would it be feasible to construct a valid 95% PI for the noise level experienced by a single office worker? Why or why not?

20. According to a study published in the *Calgary Herald* (Sep. 17, 2005), the average daily commute time for workers in Calgary is 28.5 min with a standard deviation of 24.2 min. The survey respondents constituted a random sample of 500 adults living in Calgary.

a. Construct and interpret a 99% CI for the true average daily commute time of all adults living in Calgary.

b. Calculate a 99% CI for the average *weekly* commute time of this population.

21. An article in *Issues in Accounting Education* reported on the job-changing habits of individuals who started at a Big Eight accounting firm. In a random sample of 44 such people who subsequently changed jobs, the sample mean time to change was 35.02 months with a standard deviation of 18.94 months.

a. Construct and interpret a 95% CI for the true average time to change jobs for this population.

b. Construct a 95% PI for the time to change jobs for a randomly selected person starting at a Big Eight accounting firm. Are there any extra assumptions you must make for this interval to be valid? Do those assumptions seem credible here?

22. Frontier Airlines conducted a study of passenger weights, including carry-on items (*Alaska J. Commerce*, May 25, 2003). They found an average summer weight of 183 lbs and an average winter weight of 190 lbs. Suppose that both of these surveys were based on random samples of 90 people and that the sample standard deviations for the summer and winter groups were 25 and 28, respectively.

a. Construct and interpret a 95% CI for true average passenger weight (including carry-ons) during the summer for Frontier Airlines.

b. Repeat part (a) for the winter sample.

c. Federal Aviation Administration (FAA) guidelines state that typical passenger weight should be 190 lbs in the summer and 195 lbs in the winter. Based on the confidence intervals in parts (a) and (b), do Frontier Airlines passengers appear to meet FAA recommendations? Explain.

23. Consider the following sample of fat content (in percentage) of $n = 10$ randomly selected hot dogs ("Sensory and Mechanical Assessment of the Quality of Frankfurters," *J. Texture Stud.* 1990: 395–409):

25.2 21.3 22.8 17.0 29.8 21.0 25.5 16.0 20.9 19.5

Assume that these were selected from a normal population distribution.

a. Compute a 95% confidence interval for the population mean fat content.

b. Would a 90% CI be wider or narrower than the interval you computed in (a)?

c. Would a 95% CI based on a sample of $n = 20$ hot dogs be wider or narrower than the interval you computed in (a)?

d. Calculate a 95% PI for the fat content of a single hot dog.

24. Here is a sample of ACT scores (average of the Math, English, Social Science, and Natural Science scores) for students taking college freshman calculus:

24.00	28.00	27.75	27.00	24.25	23.50	26.25
24.00	25.00	30.00	23.25	26.25	21.50	26.00
28.00	24.50	22.50	28.25	21.25	19.75	

a. Using an appropriate graph, see if it is plausible that the observations were selected from a normal distribution.

b. Calculate a two-sided 95% confidence interval for the population mean.

c. The university ACT average for entering freshmen that year was about 21. Are the calculus students better than average, as measured by the ACT?

25. A sample of 14 joint specimens of a particular type gave a sample mean proportional limit stress of 8.48 MPa and a sample standard deviation of .79 MPa ("Characterization of Bearing Strength Factors in Pegged Timber Connections," *J. Struct. Engr.* 1997: 326–332).

a. Calculate and interpret a 95% lower confidence bound for the true average proportional limit stress of all such joints. What, if any, assumptions did you make about the distribution of proportional limit stress?

b. Calculate and interpret a 95% lower prediction bound for the proportional limit stress of a single joint of this type.

26. Even as traditional markets for sweetgum lumber have declined, large section solid timbers traditionally used for bridge construction have become increasingly scarce. The article "Development of Novel Industrial Laminated Planks from Sweetgum Lumber" (*J. of Bridge Engr.* 2008: 64–66) described the manufacturing and testing of composite beams designed to add value to low-grade sweetgum lumber. Here is data on the modulus of rupture (psi; the article contained summary data expressed in MPa):

6807.99	7637.06	6663.28	6165.03	6991.41	6992.23
6981.46	7569.75	7437.88	6872.39	7663.18	6032.28
6906.04	6617.17	6984.12	7093.71	7659.50	7378.61
7295.54	6702.76	7440.17	8053.26	8284.75	7347.95
7422.69	7886.87	6316.67	7713.65	7503.33	7674.99

a. Verify the plausibility of assuming a normal population distribution.

b. Estimate the true average modulus of rupture in a way that conveys information about precision and reliability.

c. Predict the modulus for a single beam in a way that conveys information about precision and reliability. How does the resulting prediction compare to the estimate in (b)?

27. The $n = 26$ observations on escape time given in Exercise 46 of Chapter 1 give a sample mean and sample standard deviation of 370.69 s and 24.36 s, respectively. Assume the population distribution of escape times is at least approximately normal.

a. Calculate an upper confidence bound for population mean escape time using a confidence level of 95%.

b. Calculate an upper prediction bound for the escape time of a single additional worker using a prediction level of 95%. How does this bound compare with the confidence bound of part (a)?

c. Suppose that two additional workers will be chosen to participate in the

simulated escape exercise. Denote their escape times by X_{27} and X_{28}, and let $\overline{X}_{\text{new}}$ denote the average of these two values. Modify the formula for a PI for a single *x* value to obtain a PI for $\overline{X}_{\text{new}}$, and calculate a 95% two-sided interval based on the given escape data.

28. A study of the ability of individuals to walk in a straight line ("Can We Really Walk Straight?" *Amer. J. Phys. Anthropol.* 1992: 19–27) reported the accompanying data on cadence (strides per second) for a sample of $n = 20$ randomly selected healthy men.

.95 .85 .92 .95 .93 .86 1.00 .92 .85 .81
.78 .93 .93 1.05 .93 1.06 1.06 .96 .81 .96

A normal probability plot gives substantial support to the assumption that the population distribution of cadence is approximately normal.

a. Calculate and interpret a 95% confidence interval for population mean cadence.
b. Calculate and interpret a 95% prediction interval for the cadence of a single individual randomly selected from this population.

29. Return to the odometer reading scenario of Example 8.8. Calculate a prediction for an additional Boxster's odometer reading in a way that provides information about precision and reliability. The authors actually selected a 17th such vehicle and found its odometer reading to be 19,815. Is that consistent with your prediction?

30. Exercise 85 of Chapter 1 gave the following observations on a receptor binding measure (adjusted distribution volume) for a sample of 13 healthy individuals: 23, 39, 40, 41, 43, 47, 51, 58, 63, 66, 67, 69, 72.

a. Is it plausible that the population distribution from which this sample was selected is normal?
b. Predict the adjusted distribution volume of a single healthy individual by calculating a 95% prediction interval.

31. Here are the lengths (in minutes) of the 63 nine-inning games from the first week of the 2001 major league baseball season:

194 160 176 203 187 163 162 183 152 177
177 151 173 188 179 194 149 165 186 187
187 177 187 186 187 173 136 150 173 173
136 153 152 149 152 180 186 166 174 176
198 193 218 173 144 148 174 163 184 155
151 172 216 149 207 212 216 166 190 165
176 158 198

Assume that this is a random sample of nine-inning games (the mean differs by 12 s from the mean for the whole season).

a. Give a 95% confidence interval for the population mean.
b. Give a 95% prediction interval for the length of the next nine-inning game. On the first day of the next week, Boston beat Tampa Bay 3–0 in a nine-inning game of 152 min. Is this within the prediction interval?
c. Compare the two intervals and explain why one is much wider than the other.
d. Explore the issue of normality for the data and explain how this is relevant to parts (a) and (b).

32. A more extensive tabulation of *t* critical values than what appears in this book shows that for the *t* distribution with 20 df, the areas to the right of the values .687, .860, and 1.064 are .25, .20, and .15, respectively. What is the confidence level for each of the following three confidence intervals for the mean μ of a normal population distribution? Which of the three intervals would you recommend be used, and why?

a. $(\bar{x} - .687s/\sqrt{21}, \bar{x} + 1.725s/\sqrt{21})$

b. $(\bar{x} - .860s/\sqrt{21}, \bar{x} + 1.325s/\sqrt{21})$

c. $(\bar{x} - 1.064s/\sqrt{21}, \bar{x} + 1.064s/\sqrt{21})$

33. The following data on distilled alcohol content (%) for a sample of 35 port wines was extracted from the article "A Method for the Estimation of Alcohol in Fortified Wines Using Hydrometer Baumé and Refractometer Brix" (*Amer. J. Enol. Vitic.* 2006: 486–490):

```
16.35  18.85  16.20  17.75  19.58  17.73  22.75  23.78  23.25
19.08  19.62  19.20  20.05  17.85  19.17  19.48  20.00  19.97
17.48  17.15  19.07  19.90  18.68  18.82  19.03  19.45  19.37
19.20  18.00  19.60  19.33  21.22  19.50  15.30  22.25
```

a. Calculate and interpret a 99% confidence interval for the population mean.

b. Calculate and interpret a 90% lower confidence bound for the population mean.

c. It would be of interest to winemakers to obtain a prediction interval for the alcohol content of an individual port wine. Why should we hesitate to apply the PI formula (8.14) to this data? [*Hint:* If you haven't done so already, make a graph.]

34. The article "Evaluating Tunnel Kiln Performance" (*Amer. Ceramic Soc. Bull.*, Aug. 1997: 59–63) gave the following summary information for fracture strengths (MPa) of $n = 169$ ceramic bars fired in a particular kiln: $\bar{x} = 89.10, \ s = 3.73$.

a. Calculate a (two-sided) confidence interval for true average fracture strength using a confidence level of 95%. Does it appear that true average fracture strength has been precisely estimated?

b. Suppose the investigators had believed a priori that the population standard deviation was about 4 MPa. Based on

this supposition, how large a sample would have been required to estimate μ to within .5 MPa with 95% confidence?

35. As health care costs rise and health care systems worldwide become overburdened, patients are made to wait longer for critical procedures. One Canadian study of 539 cardiac patients waiting for cardiac bypass surgery found a mean wait time of 19 days with a standard deviation of ten days ("Wait Times Data Guide," Ministry of Health and Long-Term Care, Ontario, Canada, 2006; *wait time* is measured from the date a patient was recommended for surgery to the date surgery was performed). Assuming the data can be considered representative of the Ontario population, construct a 90% CI for the true mean wait time for bypass surgery in Ontario.

36. A sample of 66 obese adults was put on a low-carbohydrate diet for a year. The average weight loss was 11 lb and the standard deviation was 19 lb. Calculate a 99% lower confidence bound for the true average weight loss. What does the bound say about confidence that the mean weight loss is positive?

37. A study was done on 41 first-year medical students to see if their anxiety levels changed during the first semester. One measure used was the level of serum cortisol, which is associated with stress. For each of the 41 students the level was compared during finals at the end of the semester against the level in the first week of classes. The average difference (end of semester minus beginning) was +2.08 with a standard deviation of 7.88. Find a 95% lower confidence bound for the population mean difference μ. Does the bound suggest that the mean population stress change is necessarily positive?

38. The article "Ultimate Load Capacities of Expansion Anchor Bolts" (*J. Energy Engr.* 1993: 139–158) gave the following summary data on shear strength (kip) for a sample of 3/8-in. anchor bolts: $n = 78$, $\bar{x} = 4.25$, $s = 1.30$. Calculate a lower confidence bound using a confidence level of 90% for true average shear strength.

39. University administrators wish to estimate the mean time to graduation, for the population of students who actually graduate, to within ±3 months (one-quarter). It is known that the maximum time to graduation is eight years (96 months) and the minimum time is three years (36 months), so that a conservative estimate of the standard deviation of graduation times is range/4 = (96 − 36)/4 = 15 months. Use this standard deviation estimate to determine the sample size required to achieve the administrators' goal with 95% confidence. [*Note:* 3 months is not the desired interval width; it's the target margin of error.]

40. Young people may feel they are carrying the weight of the world on their shoulders, when what they are actually carrying too often is an excessively heavy backpack. The article "Effectiveness of a School-Based Backpack Health Promotion Program" (*Work* 2003: 113–123) reported the following data for a sample of 131 sixth graders: for backpack weight (lb), $\bar{x} = 13.83$, $s = 5.05$; for backpack weight as a percentage of body weight, a 95% confidence interval for the population mean was (13.62, 15.89).
    a. Calculate and interpret a 99% CI for population mean backpack weight.
    b. Obtain a 99% CI for population mean weight as a percentage of body weight.
    c. The American Academy of Orthopedic Surgeons recommends that backpack weight be at most 10% of body weight. What does your calculation of (b) suggest, and why?

41. Refer to the discussion below Expression (8.12) concerning one-sided large-sample confidence bounds. Determine the confidence level for each of the following large-sample one-sided confidence bounds:
    a. Upper bound: $\bar{x} + .84s/\sqrt{n}$
    b. Lower bound: $\bar{x} - 2.05s/\sqrt{n}$
    c. Upper bound: $\bar{x} + .67s/\sqrt{n}$

42. Use the results of Sections 6.3–6.4 to show that the variable T on which the PI is based does in fact have a *t* distribution with $n - 1$ df.

8.3 Intervals for a Population Proportion

The previous section focused primarily on interval estimates for a population mean, μ. In this section, we consider some methods for constructing a CI for a proportion. Let p denote the proportion of "successes" in a population: the proportion of all students at your university that graduate, the proportion of all production items that meet manufacturer specs, the proportion of all laptops that do not need warranty service, etc.

A random sample of n individuals or objects will be selected, and X will denote the number of successes in the sample. Provided that n is small relative to the population size, X can be regarded as a Bin(n, p) random variable. Moreover, as discussed in Chapter 6 in connection with the Central Limit Theorem, if both $np \geq 10$ and $n(1 - p) \geq 10$, X has approximately a normal distribution.

A natural estimator of the parameter p is the statistic $\hat{P} = X/n$, the sample proportion of successes. As seen in Example 7.4, properties of X imply that

$$E(\hat{P}) = p \quad \text{and} \quad \sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$$

Also, since $\hat{P}$ is just X multiplied by the constant $1/n$, $\hat{P}$ has an approximately normal distribution. Standardizing $\hat{P}$ by subtracting its mean and dividing by its standard deviation then implies that

$$P\left(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

(This is only an approximate equality, because $\hat{P}$ is only approximately normal.) The standardized version of $\hat{P}$ is also an approximate pivotal quantity. Proceeding as suggested in the subsection "Deriving a General Confidence Interval" (Section 8.1), the confidence limits result from replacing each $<$ by $=$ and solving for p. These equations are quadratic; sparing the reader the details, the two roots are

$$p = \frac{\hat{P} + z_{\alpha/2}^2/2n}{1 + z_{\alpha/2}^2/n} \pm z_{\alpha/2}\frac{\sqrt{\hat{P}(1-\hat{P})/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n}$$

PROPOSITION Let $\hat{p}$ denote the observed fraction of successes in a random sample of size n from a population with true success proportion p. Then an **approximate 100(1 − α)% confidence interval for p** has endpoints

$$\tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n} \tag{8.15}$$

where $\tilde{p} = [\hat{p} + z_{\alpha/2}^2/2n]/[1 + z_{\alpha/2}^2/n]$. This is often referred to as the *score CI* for p.

Example 8.13 Anyone using e-mail or surfing the web (these days, virtually everyone!) has encountered *phishing*, fraudulent e-mails or websites designed to look legitimate and thus trick people into revealing credit card numbers, passwords, etc. The article "Is Domain Highlighting Actually Helpful in Identifying Phishing Web Pages?" (*Human Factors* 2016: 640–660) describes a study in which 320 participants were shown webpages and asked to identify which were legitimate and which were fraudulent. In one phase of the study, 157 participants misidentified a phishing website as safe. Let p denote the proportion of *all* web users that would misidentify this fraudulent website under the study's settings. A point estimate for p is the sample proportion $\hat{p} = 157/320 = .491$. Using the score interval (8.15), an approximate 95% confidence interval for p is

$$\frac{(.491) + (1.96)^2/2(320)}{1 + (1.96)^2/320} \pm 1.96\frac{\sqrt{(.491)(.509)/320 + (1.96)^2/4(320)^2}}{1 + (1.96)^2/320}$$
$$= .491 \pm .054 = (.437, .545)$$

With 95% confidence, we conclude that between 43.7% and 54.5% of all web users would fall for this fraudulent website under the study's settings.

   The point of the article was to determine whether looking at a site's URL in the browser's address bar can help people detecting phishing sites. In the part of the study just described, participants could not see the URL; in a second phase, participants were shown different websites along with their addresses, and only 31.6% of them made the same mistake. Using (8.15) again, with 95% confidence we infer that between 26.7 and 36.8% of *all* web users would mistakenly think a particular fraudulent website was safe even if they could see its web address.   ∎

   One-sided confidence bounds are available for p, and they are constructed in a similar fashion to those discussed in Section 8.1. To obtain an approximate $100(1 - \alpha)\%$ upper confidence bound for p, simply replace $\pm$ with $+$ and $z_{\alpha/2}$ with z_α in (8.15). For a lower confidence bound, $\pm$ becomes a $-$ sign.

The "Traditional" Interval for *p*

If the sample size n is very large, then the terms $z^2/2n$, and z^2/n, and $z^2/4n^2$ in (8.15) are generally quite negligible (small) compared to the other terms in the expression. Removing those "lesser" terms simplifies the score interval to

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \tag{8.16}$$

Expression (8.16), known as the *Wald CI* for p, is the one that for decades has appeared in introductory statistics textbooks. It clearly has a much simpler and more appealing form than the score CI. So, why bother with (8.15)?

   Suppose we use $z_{.025} = 1.96$ in the Wald interval (8.16). Then our *nominal* confidence level (the one we *think* we're buying by using 1.96) is approximately 95%. So, before a sample is collected, the chance that the random interval includes the actual value of p—i.e., the *coverage probability*—should be about .95. But, as Figure 8.8 shows for the case $n = 100$, the actual coverage probability for this interval can differ considerably from the nominal probability .95, particularly when p is not close to .5. (The graph of coverage probability versus p is very jagged because the underlying binomial distribution is discrete rather than continuous.) This is a serious deficiency of the Wald interval: the actual confidence level can be considerably less than the nominal level, even for fairly large-sample sizes.
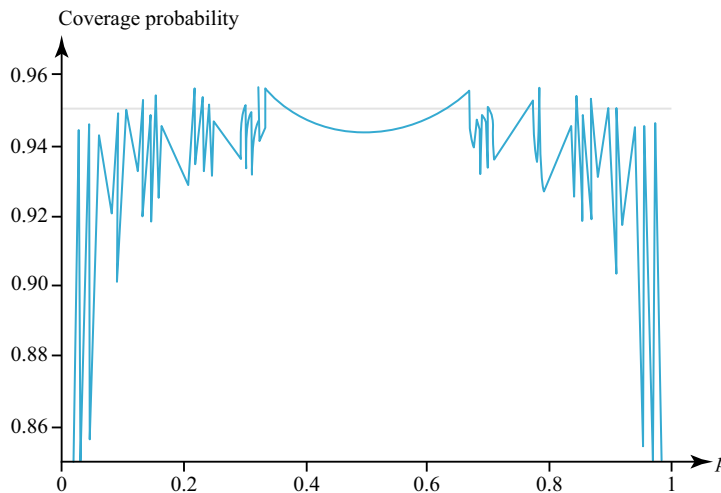
Figure 8.8   Actual coverage probability for the Wald interval (8.16) for varying values of p when $n = 100$

Research has shown that the score interval (8.15) rectifies this behavior: for virtually all sample sizes and values of p, the actual confidence level will be quite close to the nominal level specified by the choice of $z_{\alpha/2}$. This is due largely to the fact that the score interval is shifted a bit toward .5 compared to the Wald interval. In particular, the midpoint $\tilde{p}$ of the score interval is always a bit closer to .5 than is the midpoint $\hat{p}$ of the Wald interval; this is especially important when p is close to 0 or 1.

Sample Size Determination

Equating the width of the CI for p to a prespecified width w gives a quadratic equation for the sample size n necessary to give an interval with a desired degree of precision. The solution is

$$n = \frac{z_{\alpha/2}^2 \left[2\hat{p}\hat{q} - w^2 + \sqrt{(2\hat{p}\hat{q})^2 + (1 - 4\hat{p}\hat{q})w^2} \right]}{w^2} \tag{8.17}$$

where $\hat{q} = 1 - \hat{p}$. Neglecting the terms in the numerator involving w^2, which will be quite small because w is a (typically small) decimal value, gives

$$n = \frac{4z_{\alpha/2}^2 \hat{p}\hat{q}}{w^2}$$

This latter expression is what results from equating the width of the Wald interval to w.

These formulas unfortunately include $\hat{p}$, which a researcher does not know in advance of the study when trying to determine what sample size is required. The most conservative approach is to use the fact that both expressions are maximized when $\hat{p} = \hat{q} = .5$. Thus if $\hat{p}\hat{q} = (.5)(.5) = .25$ is used in (8.17), the width of the CI will be at most w regardless of what value of $\hat{p}$ results from the sample. Alternatively, if the investigator has a rough estimate p_0 from some prior study, p_0 can be used in place of $\hat{p}$.

Example 8.14 Using the conservative method $\hat{p} = \hat{q} = .5$ proposed above, the sample size formula (8.17) simplifies to

$$n = \frac{z^2 \left[2(.25) - w^2 + \sqrt{(2(.25))^2 + (1 - 4(.25))w^2} \right]}{w^2} = \frac{z^2 \left[.5 - w^2 + \sqrt{(.5)^2} \right]}{w^2} = \frac{z^2(1 - w^2)}{w^2}$$

The width of the 95% CI in Example 8.13 is .108. The value of n necessary to ensure a width of no more than .08, irrespective of the value of $\hat{p}$, is

$$n = \frac{1.96^2(1 - .08^2)}{.08^2} = 596.4$$

Thus, a sample size of 597 should be used. The expression for n based on the Wald CI gives a slightly larger value of 601. ∎

Alternative Intervals for *p*

Even the score interval (8.15) is not perfect: because it relies on a normal approximation to the binomial, it isn't necessarily suited to situations where that approximation is poor (although it fares much better than the Wald interval!). Interval estimation methods exist that do not rely on this normal approximation and, hence, are reliable for all values of n and p. The so-called *exact method*, based on

the binomial distribution, is guaranteed to produce an interval having coverage probability at least as great as the nominal confidence level. But, the exact method tends to produce CIs that are very wide—an undesirable property for a confidence interval.

These issues have been largely resolved by research reported in the 2014 article "A Coverage Probability Approach to Finding an Optimal Binomial Confidence Procedure" (*The Amer. Stat.*, Schilling and Doi). The article describes a computer-intensive method for determining what the authors call the *length/coverage optimal* interval for p. That is, for any given sample data, their method produces an interval that (1) has coverage probability at least as great as the specified confidence level, regardless of the value of p, and (2) is the *shortest* among all such intervals. There is no explicit "formula" for the length/coverage optimal interval, but the article's authors have created online software to compute the interval automatically. This can be accessed at http://shiny.stat.calpoly.edu/LCO-CI/.

Example 8.15 The Super Bowl is one of the most watched events in the country, but not *everyone* watches it. Imagine that in a sample of 25 students at a university, all 25 said they watched the most recent Super Bowl. What can be said about the parameter p = the proportion of all students at this university who watched the game? Though $\hat{p} = 25/25 = 1$, it seems unrealistic to infer that 100% of *all* students saw the most recent Super Bowl.

The Wald interval clearly should not be applied here; doing so results in a CI of $1 \pm 0 = 1$, suggesting p is known to be 100% exactly (which, again, is just silly). Using the Schilling-Doi website, with $n = 25$ and $x = 25$ the length/coverage optimal 95% confidence interval for p is (.866, 1). That is, we are 95% confident that between 86.6% and 100% of all students at this university watched the most recent Super Bowl. The score interval based on this data is nearly identical, (.867, 1), suggesting that even for small-sample sizes the score interval is a wise choice. ∎

Exercises: Section 8.3 (43–56)

43. According to Oklahoma State University's 2015 Food Demand Survey, 859 of 1044 randomly selected adults support mandatory labels on foods produced with genetic engineering (popularly called GMO products). Construct an approximate 95% confidence interval for the proportion of all adults who support mandatory labeling of GMO products.

44. In a survey of 1100 drivers, 90% admitted to careless or aggressive driving during the previous six months ("Nine out of Ten Drivers Admit in Survey to Having Done Something Dangerous," *Knight Ridder Newspapers*, Jul. 8, 2005). Assuming these 1100 drivers may be treated as a random sample of all drivers in the USA, construct and interpret a 95% CI for the true proportion of drivers who have engaged in "dangerous" driving in the past six months.

45. A June, 2019 Gallup survey of 1018 randomly selected US adults found that 53% supported the government sponsoring a manned mission to Mars.

   a. Construct and interpret a 95% lower confidence bound for the proportion of all US adults who support such a mission.

   b. Does your answer to part (a) clearly indicate that a majority of all US adults feel this way (at least as of June, 2019)? Explain.

46. The article "Teens and Young Adults Embrace Online Multiplayer and Competitive Video Games" (*Washington Post*, Apr. 3, 2018) reported that, in a survey of 522 Americans age 14–21, 38% said they consider themselves a fan of competitive gaming. Construct a 99% confidence interval for the proportion of all Americans in this age group that are fans of esports or competitive gaming.

47. As reported by CNBC (Dec. 11, 2018), 57% of people surveyed admitted to shopping online while at work. The survey was based on a random sample of $n = 2020$ US adults. Construct and interpret a 90% upper confidence bound for the proportion of all US adults who shop online while at work.

48. The article "Repeatability and Reproducibility for Pass/Fail Data" (*J. Testing Eval.* 1997: 151–153) reported that in $n = 48$ trials in a particular laboratory, 16 resulted in ignition of a particular type of substrate by a lighted cigarette. Let p denote the long-run proportion of all such trials that would result in ignition.

    a. Use the score interval (8.15) to construct a 95% CI for p.
    b. Use the Wald interval (8.16) to construct a 95% CI for p.
    c. How do the intervals in parts (a) and (b) compare? Is the narrower interval preferable here? Why or why not?

49. The article "Limited Yield Estimation for Visual Defect Sources" (*IEEE Trans. Semicon. Manuf.* 1997: 17–23) reported that, in a study of a particular wafer inspection process, 356 dies were examined by an inspection probe and 201 of these passed the probe. Assuming a stable process, calculate a 95% (two-sided) confidence interval for the proportion of all dies that pass the probe.

50. There is increasing concern within the health sciences community over the use of electronic tobacco products. The article "Exposure to Tobacco and Nicotine Product Advertising: Associations with Perceived Prevalence of Use Among College Students" (*Amer. J. College Health*, Volume 66, 2018, Issue 8) reported on a study based on the Texas College Tobacco Project survey administered in 2016. In a sample of 5767 undergraduates ages 18–25, 9.1% said they had used electronic cigarettes at least once during the previous 30 days. Calculate and interpret a confidence interval using a 99% confidence level for the proportion of all students in the population sampled who used e-cigarettes during the previous 30 days.

51. The article "Broad Agreement on Most Ideas to Curb School Shootings" from the Gallup.com website reported on a survey carried out from March 5–11, 2018. There was overwhelming support for more training of school and security personnel, and for background checks for all gun sales. However, opinion was more evenly split on the issue of arming teachers; 42% of the 1515 adults in the sample favored providing teachers with weapons. Using a 95% confidence level, calculate an upper confidence bound for the percentage of all adults in the USA who favor arming teachers. Based on your interval, can you be confident that a majority of the population does not favor such a policy?

52. In a sample of 1000 randomly selected consumers who had opportunities to send in a rebate claim form after purchasing a product, 250 of these people said they never did so ("Rebates: Get What You Deserve," *Consumer Reports*, May 2009: 7). Reasons cited for their behavior included too many steps in the process, amount too small, missed deadline, fear of being placed on a mailing list, lost receipt, and doubts about receiving the money. Calculate an upper confidence bound at the 95% confidence level for the true proportion of such consumers who never apply for a rebate. Based on this bound, is there compelling evidence that the true proportion of such consumers is smaller than 1/3? Explain your reasoning.

53. A state legislator wishes to survey residents of her district to see what proportion of the electorate is aware of her position on using state funds to pay for abortions.

    a. What sample size is necessary if the 95% CI for p is to have width of at most .10 irrespective of p?
    b. If the legislator has strong reason to believe that at least 2/3 of the electorate know of her position, how large a sample size would you recommend?

54. A mortgage company wishes to estimate the proportion of all borrowers who default on their home loans, to within a margin of error of 2 percentage points ($\pm.02$). What sample size is required to achieve this at the 90% confidence level? How does the answer change if it is believed initially that roughly 15% of all customers default?

55. A recent student project asked students at Cal Poly (where one of the authors teaches) whether they have ever been arrested on alcohol- or drug-related charges, including drunk driving. Out of 57 students surveyed in the College of Business, only four reported they had been arrested (the surveys were made anonymous to encourage truthful responses). Assuming these 57 students comprise a random sample of all business students at the school, and assuming students answered truthfully, estimate with 95% confidence the proportion of all business students who have been arrested on such charges.

56. Reconsider the score CI (8.15) for p, and focus on a confidence level of 95%. Show that the confidence limits agree quite well with those of the Wald interval (8.16) once two successes and two failures have been appended to the sample, i.e., (8.16) based on $(x + 2)$ S's in $(n + 4)$ trials. [*Hint*: $1.96 \approx 2$.]

8.4 Confidence Intervals for the Population Variance and Standard Deviation

Although inferences concerning a population variance σ^2 or standard deviation σ are usually of less interest than those about a mean or proportion, there are occasions when such procedures are needed. In the case of a normal population distribution, inferences are based on a result from Section 6.4 concerning the sample variance S^2: if $X_1, \ldots, X_n$ is a random sample from a normal distribution with variance σ^2, then the random variable

$$\frac{(n-1)S^2}{\sigma^2} \tag{8.18}$$

has a chi-squared (χ^2) distribution with $n - 1$ df.

As discussed in Section 6.3, the chi-squared distribution is a continuous probability distribution with a single parameter v, the number of degrees of freedom. To specify inferential procedures that use the chi-squared distribution, recall the notation for critical values from Section 6.3.

NOTATION Let $\chi^2_{\alpha,v}$, called a **chi-squared critical value**, denote the number on the measurement axis such that α of the area under the chi-squared curve with v df lies to the right of $\chi^2_{\alpha,v}$. See Figure 8.9a.

It was necessary to tabulate only upper-tail critical values for the t distribution ($t_{\alpha,v}$ for small values of α), because t density curves are symmetric. But chi-squared distributions are *not* symmetric, so Appendix Table A.5 contains values of $\chi^2_{\alpha,v}$ for α both near 0 and near 1, as illustrated in Figure 8.9b. For example, $\chi^2_{.025,14} = 26.119$ and $\chi^2_{.95,20}$ (the 5th percentile) $= 10.851$.

The rv $(n - 1)S^2/\sigma^2$ in (8.18) satisfies the two properties of being a pivotal quantity: it is a function of the parameter of interest σ, yet its probability distribution, χ^2_{n-1}, does not depend on this parameter. So, the methods described in Section 8.1 can be applied to this rv in order to construct a confidence interval for σ. Analogous to Figure 8.9b, the area under a χ^2_{n-1} curve to the right of $\chi^2_{\alpha/2,n-1}$ is $\alpha/2$, as
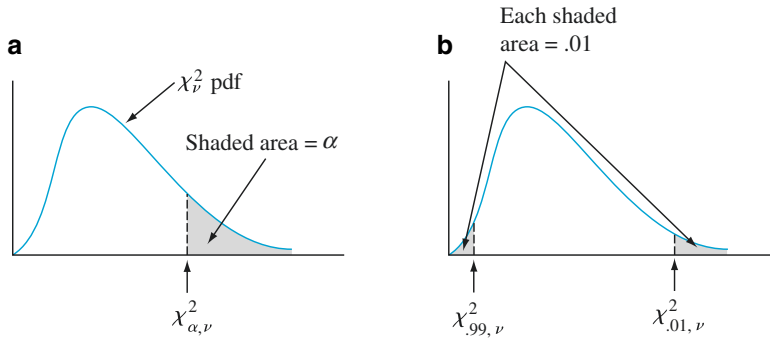
Figure 8.9 $\chi^2_{\alpha,v}$ notation illustrated

is the area to the left of $\chi^2_{1-\alpha/2,n-1}$. Thus the area captured between these two critical values is $1 - \alpha$, from which we may infer

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1 - \alpha \qquad (8.19)$$

The inequalities in (8.19) are equivalent to

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$$

Substituting the computed value s^2 into the limits gives a CI for σ^2, and taking square roots gives an interval for σ.

PROPOSITION   A **100(1 − α)% confidence interval for the variance σ^2 of a normal population** is given by the endpoints

$$\left(\frac{n-1}{\chi^2_{\alpha/2,n-1}} \cdot s^2, \; \frac{n-1}{\chi^2_{1-\alpha/2,n-1}} \cdot s^2\right) \qquad (8.20)$$

A **100(1 − α)% confidence interval for σ** has lower and upper limits that are the square roots of the corresponding limits in (8.20).

An upper confidence bound for σ^2 is obtained from the right endpoint of (8.20) by substituting α for α/2 in the χ^2 critical value; taking the square root of that quantity results in an upper confidence bound for σ. The left endpoint of (8.20) can be modified similarly to achieve lower confidence bounds.

Recall from Section 6.3 that the expected value of a chi-squared rv is its df; here, df = $n - 1$. As a result, the upper critical value $\chi^2_{\alpha/2,n-1}$ should exceed $n - 1$, and so the fraction appearing in the left endpoint of (8.20) should be less than 1. Similarly, the fraction in the right endpoint should be greater than 1, so that the interval's endpoints lie on either side of the point estimate s^2 (albeit not equidistant from s^2).

Example 8.16   The report "Strand Debonding for Pretensioned Girders" (*NCHRP Research Report 849:* 2017) includes the following yield strength measurements (ksi) for a sample of 16 reinforcing bars of the type commonly used in bridges:

67.6	63.6	69.2	82.1	69.7	79.0	67.1	65.9
75.1	65.4	70.5	75.6	72.7	63.6	70.1	75.6

Figure 8.10 shows a normal probability plot, which indicates that a normal model for the population of yield strength measurements is plausible.
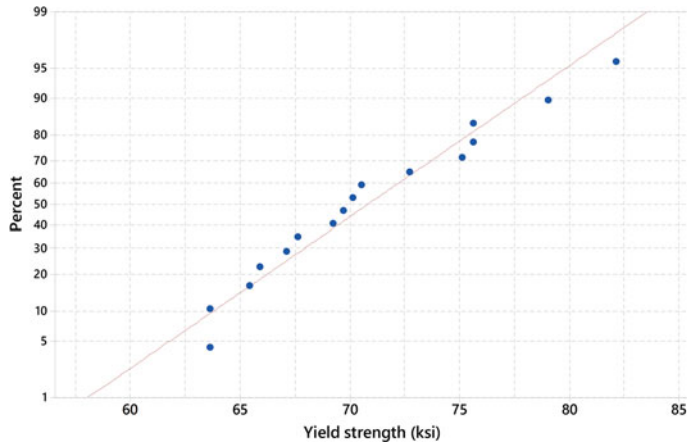
Figure 8.10   Normal probability plot of the yield strength data in Example 8.16

Let σ denote the true standard deviation of the yield strength distribution. The computed value of the sample sd is $s = 5.47$; this is a point estimate of σ. With df $= 16 - 1 = 15$, a 95% CI requires the critical values $\chi^2_{.025,15} = 27.488$ and $\chi^2_{.975,15} = 6.262$. The resulting interval for σ^2 is

$$\left(\frac{16-1}{27.488} \cdot (5.47)^2, \frac{16-1}{6.262} \cdot (5.47)^2 \right) = (16.32,\ 71.67)$$

Taking the square root of the endpoints yields (4.04, 8.47) as the 95% CI for σ. At the 95% confidence level, the true standard deviation of yield strength for this type of bridge reinforcing bar is between 4.04 and 8.47 ksi.   ∎

The confidence interval illustrated in the preceding example relies heavily on the normality assumption. Research has shown that using (8.20) with data from nonnormal populations can result in highly unreliable intervals, even when the sample size n is large. (For example, the coverage probability of an ostensible 95% CI can be far less than .95.) A more robust method is presented in the article "Approximate Confidence Interval for Standard Deviation of Nonnormal Distributions" (*Comp. Stat. & Data Analysis* 2006: 775–782) by D. Bonett. This interval, though typically wider than (8.20), has been shown to achieve much better coverage probability for a wide variety of nonnormal population distributions. It is now incorporated into the Minitab software package, which

produces (8.20) and Bonett's interval when users request a CI for population variance. See Exercise 92 for more information.

Alternatively, the *bootstrap* method presented in the next section can produce an interval estimate for population standard deviation (or variance) without requiring population normality.

Exercises: Section 8.4 (57–62)

57. Determine the values of the following quantities:

    a. $\chi^2_{.1,15}$ b. $\chi^2_{.1,25}$ c. $\chi^2_{.01,25}$ d. $\chi^2_{.99,25}$

    e. $\chi^2_{.995,25}$

58. Determine the following:

    a. The 95th percentile of the χ^2_{10} distribution
    b. The 5th percentile of the χ^2_{10} distribution
    c. $P(10.98 \leq Y \leq 36.78)$, where Y is a χ^2_{22} rv
    d. $P(Y < 14.611 \text{ or } Y > 37.652)$, where Y is a χ^2_{25} rv

59. Exercise 17 provided alcohol percentage data for a sample of 16 beers. The sample standard deviation of those measurements was $s = .8483$. Construct a 90% CI for the population variance σ^2 of alcohol percentage in beers, and then a 90% CI for σ.

60. Exercise 24 gave a random sample of 20 ACT scores from students taking college freshman calculus. Calculate a 99% CI for the standard deviation of the population distribution. Is this interval valid whatever the nature of the distribution? Explain.

61. Here are the names of 12 orchestra conductors and their performance times in minutes for Beethoven's Ninth Symphony:

Bernstein	71.03	Furtwängler	74.38
Leinsdorf	65.78	Ormandy	64.72
Solti	74.70	Szell	66.22
Bohm	72.68	Karajan	66.90
Masur	69.45	Rattle	69.93
Steinberg	68.62	Tennstedt	68.40

    a. Check to see that normality is a reasonable assumption for the performance time distribution.
    b. Compute a 95% CI for the population standard deviation, and interpret the interval.
    c. Supposedly, classical music is 100% determined by the composer's notation, including all timings. Based on your results, is this true or false?

62. Refer to the baseball game times in Exercise 31. Calculate an upper confidence bound with confidence level 95% for the population standard deviation of game time. Interpret your interval. Explore the issue of normality for the data and explain how this is relevant to your interval.

8.5 Bootstrap Confidence Intervals

How can a confidence interval for the mean be constructed if the population distribution is not normal and the sample size n is small? Can we find confidence intervals for other parameters, such as the population median or the 90th percentile of the population distribution? The **bootstrap**, developed by Bradley Efron in the late 1970s, facilitates calculating estimates in situations where statistical theory does not produce a formula for a confidence interval. The method substitutes heavy computation for theory, and many statistical software packages now implement various bootstrap methods (this includes SAS, R, JMP Pro, and Minitab). The *parametric bootstrap*, for applications with a known (or assumed) population distribution, was briefly mentioned in Section 7.1. In this section we are concerned with the case of an unknown distribution, for which the *nonparametric bootstrap* is appropriate.

The Bootstrap Method

Traditional inference (e.g., the presentation in Sections 8.1–8.4) relies on the sampling distribution of a statistic—the distribution of the values of that statistic if we were to hypothetically take all possible random samples of size n from the parent population. This is illustrated for the sample mean in Figure 8.11a. In contrast, the bootstrap method considers what would happen if we were to draw repeatedly *from the sample at hand*. For example, if we had $n = 15$ observations from some population and we were interested in drawing inferences about the mean, the **bootstrap distribution** of $\overline{X}$ would consist of all $\bar{x}$ values that could be obtained by taking a random sample of size 15 (called a **bootstrap sample** or **resample**) from the original 15 observations. Obviously, for that to make sense, bootstrap sampling must occur *with replacement*; otherwise, we would get the same sample over and over again. Figure 8.11b diagrams the basic bootstrap method.
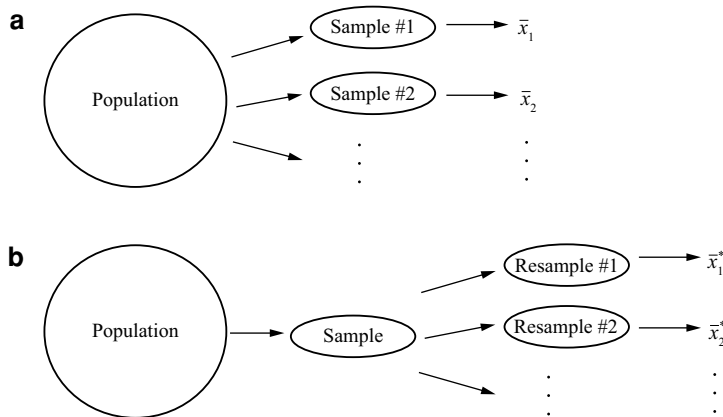
Figure 8.11   Two versions of sampling variability: (**a**) creating the sampling distribution of $\bar{X}$; (**b**) creating the bootstrap distribution of $\overline{X}$

Philosophically, the bootstrap method treats the sample at hand as if it were the population, since the sample represents in a sense everything the user knows about the underlying population. Again, the advantage of bootstrapping is that the method applies in the absence of theory (e.g., the CLT) or distributional requirements (e.g., normality). The steps in the basic bootstrap method are as follows.

BASIC BOOTSTRAP METHOD

Suppose we wish to generate the **bootstrap distribution** of a statistic $\hat{\theta}$ based upon an observed sample $x_1, x_2, \ldots, x_n$ from some population.

1. Take a random sample of size n *with replacement* from $x_1, x_2, \ldots, x_n$, resulting in $x_1^*, x_2^*, \ldots, x_n^*$.
2. Compute the value of the statistic $\hat{\theta}$ from this bootstrap sample; label the resulting value $\hat{\theta}^*$.
3. Repeat steps **1** and **2** a large number of times (say, B times), giving values $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$ for the statistic of interest.

These values $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$ approximate the bootstrap distribution of $\hat{\theta}$.

We use the word "approximate" above only because the process terminates after obtaining B re-samples and B resulting values $\hat{\theta}^*$. The complete bootstrap distribution of $\hat{\theta}$ consists of *all* $\hat{\theta}^*$ values from *all* possible bootstrap samples, but the number of such samples can be unwieldy for even moderate sample sizes. It can be shown that the number of bootstrap samples from an original sample of size n is $\binom{2n-1}{n-1}$; even for $n = 15$, this is more than 77 million, and the number of possible bootstrap samples increases rapidly with n. In practice, $B = 1000$ is often used.

It has been shown experimentally that the bootstrap distribution of a statistic quite often resembles the actual sampling distribution of that statistic. In particular, the standard error of a statistic $\hat{\theta}$, $\sigma_{\hat{\theta}}$, can often be well approximated by its **bootstrap standard error**, defined to be the sample standard deviation of the $\hat{\theta}_i^*$'s:

$$s_{\text{boot}} = \sqrt{\frac{1}{B-1} \sum \left(\hat{\theta}_i^* - \bar{\theta}^* \right)^2} \tag{8.21}$$

The symbol $\bar{\theta}^*$ in (8.21) denotes the mean of the bootstrap values of $\hat{\theta}$, i.e., $\bar{\theta}^* = \sum \hat{\theta}_i^*/B$. In the bootstrap literature, B is sometimes used in place of $B - 1$ in (8.21); for typical values of B, there is usually little difference between the resulting estimates.

Example 8.17 In a student project, Erich Brandt studied tips at a restaurant. Here is a random sample of 30 observed tip percentages:

| 22.7 | 16.3 | 13.6 | 16.8 | 29.9 | 15.9 | 14.0 | 15.0 | 14.1 | 18.1 | 22.8 | 27.6 | 16.4 | 16.1 | 19.0 |
| 13.5 | 18.9 | 20.2 | 19.7 | 18.2 | 15.4 | 15.7 | 19.0 | 11.5 | 18.4 | 16.0 | 16.9 | 12.0 | 40.1 | 19.2 |

We would like to get a confidence interval for μ, the population mean tip percentage at this restaurant. However, this is not a very large sample and there is a problem with positive skewness, as shown in the normal probability plot of Figure 8.12. Most of the tips are between 10 and 20%, but a few big tips cause enough skewness to invalidate the normality assumption. The one-sample t interval applied to this data would not be trustworthy.

Mean	18.43
StDev	5.761
N	30
AD	1.828
P-Value	<0.005
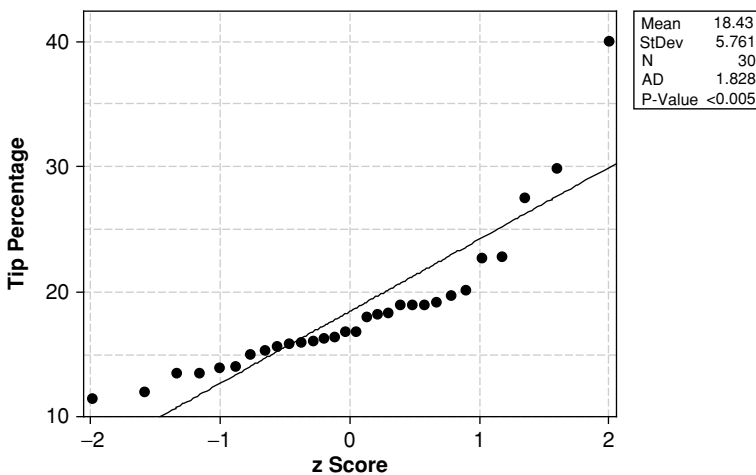
Figure 8.12 Normal probability plot (from Minitab) of the tip percentages

To implement the bootstrap method here, regard the 30 observations as constituting a population. Then take a large number of random resamples with replacement, each of size 30, from this "population." For each of these resamples compute the sample mean (because the population mean is the parameter of interest). Then use the distribution of these resample means to get a confidence interval for the population mean (we'll explain how shortly). To help get a feeling for how this works, here is the first of $B = 1000$ resamples generated using software

| 22.8 | 16.8 | 16.0 | 19.0 | 19.2 | 20.2 | 13.6 | 15.9 | 22.8 | 11.5 | 15.9 | 14.0 | 29.9 | 19.2 | 16.0 |
| 27.6 | 14.1 | 13.5 | 16.8 | 15.4 | 20.2 | 16.4 | 20.2 | 16.9 | 16.8 | 22.8 | 19.7 | 18.2 | 22.7 | 18.2 |

That is, $x_1^* = 22.8$, $x_2^* = 16.8$, …, $x_{30}^* = 18.2$ for this bootstrap sample. Notice that some values from the original sample are repeated (due to sampling with replacement), while some values don't appear at all. This first bootstrap sample has mean $\bar{x}_1^* = 18.41$; the asterisk emphasizes that this is the mean of a bootstrap resample and not of the original sample of 30 tip percentages. This process was repeated 1000 times, resulting in resample means $\bar{x}_1^*, \ldots, \bar{x}_{1000}^*$. Figure 8.13 displays a histogram of these 1000 $\bar{x}^*$ values, the approximate bootstrap distribution of the statistic $\overline{X}$. Notice that the bootstrap distribution of $\overline{X}$ is somewhat right-skewed, inconsistent with a normal distribution.
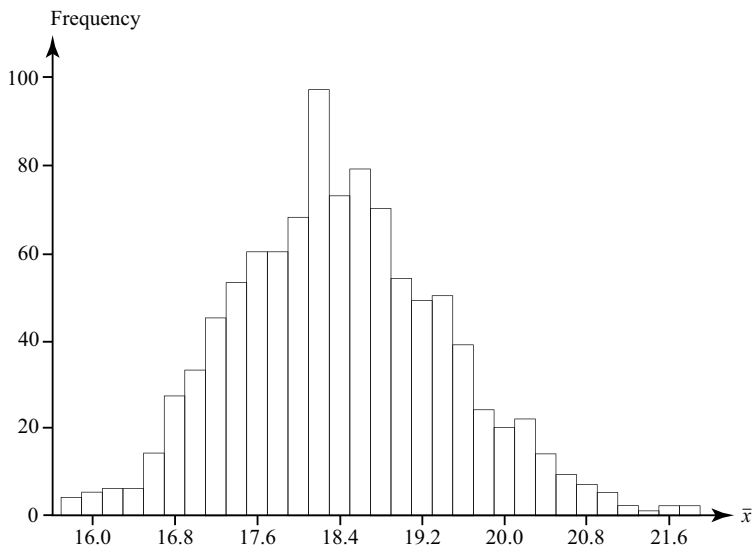
Figure 8.13   Histogram of the bootstrap distribution of $\bar{X}$ for Example 8.17   ∎

In Example 8.17, the mean for the original 30 observations is $\bar{x} = 18.43$. On the other hand, the mean of the bootstrap distribution displayed in Figure 8.13 is 18.416. This is due to only taking 1000 bootstrap samples; it can be shown that the complete bootstrap distribution of $\overline{X}$ will always be centered at the mean of the original sample. However, this is not the case for other statistics. For example, the mean value of the bootstrap distribution of a trimmed mean is not necessarily the "true" value of $\bar{x}_{tr}$ (i.e., the trimmed mean of the original sample). The **bias** of a bootstrap distribution is defined to be the difference between these two values. In practice, if this bias is small relative to the magnitude of the data itself, there is little cause for concern.

Example 8.18 As data proliferates across every business sector, data base administrator (DBA) has become an increasingly lucrative career choice. Figure 8.14a shows a histogram of the salaries for 115 DBAs with 0–2 years of experience ("The 2019 Data Professional Salary Survey Results," www. brentozar.com). Because some salaries are unusually high (both in the sample and the population), a 10% trimmed mean might be considered an appropriate measure of center. To estimate the population trimmed mean, we must first understand the variability of the statistic $\overline{X}_{\mathrm{tr}}$; the bootstrap method is appropriate because a theoretical description of the sampling distribution of trimmed means is not available. Figure 8.14b shows the bootstrap distribution of $\overline{X}_{\mathrm{tr}}$ based on $B = 1000$ resamples. Interestingly, this bootstrap distribution appears to be approximately normal.
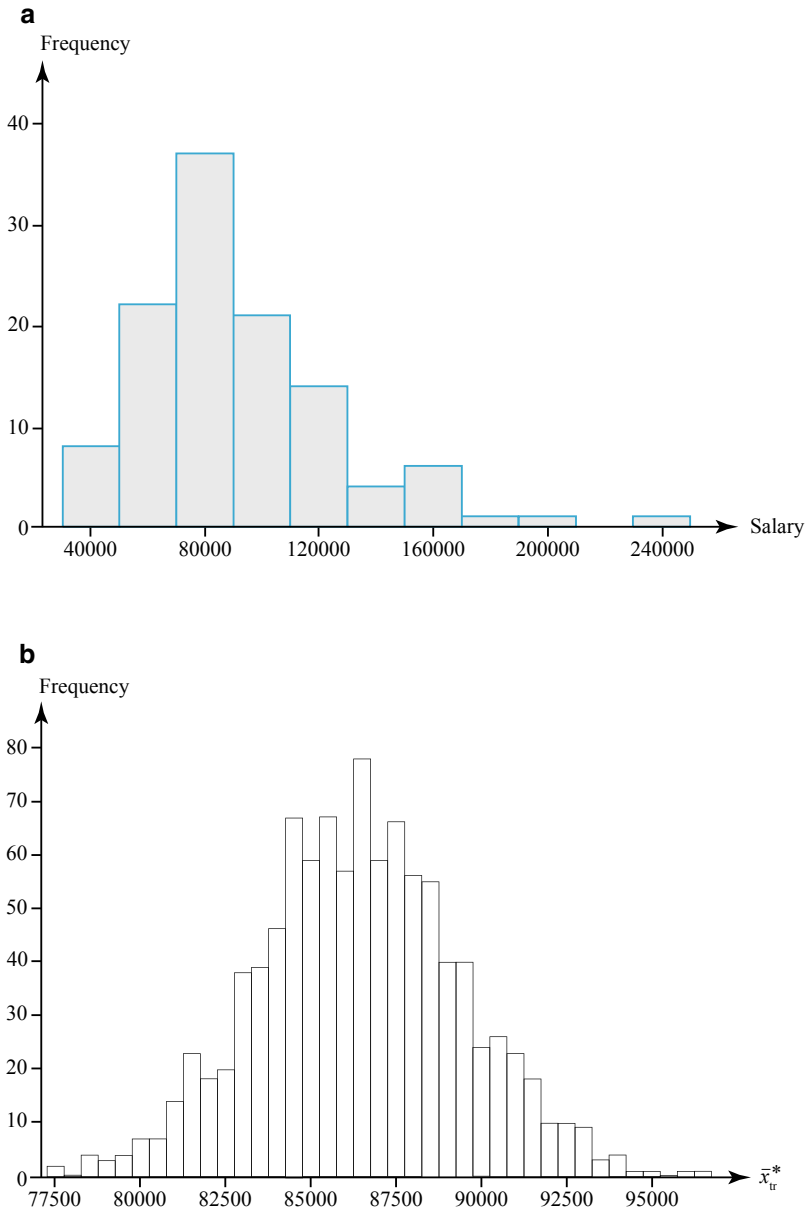
Figure 8.14 Graphs for Example 8.18: (**a**) histogram of 115 salaries; (**b**) bootstrap distribution of $\overline{X}_{\mathrm{tr}}$ from this sample

The 10% trimmed mean for the original 115 salaries is $\bar{x}_{tr}$ = \$86,320, while the center (i.e., mean) value of the associated bootstrap distribution is \$86,422. The bias of –\$102 in this simulated bootstrap distribution is small relative to the salaries themselves, suggesting the bootstrap distribution will be a reasonable tool for inference on the *population* trimmed mean.   ■

Bootstrap Interval Estimation

Once we have the bootstrap distribution of a statistic, several different methods can be used to obtain a confidence interval for the corresponding parameter. If, as in Example 8.18, the bootstrap distribution of the statistic appears reasonably bell-shaped, then a variation on the t interval from Section 8.2 may be employed. Recall that the t interval for a mean μ, assuming a normal sampling distribution for $\overline{X}$, is $\bar{x} \pm t_{n-1,\alpha/2} \cdot s/\sqrt{n}$; the $s/\sqrt{n}$ term is the estimated standard error of $\overline{X}$. By analogy, a confidence interval for a parameter θ based on a bootstrapped statistic $\hat{\theta}$ could be obtained by replacing $\bar{x}$ in the one-sample t interval with the calculated value of $\hat{\theta}$ from the sample and replacing $s/\sqrt{n}$ with the bootstrap standard error of $\hat{\theta}$.

DEFINITION   Suppose we wish to estimate a parameter θ by the corresponding sample statistic $\hat{\theta}$. A **bootstrap t confidence interval** for θ with confidence level $100(1 - \alpha)\%$ is

$$\hat{\theta} \pm t_{\alpha/2,n-1} \cdot s_{\text{boot}} \tag{8.22}$$

where the value of the statistic $\hat{\theta}$ in (8.22) is obtained from the original sample.

The bootstrap t confidence interval is appropriate when the bootstrap distribution of the statistic is approximately normal and the bias of the bootstrap distribution is small.

Example 8.19   (Example 8.18 continued) Let's construct an interval estimate for the parameter μ_{tr}, the population 10% trimmed mean salary for all database administrators with 0–2 years of experience. Since the bootstrap distribution of $\overline{X}_{tr}$ in Figure 8.14b appears approximately normal, we may reasonably apply the bootstrap t interval.

The 10% trimmed mean of the 115 salaries in the sample is $\bar{x}_{tr}$ = \$86,320. The bootstrap standard error of $\overline{X}_{tr}$— that is, the sample standard deviation of the bootstrap values displayed in Figure 8.14b— is s_{boot} = \$2994 (the software that performed the bootstrapping provided this value). A 95% confidence level requires $t_{.025,114} = 1.981$, giving a CI of

$$86{,}320 \pm 1.981(2994) = 86{,}320 \pm 5931 = (80{,}389,\ 92{,}251)$$

We are 95% confident that the 10% trimmed mean for the salary distribution of this population of DBAs is between \$80,389 and \$92,251.   ■

If the bootstrap distribution of a statistic is *not* normal, this casts doubt on the normality of its sampling distribution and suggests that a z- or t-based interval is not appropriate. Instead, we can use percentiles of the bootstrap distribution itself to form an interval. After all, critical values such as $z = \pm 1.96$ are used because they bound the "middle 95%" of a certain standardized distribution. Even if a distribution is not symmetric, we can still identify the endpoints of the "middle 95%" of a distribution.

DEFINITION   Suppose we have the bootstrap distribution of a statistic that estimates a certain
parameter. A **95% confidence bootstrap percentile interval** for that parameter
has endpoints equal to the 2.5th percentile and the 97.5th percentile of this
bootstrap distribution.

Similarly, a bootstrap percentile interval with confidence level $100(1 - \alpha)\%$ for
a parameter has endpoints equal the $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap
distribution of the corresponding statistic.

Bootstrap percentile intervals are appropriate when the bias of the bootstrap
distribution is low.

The .025 and .975 quantiles of a bootstrap distribution must be estimated from the B bootstrap
resamples actually obtained. For $B = 1000$ resamples, one typically uses the 25th-smallest and 25th-
largest values among $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_{1000}^*$; that is, the endpoints of the 95% confidence bootstrap per-
centile interval are the 25th and 976th ordered $\hat{\theta}_i^*$ values. A similar approach may be applied to other
values of B and other confidence levels.

Example 8.20 (Example 8.17 continued) Figure 8.13 shows the approximate bootstrap distribution
of $\bar{x}$ based on $B = 1000$ bootstrap resamples. The distribution does not appear normal. The 25th-
smallest and 25th-largest of the 1000 $\bar{x}^*$ values are 16.56 and 20.58, respectively. Thus, with 95%
confidence, we estimate that the true mean tip at the restaurant where Erich worked is between
16.56% and 20.58%.   ∎

A Refined Interval

Some caution must be taken when using a percentile interval. It is known that percentile intervals
sometimes have lower confidence levels than advertised. When the bootstrap distribution is skewed,
bias tends to be greater, and the percentile intervals are *not* equally likely to "miss" the value of a
parameter on the high and low sides. A somewhat sophisticated adjustment to the traditional per-
centile interval corrects for these problems: the **bias-corrected and accelerated (BCa) interval** is
almost always superior to the basic percentile interval and should be used whenever software is
available. BCa intervals are generally accurate unless the sample size is extremely small.

The acceleration aspect of the BCa interval is an adjustment for dependence of the standard error
of the estimator on the parameter that is being estimated. For example, suppose we are trying to
estimate the mean in the case of exponential data. In this case the standard deviation is equal to the
mean, and the standard error of $\bar{X}$ is $\sigma/\sqrt{n} = \mu/\sqrt{n}$, so the standard error of the estimator $\bar{X}$ depends
strongly on the parameter μ that is being estimated. If the histogram in Figure 8.13 resembled the
exponential pdf, we would expect the BCa method to make a substantial correction to the percentile
interval.

Bootstrapping the Median

The sample median $\tilde{X}$ is less sensitive than $\bar{X}$ to the influence of individual observations. For the 30
tip percentages in Example 8.17, the median is 16.85, substantially less than the mean of 18.43. The
mean is pulled upward by the few large values, but these extremes have little effect on the median.
Unfortunately, it is more difficult to get confidence intervals for the population median than for the
mean, in part because we can easily estimate the standard error of a sample mean $(s/\sqrt{n})$ but no
analogous formula exists for the sample median.

Example 8.21 (Example 8.17 continued) Let's use the bootstrap method to get a confidence interval for the true *median* tip percentage, $\tilde{\mu}$. As before, 1000 resamples of the original 30 observations are taken with replacement, but now for each resample the sample median $\tilde{x}^*$ is calculated. A histogram of the bootstrap medians $\tilde{x}_2^*, \ldots, \tilde{x}_{1000}^*$ is shown in Figure 8.15.
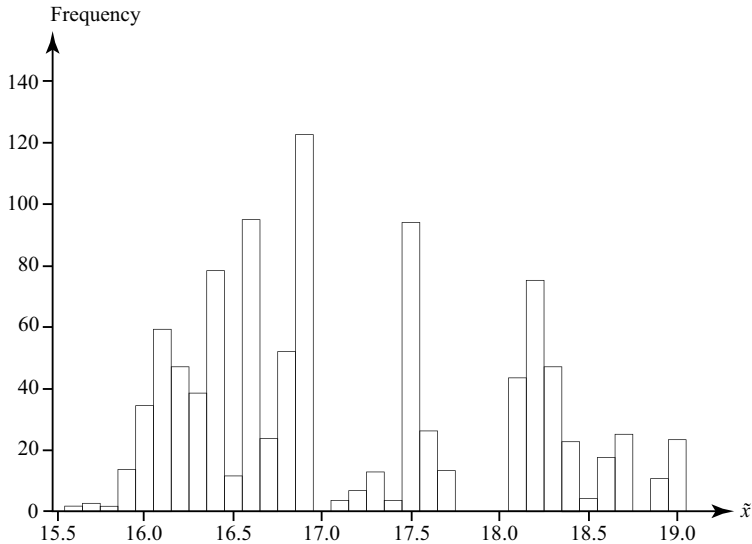
Figure 8.15 Histogram of the bootstrap medians for Example 8.21

It should be apparent that the distribution of the 1000 bootstrap medians is far from normal. As is often the case with the median, the bootstrap distribution takes on just a few values and there are many repeats. Instead of 1000 different values, as would be expected if we took 1000 samples from a true continuous distribution, here there are only a handful of distinct values.

Because the bootstrap distribution is so nonnormal, we should use the percentile interval in which the confidence limits for a 95% CI are taken from the 2.5 and 97.5 percentiles of the bootstrap distribution. When the 1000 bootstrap medians displayed in Figure 8.15 are sorted, the 25th value is 15.95 and the 976th value is 18.90, so the 95% confidence interval for the population median is (15.95, 18.90). ∎

We should be a bit uncomfortable with the results of bootstrapping the median. Given that the bootstrap distribution takes on just a few values but the true sampling distribution is continuous, we should worry a little about how well the bootstrap distribution approximates the true sampling distribution. On the other hand, the situation here is nowhere near as bad as it could be. Sometimes, especially when the sample size is smaller, the bootstrap distribution has far fewer values.

Exercise 88 presents an alternative method for constructing a confidence interval for a population median that can be applied to data from any continuous distribution, irrespective of the sample size.

Further Comments on Bootstrapping

Is the bootstrap guaranteed to work, or is it possible that the method can give grossly incorrect estimates? The key here is how closely the original sample represents the whole distribution of the random variable X. When the sample is small, then there is a possibility that important features of the distribution are not included in the data set. In Example 8.17 the value 40.1% is highly influential. If we drew another sample of 30 observations independent of this sample, the luck of the draw might give no values above 25, and the sample would yield very different conclusions. The bootstrap is a

useful method for making inferences from data, but it is dependent on a good sample. If this is all the data that we can get, we will never know how well our sample represents the distribution, and therefore how good our answer is. Of course, no statistical method will give good answers if the sample is not representative of the population.

Exercises: Section 8.5 (63–70)

63. In a survey, students gave their study time per week (h), and here are the 22 values:

15.0	10.0	10.0	15.0	25.0	7.0	3.0	8.0	10.0
10.0	11.0	7.0	5.0	15.0	7.5	7.5	12.0	7.0
10.5	6.0	10.0	7.5					

A 95% confidence interval for the population mean μ is desired.

a. Compute the t-based confidence interval of Section 8.2.
b. Create a normal probability plot. Is it apparent that the data set is not normal, so the t-based interval is of questionable validity?
c. Use software to generate a bootstrap sample of means. Create a histogram of the resulting $\bar{x}^*$ values.
d. Use the standard deviation for part (c) to get a 95% bootstrap t confidence interval for μ. Based on the histogram in part (c), is this CI valid?
e. Use part (c) to form the 95% confidence bootstrap percentile interval for μ.
f. Which interval should be used, and why?

64. Consider obtaining a 95% confidence interval for the population median $\tilde{\mu}$ of the study hours data in the previous exercise.

a. Use software to generate a bootstrap sample of medians.
b. Use the standard deviation for part (a) to get a 95% bootstrap t confidence interval for $\tilde{\mu}$.
c. Investigate the distribution of the bootstrap medians and discuss the validity of part (b).
d. Use the results of part (a) to form a 95% confidence bootstrap percentile interval for $\tilde{\mu}$.

e. For the study hours data, state your preference between the median and the mean, and explain your reasoning.

65. Here are 68 weight gains (lb) for pregnant women from conception to delivery ("Classifying Data Displays with an Assessment of Displays Found in Popular Software," *Teach. Statist.*, Autumn 2002: 96–101). A 95% CI for the population mean weight gain μ is desired.

25	14	20	38	21	22	36	38	35	37
35	24	31	28	25	32	23	30	39	26
38	20	21	11	35	42	31	25	59	23
43	38	21	76	22	26	10	19	25	25
15	31	34	36	35	33	24	44	35	43
7	32	25	27	31	14	25	16	25	47
35	−14	65	40	35	45	27	24		

a. Compute the t-based confidence interval of Section 8.2.
b. Check for normality to see if part (a) is valid. Is the sample large enough that the interval might be valid anyway?
c. Use software to generate a bootstrap sample of means. Create a histogram of the resulting $\bar{x}^*$ values.
d. Use the standard deviation for part (c) to get a 95% bootstrap t confidence interval for μ. Based on the histogram in part (c), is this CI valid?
e. Use part (c) to form the 95% confidence bootstrap percentile interval for μ.
f. Compare all three intervals. [*Note:* If they are all close, then the bootstrap supports the CI of part (a).]

66. Consider again the weight gain data from the previous exercise.

a. Use the method of Section 8.4 to obtain a 95% confidence interval for σ. Discuss

normality for the weight gain data: do you have reason to be concerned about the validity of this CI?

b. Use software to generate a bootstrap distribution of standard deviations. (That is, generate many resamples from the given data, and for each one compute the sample standard deviation s_i^*.)

c. Use the bootstrap standard deviation for part (a) to get a 95% bootstrap t confidence interval for σ.

d. Investigate the distribution of the bootstrap standard deviations and discuss the validity of part (c).

e. Use part (b) to form the 95% confidence bootstrap percentile interval for σ.

67. Nine Australian soldiers were subjected to extreme conditions, which involved a 100-min walk with a 25-lb pack when the temperature was 40 °C (104 °F). One of them overheated (above 39 °C) and was removed from the study. Here are the rectal Celsius temperatures of the other eight at the end of the walk ("Neural Network Training on Human Body Core Temperature Data," Combatant Protection and Nutrition Branch, Aeronautical and Maritime Research Laboratory of Australia, DSTO TN-0241, 1999):

38.4   38.7   39.0   38.5   38.5   39.0   38.5   38.6

a. Compute the t-based confidence interval of Section 8.2 for the population mean μ.

b. Check for the validity of part (a).

c. Use software to generate a bootstrap sample of means. Create a histogram of the resulting $\bar{x}^*$ values.

d. Use the standard deviation for part (c) to get a 95% bootstrap t confidence interval for μ. Based on the histogram in part (c), is this CI valid?

e. Use part (c) to form the 95% confidence bootstrap percentile interval for μ.

f. Compare the intervals and explain your preference.

g. Based on your knowledge of normal body temperature, would you say that body temperature can be influenced by environment?

68. Refer back to the body temperature data in the previous exercise.

a. Obtain a bootstrap sample of 12.5% trimmed means. [*Hint:* With $n = 8$, a 12.5% trimmed mean entails deleting the largest and smallest value in each resample.]

b. Use the standard deviation from the bootstrap samples in part (a) to get a 95% bootstrap t confidence interval for the population 12.5% trimmed mean μ_{tr}.

c. Investigate the distribution of the bootstrap trimmed means and discuss the validity of the interval in part (b).

d. Use the results of part (a) to form a 95% confidence bootstrap percentile interval for μ_{tr}.

e. Compare all the intervals for the mean μ and trimmed mean μ_{tr}. Are they fairly similar? How do you explain that?

69. If you go to a major league baseball game, how long do you expect the game to be? From the 2430 games played in 2018, here is a random sample of 25 times (min):

168	187	161	205	162	183	186	190	136
177	182	185	185	194	169	151	192	181
194	162	194	171	172	168	174		

This is one of those rare instances in which we can calculate a confidence interval and compare with the actual population mean. The mean duration of all 2430 games was $\mu = 184.94$ min (a little more than 3 h), but pretend we don't know that.

a. Compute the t-based confidence interval of Section 8.2.

b. Use a normal probability plot to see if part (a) is valid.

c. Use software to generate a bootstrap sample of means.

d. Use the standard deviation for part (c) to get a 95% bootstrap t confidence interval for μ.

e. Use part (c) to form the 95% confidence bootstrap percentile interval for μ.

f. Say which interval should be used and explain why. Does your interval include the true value, $\mu = 184.94$ min?

70. Because of extra-inning games, the median might be a more meaningful statistic for the length-of-game data in the previous exercise. The median length of all 2430 MLB games in 2018 was $\tilde{\mu} = 182$ min.

a. Use software and the data in the previous exercise to obtain a bootstrap sample of medians.

b. Obtain a 95% confidence bootstrap t interval for the population median.

c. Investigate the distribution of the bootstrap medians and discuss the validity of part (b).

d. Determine a 95% confidence bootstrap percentile interval for the median. Compare your answer with the population median.

Supplementary Exercises (71–92)

71. A manufacturer of college textbooks is interested in estimating the strength of the bindings produced by a particular binding machine. Strength can be measured by recording the force required to pull the pages from the binding. If this force is measured in pounds, how many books should be tested to estimate the average force required to break the binding to within .1 lb with 95% confidence? Assume that σ is known to be .8.

72. According to the article "Fatigue Testing of Condoms" (*Polymer Testing* 2009: 567–571), "tests currently used for condoms are surrogates for the challenges they face in use," including a test for holes, an inflation test, a package seal test, and tests of dimensions and lubricant quality (all fertile territory for the use of statistical methodology!). The investigators developed a new test that adds cyclic strain to a level well below breakage and determines the number of cycles to break. A sample of 20 condoms of one particular type resulted in a sample mean number of 1584 and a sample standard deviation of 607. Calculate and interpret a confidence interval at the 99% confidence level for the true average number of cycles to break. [*Note*: The article presented the results of hypothesis tests based on the t distribution; the validity of these depends on assuming normal population distributions.]

73. Before opening a new location, franchise companies conduct market research to determine if sufficient demand exists for their products. A national sandwich chain recently conducted a survey to investigate opening a franchise in a particular town. Among 300 households contacted through random-digit dialing, 198 respondents indicated they would patronize this shop.

a. Let $p =$ the proportion of all households in this town that would patronize the sandwich franchise. Calculate and interpret a 95% lower confidence bound for p.

b. From years of marketing experience, the company knows they need more than 5000 households in the population to patronize the shop—this accounts for competing local businesses and variation in frequency of visitation by potential patrons. This particular town has 7700 households. Determine a 95% lower confidence bound for the *number* of households that will eat at the new store. Can the company be confident they will have enough customers?

c. Imagine the company ignored sampling variability and simply used the sample proportion from the survey to determine the expected number of customers (rather than the lower confidence bound). Would that change their opinion regarding the viability of the new location? Explain.

74. The Pew Forum on Religion and Public Life reported on Dec. 9, 2009 that in a survey of 2003 American adults, 25% said they believed in astrology.

   a. Calculate and interpret a confidence interval at the 99% confidence level for the proportion of all adult Americans who believe in astrology.

   b. What sample size would be required for the width of a 99% CI to be at most .05 irrespective of the value of $\hat{p}$?

   c. The upper limit of the CI in (a) gives an upper confidence bound for the proportion being estimated. What is the corresponding confidence level?

75. There were 12 first-round heats in the men's 100-m race at the 1996 Atlanta Summer Olympics. Here are the reaction times in seconds (time to first movement) of the top four finishers of each heat. The first 12 are the 12 winners, then the second-place finishers, and so on.

1st	.187	.152	.137	.175	.172	.165
	.184	.185	.147	.189	.172	.156
2nd	.168	.140	.214	.163	.202	.173
	.175	.154	.160	.169	.148	.144
3rd	.159	.145	.187	.222	.190	.158
	.202	.162	.156	.141	.167	.155
4th	.156	.164	.160	.145	.163	.170
	.182	.187	.148	.183	.162	.186

Because reaction time has little if any relationship to the order of finish, it is reasonable to view the times as coming from a single population.

   a. Estimate the population mean in a way that conveys information about precision and reliability. [*Note*: $\sum x_i = 8.08100$, $\sum x_i^2 = 1.37813$.]

   b. Calculate a 95% confidence interval for the population proportion of reaction times that are below .15. (Reaction times below .10 are regarded as false starts, meaning that the runner anticipates the starter's gun, because such times are considered physically impossible. Linford Christie, who had a reaction time of .160 in placing second in his first-round heat, had two such false starts in the finals and was disqualified.)

76. Aphid infestation of fruit trees can be controlled either by spraying with pesticide or by inundation with ladybugs. In a particular area, four different groves of fruit trees are selected for experimentation. The first three groves are sprayed with pesticides 1, 2, and 3, respectively, and the fourth is treated with ladybugs, with the following results on yield:

Treatment	n_i (number of trees)	$\bar{x}_i$ (bushels/tree)	s_i
1	100	10.5	1.5
2	90	10.0	1.3
3	100	10.1	1.8
4	120	10.7	1.6

Let μ_i = the true average yield (bushels/tree) after receiving the ith treatment. Then

$$\theta = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) - \mu_4$$

measures the difference in true average yields between treatment with pesticides and treatment with ladybugs. When n_1, n_2, n_3, and n_4 are all large, the estimator $\hat{\theta}$ obtained by replacing each μ_i by $\bar{X}_i$ is approximately normal. Use this to derive a large-sample $100(1 - \alpha)\%$ CI for θ, and compute the 95% interval for the given data.

77. It is important that face masks used by firefighters be able to withstand high temperatures because firefighters commonly work in temperatures of 200–500 °F. In a test of one type of mask, 11 of 55 masks had lenses pop out at 250°. Construct a 90% CI for the true proportion of masks of this type whose lenses would pop out at 250°.

78. A journal article reports that a sample of size 5 was used as a basis for calculating a 95% CI for the true average natural frequency (Hz) of delaminated beams of a

certain type. The resulting interval was (229.764, 233.504). You decide that a confidence level of 99% is more appropriate than the 95% level used. What are the limits of the 99% interval? [*Hint*: Use the center of the interval and its width to determine $\bar{x}$ and s.]

79. The article "The Association Between Television Viewing and Irregular Sleep Schedules Among Children Less Than 3 Years of Age" (*Pediatrics* 2005: 851–856) reported the following 95% confidence intervals for average TV viewing time (hours per day) for three different age groups.

0–11 months old	12–23 months old	24–35 months old
(0.8, 1.0)	(1.4, 1.8)	(2.1, 2.5)

  a. Interpret each of these three intervals.
  b. The three intervals are not the same width. What might explain this?
  c. Do the intervals suggest a relationship between age and TV viewing time among children of this age range? Explain.

80. In Example 7.12, we introduced the concept of a censored experiment in which n components are put on test and the experiment terminates as soon as r of the components have failed. Suppose component lifetimes are independent, each having an exponential distribution with parameter λ. Let Y_1 denote the time at which the first failure occurs, Y_2 the time at which the second failure occurs, and so on, so that $T_r = Y_1 + \cdots + Y_r + (n - r)Y_r$ is the total accumulated lifetime at termination. Then it can be shown that $2\lambda T_r$ has a chi-squared distribution with $2r$ df. Use this fact to develop a $100(1 - \alpha)\%$ CI formula for true average lifetime $1/\lambda$. Compute a 95% CI from the data in Example 7.12.

81. Exercises 77–78 from Chapter 7 introduced "regression through the origin" to relate a dependent variable y to an independent variable x. The assumption there was that for any fixed x value, the dependent variable is a random variable Y with mean value βx and variance σ^2 (so that Y has mean value zero when $x = 0$). The data consists of n independent (x_i, Y_i) pairs, where each Y_i is normally distributed with mean βx_i and variance σ^2. The likelihood is then a product of normal pdfs with different mean values but the same variance.

  a. Show that the mle of β is $\hat{\beta} = \Sigma x_i Y_i / \Sigma x_i^2$.
  b. Verify that the mle of (a) is unbiased.
  c. Obtain an expression for $V(\hat{\beta})$ and then for $\sigma_{\hat{\beta}}$.
  d. For purposes of obtaining a precise estimate of β, is it better to have the x_i's all close to 0 (the origin) or quite far from 0? Explain your reasoning.
  e. The natural prediction of Y_i is $\hat{\beta} x_i$. Let $S^2 = \Sigma (Y_i - \hat{\beta} x_i)^2 / (n - 1)$, which is analogous to sample variance. It can be shown that $T = (\hat{\beta} - \beta) / \left(S / \sqrt{\Sigma x_i^2} \right)$ has a t distribution with $n - 1$ df. Use this to obtain a CI formula for estimating β, and calculate a 95% CI using the data from the cited exercises.

82. Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $[0, \theta]$ and $Y = \max(X_1, \ldots, X_n)$. Then methods from Section 5.7 can be used to show that the rv $U = Y/\theta$ has pdf

$$f_U(u) = nu^{n-1} \quad 0 \le u \le 1$$

  a. Verify that

$$P\left[(\alpha/2)^{1/n} \le \frac{Y}{\theta} \le (1 - \alpha/2)^{1/n} \right] = 1 - \alpha$$

  and use this to derive a $100(1 - \alpha)\%$ CI for θ.
  b. Verify that $P(\alpha^{1/n} \le Y/\theta \le 1) = 1 - \alpha$, and derive a $100(1 - \alpha)\%$ CI for θ based on this probability statement.

c. Which of the two intervals derived in (a) and (b) is shorter? If your waiting time for a morning bus is uniformly distributed and observed waiting times are $x_1 = 4.2$, $x_2 = 3.5$, $x_3 = 1.7$, $x_4 = 1.2$, and $x_5 = 2.4$, obtain a 95% CI for θ by using the shorter of the two intervals.

83. Let $0 < \gamma < \alpha$. Then a $100(1 - \alpha)$% CI for μ when n is large is

$$\left(\bar{x} - z_\gamma \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha-\gamma} \cdot \frac{s}{\sqrt{n}}\right)$$

The choice $\gamma = \alpha/2$ yields the large-sample interval derived in Section 8.2; if $\gamma \neq \alpha/2$, this confidence interval is not symmetric about $\bar{x}$. The width of the interval is $w = s(z_\gamma + z_{\alpha-\gamma})/\sqrt{n}$. Show that w is minimized for the choice $\gamma = \alpha/2$, so that the symmetric interval is the shortest. [*Hints*: (1) By definition of z_α, $\Phi(z_\alpha) = 1 - \alpha$, so that $z_\alpha = \Phi^{-1}(1 - \alpha)$; (2) the relationship between the derivative of a function $y = f(x)$ and the inverse function $x = f^{-1}(y)$ is $(d/dy)f^{-1}(y) = 1/f'(x)$.]

84. Suppose x_1, x_2, …, x_n are observed values resulting from a random sample from a symmetric but possibly heavy-tailed distribution. Chapter 11 of *Understanding Robust and Exploratory Data Analysis* (see the bibliography) suggests the following robust 95% CI for the population mean (point of symmetry):

$$\tilde{x} \pm \left(\frac{\text{conservative } t \text{ critical value}}{1.075}\right) \cdot \frac{\text{iqr}}{\sqrt{n}}$$

The value of the quantity in parentheses is 2.10 for $n = 10$, 1.94 for $n = 20$, and 1.91 for $n = 30$. Compute this CI for the restaurant tip data of Example 8.17, and compare to the t CI appropriate for a normal population distribution.

85. a. Use the results of Example 8.5 to obtain a 95% lower confidence bound for the parameter λ of an exponential

distribution, and calculate the bound based on the data given in the example.

b. If lifetime X has an exponential distribution, the probability that lifetime exceeds t is given by $P(X > t) = e^{-\lambda t}$. Use the result of part (a) to obtain a 95% lower confidence bound for the probability that lifetime exceeds 100 min.

86. Let θ_1 and θ_2 denote the mean weights for animals of two different species. A biologist wishes to estimate the ratio θ_1/θ_2. Unfortunately the species are extremely rare, so the estimate will be based on finding a single animal of each species. Let X_i denote the weight of the species i animal ($i = 1$, 2), assumed to be normally distributed with mean θ_i and standard deviation 1.

a. Show that the rv $h(X_1, X_2; \theta_1, \theta_2) = (\theta_2 X_1 - \theta_1 X_2)/\sqrt{\theta_1^2 + \theta_2^2}$ is a pivotal quantity by determining the distribution of h.

b. Show that h depends on θ_1 and θ_2 only through θ_1/θ_2. [*Hint:* Divide numerator and denominator by θ_2.]

c. Consider Expression (8.7) from the first section of this chapter with $a = -1.96$ and $b = 1.96$. Now replace $<$ by $=$ and solve for θ_1/θ_2. Then show that a confidence interval results if $x_1^2 + x_2^2 \geq 1.96^2$, whereas if this inequality is not satisfied, the resulting *confidence set* is the complement of an interval.

87. The one-sample CI for a normal mean and PI for a single observation from a normal distribution were both based on the *central t* distribution. A CI for a particular percentile (e.g., the 1st percentile or the 95th percentile) of a normal population distribution is based on the *noncentral t* distribution. A particular distribution of this type is specified by both df and the value of the noncentrality parameter δ ($\delta = 0$ gives the central t distribution). The key result is that the variable

$$T = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} - (z \text{ percentile})\sqrt{n}}{S/\sigma}$$

has a noncentral t distribution with df $=$ $n - 1$ and $\delta = -(z \text{ percentile})\sqrt{n}$.

Let $t_{.025,v,\delta}$ and $t_{.975,v,\delta}$ denote the critical values that capture upper-tail area .025 and lower-tail area .025, respectively, under the noncentral t curve with v df and noncentrality parameter δ (when $\delta = 0$, $t_{.975} = -t_{.025}$, since central t distributions are symmetric about 0).

a. Use the given information to obtain a formula for a 95% confidence interval for the $(100p)$th percentile of a normal population distribution.

b. For $\delta = 6.58$ and df = 15, $t_{.975}$ and $t_{.025}$ are (from software) 4.1690 and 10.9684, respectively. Use this information to obtain a 95% CI for the 5th percentile of the beer alcohol distribution considered in Exercise 17.

88. In this exercise, we develop a CI for $\tilde{\mu}$ that is valid whatever the shape of the population distribution as long as it is continuous. Let $X_1, \ldots, X_n$ be a random sample from the distribution and $Y_1 < \cdots < Y_n$ denote the corresponding ordered values (smallest observation, second smallest, and so on).

a. What is $P(X_1 < \tilde{\mu})$? What is $P(\{X_1 < \tilde{\mu}\} \cap \{X_2 < \tilde{\mu}\})$?

b. What is $P(Y_n < \tilde{\mu})$? What is $P(Y_1 > \tilde{\mu})$? [Hint: What condition involving all of the X_i's is equivalent to the largest being smaller than the population median?]

c. What is $P(Y_1 < \tilde{\mu} < Y_n)$? What does this imply about the confidence level associated with the CI (y_1, y_n) for $\tilde{\mu}$?

d. An experiment carried out to study the time (min) necessary for an anesthetic to produce the desired result yielded the following data: 31.2, 36.0, 31.5, 28.7, 37.2, 35.4, 33.3, 39.3, 42.0, 29.9. Determine the confidence interval of (c) and the associated confidence level.

89. Consider the situation described in the previous exercise.

a. What is $P(\{X_1 < \tilde{\mu}\} \cap \{X_2 > \tilde{\mu}\} \cap \cdots \cap \{X_n > \tilde{\mu}\})$, that is, the probability that only the first observation is smaller than the median?

b. What is the probability that exactly one of the n original observations is smaller than the median?

c. What is $P(\tilde{\mu} < Y_2)$? [Hint: The event in parentheses occurs if all n of the observations exceed the median. How else can it occur?]

d. What is $P(Y_2 < \tilde{\mu} < Y_{n-1})$? What does this imply about the confidence level associated with the CI (y_2, y_{n-1}) for $\tilde{\mu}$?

e. Determine the confidence level and CI using part (d) with the data given in the previous exercise.

90. The previous two exercises considered a CI for a population median $\tilde{\mu}$ based on the ordered values from a random sample. Let's now consider a *prediction* interval for the next observation X_{n+1}, which is assumed to be independent of $X_1, \ldots, X_n$.

a. What is $P(X_{n+1} < X_1)$? What is $P(\{X_{n+1} < X_1\} \cap \{X_{n+1} < X_2\})$?

b. What is $P(X_{n+1} < Y_1)$? What is $P(X_{n+1} > Y_n)$?

c. What is $P(Y_1 < X_{n+1} < Y_n)$? What does this say about the prediction level for the PI (y_1, y_n)? Determine the prediction level and interval for the data in the previous two exercises.

91. Consider 95% CIs for two different parameters θ_1 and θ_2, and let A_i ($i = 1, 2$) denote the event that the value of θ_i is included in the random interval that results in the CI. Thus $P(A_i) = .95$.

a. Suppose that the data on which the CI for θ_1 is based is independent of the data used to obtain the CI for θ_2 (e.g., we might have $\theta_1 = \mu$, the population mean height for American females, and $\theta_2 = p$, the proportion of all iPhones that don't need warranty service). What

can be said about the *simultaneous* confidence level for the two intervals? That is, how confident can we be that the first interval contains the value of θ_1 *and* that the second contains the value of θ_2? [*Hint*: Consider $P(A_1 \cap A_2)$.]

b. Now suppose the data for the first CI is not independent of that for the second one. What now can be said about the simultaneous confidence level for both intervals? [*Hint*: Consider $P(A_1' \cup A_2')$, the probability that at least one interval fails to include the value of what it is estimating. Now use the fact that $P(A_1' \cup A_2') \leq P(A_1') + P(A_2')$. The generalization of the bound on $P(A_1' \cup A_2')$ to the probability of a k-fold union is one version of the *Bonferroni* inequality.]

c. What can be said about the simultaneous confidence level if the confidence level for each interval separately is $100(1 - \alpha)\%$? What can be said about the simultaneous confidence level if a $100(1 - \alpha)\%$ CI is computed separately for each of k parameters $\theta_1, \ldots, \theta_k$?

92. The *Bonett CI* for a population variance σ^2 mentioned at the end of Section 8.4, unlike the chi-squared method, does not hinge on population normality. This interval involves a transformation along with an estimate of the *kurtosis* of the underlying

distribution, a measure of its "tail" behavior. Specifically, Bonett defines a kurtosis estimate by

$$\bar{\gamma}_4 = \frac{n \sum (x_i - \bar{x}_{\text{tr}})^4}{\left(\sum (x_i - \bar{x})^2 \right)^2}$$

where $\bar{x}_{\text{tr}}$ is the trimmed mean with trim proportion $1/[2\sqrt{n - 4}]$. Then the Bonett CI for σ^2 with confidence level $100(1 - \alpha)\%$ has endpoints

$$\exp\left[\ln(c \cdot S^2) \pm z_{\alpha/2} \cdot c \cdot \sqrt{\frac{(n - 3)\bar{\gamma}_4}{n(n - 1)}} \right]$$

where $c = n/(n - z_{\alpha/2})$ is "an empirically determined, small-sample adjustment" (meaning Bonett found this value by trial and error).

a. For the study hours data in Exercise 63, $n = 22$, $s = 4.603$ and $\bar{\gamma}_4 = 7.003$. Use Bonett's formula to calculate a 95% CI for the population variance σ^2.

b. Use part (a) to determine a 95% CI for σ.

c. Show that as $n \to \infty$, both endpoints of the Bonett CI converge to σ^2. [*Hint:* The kurtosis estimate $\bar{\gamma}_4$ converges to a constant, while $S^2 \to \sigma^2$.]

Tests of Hypotheses Based on a Single Sample

Introduction

A parameter can be estimated from sample data either by a single number (a point estimate) or an entire interval of plausible values (a confidence interval). Frequently, however, the objective of an investigation is not to estimate a parameter but to decide which of two contradictory claims about the parameter is correct. Methods for accomplishing this comprise the part of statistical inference called *hypothesis testing*. In this chapter, we first discuss some of the basic concepts and terminology in hypothesis testing and then develop decision procedures for the most frequently encountered testing problems based on a sample from a single population.

9.1 Hypotheses and Test Procedures

A **statistical hypothesis**, or just *hypothesis*, is a claim or assertion either about the value of a single parameter (i.e., a characteristic of a population or a probability distribution), about the values of several parameters, or about the form of an entire probability distribution. Examples include

- The claim $\mu = \$311$, where μ is the true average one-term textbook expenditure for students at a university
- The statement $p < .50$, where p is the proportion of adults who approve of the job that the President is doing
- The assertion that $\mu_1 - \mu_2 > 5$, where μ_1 and μ_2 denote the true average decreases in systolic blood pressure for two different drugs
- The claim that stopping distance for a car under particular conditions has a normal distribution.

Hypotheses of the last sort will be considered briefly in Chapter 13. In this and the next several chapters, we concentrate on hypotheses about parameters.

In any hypothesis-testing problem, there are two contradictory hypotheses under consideration. One hypothesis might be the claim $\mu = \$311$ and the other $\mu \neq \$311$, or the two contradictory statements might be $p \geq .50$ and $p < .50$. The objective is to decide, based on sample information, which of the two hypotheses is correct. In statistics, hypothesis-testing problems are formulated so that one of the claims is initially assumed to be true. This initial claim will not be rejected in favor of the alternative claim unless sample evidence provides strong evidence for the latter.

DEFINITION   The **null hypothesis**, denoted by H_0, is the claim that is initially assumed to be true (the "prior belief" claim). The **alternative hypothesis**, denoted by H_a, is the assertion that is contradictory to H_0.

The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, we will continue to believe in the plausibility of the null hypothesis. The two possible conclusions from a hypothesis-testing analysis are then *reject H_0* or *fail to reject H_0*.

A **test of hypotheses** is any method for using sample data to decide whether the null hypothesis should be rejected. Thus if we test H_0: $\mu = \$311$ against the alternative H_a: $\mu \neq \$311$, the null hypothesis should be rejected only if sample data strongly suggests that μ is something other than $311. In the absence of strong evidence, H_0 should not be rejected since it is still judged to be plausible.

There is a familiar analogy to this in a criminal trial. One claim is the assertion that the defendant is innocent. In the U.S. judicial system, this is the claim that is initially believed to be true. Only in the face of strong evidence to the contrary should the jury reject this claim in favor of the alternative assertion that the accused is guilty. In this sense, the claim of innocence is the favored or protected hypothesis, and the burden of proof is placed on those who believe in the alternative claim.

Formulating Hypotheses

Sometimes an investigator does not want to accept a particular assertion unless and until data can provide strong support for the assertion. In that situation, this assertion will be the investigator's alternative hypothesis H_a. (Examples will be given shortly.) Scientific research often involves trying to decide whether a current theory should be replaced by a more plausible and satisfactory explanation of the phenomenon under investigation. A conservative approach is to identify the current theory with H_0 and the researcher's alternative explanation with H_a. Rejection of the current theory will then occur only when evidence is much more consistent with the new theory. In many situations, H_a is referred to as the "research hypothesis," since it is the claim that the researcher would really like to validate. The word *null* means "of no value, effect, or consequence," which suggests that H_0 should be identified with the hypothesis of no change (from current opinion), no difference, no improvement, and so on.

Example 9.1 Many have heard the claim that college students gain an average of 15 lb during their first year, but is this popular legend rooted in reality? This was the subject of the article "The Effects of College on Weight: Examining the 'Freshman 15' Myth and Other Effects of College Over the Life Cycle" (*Demography* 2017: 311–336). Let μ denote the true average weight gain of students over the course of their first year in college. We initially give the "freshman 15" story the benefit of the doubt, so that our null hypothesis is H_0: $\mu = 15$. It would be noteworthy if 15 were an underestimate or an overestimate, suggesting that the alternative hypothesis should be H_a: $\mu \neq 15$.   ∎

Example 9.2 Vintners and wine consumers continue to debate whether to seal wine bottles with corks or screwtops. Screwtops can reduce spoilage, but many wine enthusiasts associate them with cheap or otherwise undesirable wines. An often-cited 2011 survey by Rebecca Bleibaum (Tragon Corp.) found that about 20% of wine consumers would not buy screwtop wine, but negative attitudes toward screwtops have abated over time. Let p denote the proportion of wine consumers *today* who refuse to purchase screwtop wine. The null hypothesis is that no change has occurred since that previous survey, H_0: $p = .2$. A winery that was considering switching one of its wines from cork to

screwtop bottling would naturally be interested in the alternative hypothesis that this proportion has decreased, H_a: $p < .2$. ∎

Example 9.3 *Consumer Reports* (Nov. 26, 2018) reported that some American automakers are reducing sedan production in response to the increasing popularity of trucks and SUVs, despite the fact that sedans typically get better gas mileage. Extensive experience with engines for a certain type of light-duty truck indicates that highway fuel efficiency (miles per gallon) is normally distributed with a mean value of 25 and a standard deviation of 3. The manufacturer is considering a modification to increase average fuel efficiency. Let μ denote true average efficiency for the new, modified engines. The appropriate null (no-improvement) hypothesis is H_0: $\mu = 25$. The alternative hypothesis asserts that there has, in fact, been an improvement: H_a: $\mu > 25$.

Sample data will be collected from modified engines. Because of the expense of changing the manufacturing process, the new engine design will only be adopted if the data provides *convincing* evidence that μ really is greater than 25 mpg. ∎

In our treatment of hypothesis testing, H_0 will generally be stated as an equality claim. If θ denotes the parameter of interest, the null hypothesis will have the form H_0: $\theta = \theta_0$, where θ_0 is a specified number called the **null value** of the parameter (i.e., the value claimed for θ by the null hypothesis). For instance, consider the truck gas mileage situation of Example 9.3. The alternative hypothesis is H_a: $\mu > 25$, the claim that the mean fuel efficiency is improved by the engine modification. The null hypothesis was stated as H_0: $\mu = 25$, so the null value of the parameter is $\mu_0 = 25$. But it would be more mathematically natural to write H_0: $\mu \leq 25$, according to which the new engine either is no better *or* is worse than the one currently used. The rationale for using a simplified null hypothesis is that any reasonable procedure for deciding between H_0: $\mu = 25$ and H_a: $\mu > 25$ will also be reasonable for deciding between the claim that $\mu \leq 25$ and H_a, and should lead to exactly the same conclusion for any particular sample. The use of a simplified H_0 is preferred because it has certain technical benefits, which will become apparent shortly.

The alternative to the null hypothesis H_0: $\theta = \theta_0$ will look like one of the following three assertions:
1. H_a: $\theta > \theta_0$ (in which case the implicit null hypothesis is $\theta \leq \theta_0$)
2. H_a: $\theta < \theta_0$ (so the implicit null hypothesis states that $\theta \geq \theta_0$)
3. H_a: $\theta \neq \theta_0$

Test Procedures

A test procedure is a rule, based on sample data, for deciding whether to reject H_0. A test of H_0: $p = .2$ versus H_a: $p < .2$ in Example 9.2 might be based on surveying a random sample of $n = 200$ current wine consumers. Let X denote the number of people in the sample who refuse to buy screwtop wine, a binomial random variable (at least approximately); let x represent the observed value of X. If H_0 is true, $E(X) = np = 200(.2) = 40$, whereas we can expect fewer than 40 refusers if H_a is true. An x value just a bit below 40 does not *strongly* contradict H_0, so it is reasonable to reject H_0 in favor of H_a only if x is substantially less than 40. One such test procedure is to reject H_0 if $x \leq 35$ and not reject H_0 otherwise. This procedure has two elements: (1) a *test statistic*, or function of the sample data, used to make a decision; and (2) a *rejection region* consisting of those test statistic values for which H_0 will be rejected in favor of H_a. In the wine scenario, X is the test statistic and the rejection region consists of $x = 0, 1, 2, \ldots, 35$; H_0 will not be rejected if $x = 36, 37, \ldots, 199$, or 200.

DEFINITION   A test procedure is specified by the following:

1. A **test statistic**, a function of the sample data on which the decision (reject H_0 or do not reject H_0) is to be based
2. A **rejection region**, the set of all test statistic values for which H_0 will be rejected

The null hypothesis will then be rejected if and only if the observed or computed test statistic value falls in the rejection region.

In the context of Example 9.3, let $\bar{X}$ denote the sample average highway fuel efficiency of a random sample of 10 trucks with the new, modified engine. If H_0 is true, $E(\bar{X}) = \mu = 25$, whereas if H_0 is false, we expect $\bar{X}$ to exceed 25. Strong evidence against H_0 is provided by a value $\bar{x}$ that considerably exceeds 25. Thus we might use $\bar{X}$ as a test statistic along with the rejection region $\bar{x} \geq 30$.

In both the wine and truck examples, the choices of the test statistic and the *form* of the rejection region make sense intuitively. However, the choice of cutoff value used to specify the rejection region was somewhat arbitrary. Instead of rejecting H_0: $p = .2$ in favor of H_a: $p < .2$ when $x \leq 35$, we could use the rejection region $x \leq 30$. For this region, H_0 would not be rejected if 33 respondents refused to buy screwtop wine, whereas this occurrence would lead to rejection of H_0 if the initially suggested region were employed. Similarly, the rejection region $\bar{x} \geq 27.5$ might be used in the truck engine problem in place of the region $\bar{x} \geq 30$. We'll discuss shortly the tradeoffs between different rejection region cutoffs and how they are most often determined in practice.

Errors in Hypothesis Testing

When a jury is called upon to render a verdict in a criminal trial, there are two possible erroneous conclusions: convicting an innocent person, or letting a guilty person go free. Similarly, in statistical hypothesis testing there are two potential errors whose consequences must be considered when reaching a conclusion.

DEFINITION   A **type I error** consists of rejecting the null hypothesis H_0 when it is true.
A **type II error** involves not rejecting H_0 when it is false (i.e., H_a is true).

Since in the U.S. judicial system the null hypothesis (a priori belief) is that the accused is innocent, a type I error is analogous to convicting an innocent person, while a false acquittal (i.e., letting a guilty person go free) equates to a type II error.

Example 9.4 (Example 9.3 continued) Before selecting a test procedure and collecting data, the truck manufacturer must consider the possible type I and type II errors along with their consequences. In this scenario, a type I error means that the manufacturer concludes the modified engine design improves fuel efficiency when, in fact, it does not. Thus a type I error would lead the manufacturer to perform a very expensive but ultimately useless overhaul of its truck engines. Because this is such a consequential error, a test procedure should be selected that makes the chance of a type I error very small: if the modified engine design is truly no better than the old one (i.e., H_0 is true), this would ensure a low probability of mistakenly rejecting H_0 and proceeding with the change.

Balanced against this possibility is the threat of a type II error: failing to reject H_0 when, in fact, H_a: $\mu > 25$ is correct. That is, in a type II error the manufacturer would *fail to recognize* that the modified engine design improves fuel efficiency and would continue to use the old, inferior design. A type II error is often called an *opportunity loss* in business: the manufacturer has missed out on the opportunity to build, sell, and profit from a superior engine design. ∎

It would be nice if test procedures could be developed that offered 100% protection against committing both a type I error and a type II error. This is an impossible goal, though, because our conclusion is based on sample data rather than a census of the entire population. There is always some chance that random sampling variability will lead to an incorrect conclusion. Instead of demanding error-free procedures, we must look for procedures for which both types of error are unlikely to occur. That is, a good procedure is one for which the probability of making either type of error is small. The choice of a particular rejection region cutoff value fixes the probabilities of type I and type II errors. These error probabilities are traditionally denoted by α and β, respectively. Because H_0 specifies a unique value of the parameter, there is a single value of α. However, there is a different value of β for each value of the parameter consistent with H_a.

Example 9.5 (Example 9.2 continued) A small winemaker will conduct a pilot study by surveying $n = 25$ randomly selected customers about their views on screwtop wine bottles. The parameter of interest is now p = the proportion of *this winery's* customers who refuse to buy screwtop wine, but the hypotheses will remain H_0: $p = .2$ versus H_a: $p < .2$. Consider the following test procedure:
   Test statistic: X = the number of surveyed customers who will not buy screwtop wine bottles
   Rejection region: $R_3 = \{0, 1, 2, 3\}$; that is, reject H_0 if $x \leq 3$,
            where x is the observed value of the test statistic.
This rejection region is called *lower-tailed* because it consists only of small values of the test statistic.
   When H_0 is true, X has a binomial probability distribution with $n = 25$ and $p = .2$. Then

$$\begin{aligned} \alpha &= P(\text{type I error}) = P(H_0 \text{ is rejected when it is true}) \\ &= P(X \leq 3 \text{ when } X \sim \text{Bin}(25, .2)) = B(3;\ 25,\ .2) \\ &= .234 \end{aligned}$$

That is, if H_0 is actually true, 23.4% of all pilot surveys consisting of 25 customers would result in H_0 being incorrectly rejected (a type I error). This error probability is quite large; we will consider shortly how it can be made smaller.
   In contrast to α, there is not a single β. Instead, there is a different β for each different p less than .2. Thus there is a value of β for $p = .15$ [in which case $X \sim \text{Bin}(25, .15)$], another value of β for $p = .1$, and so on. For example,

$$\begin{aligned} \beta(.1) &= P(\text{type II error when } p = .1) \\ &= P(H_0 \text{ is not rejected when it is false because } p = .1) \\ &= P(X > 3 \text{ when } X \sim \text{Bin}(25,\ .1)) = 1 - B(3;\ 25, .1) = .236 \end{aligned}$$

When p is actually .1 rather than .2 (a rather large departure from H_0), roughly 24% of all surveys of this type would result in H_0 being incorrectly not rejected.
   The accompanying table displays β for selected values of p (each calculated for the rejection region R_3). Clearly, β decreases as the value of p moves farther below the null value .2. Intuitively, the greater the departure from H_0, the less likely it is that such a departure will sneak past undetected.

p	.19	.15	.12	.10	.08	.04
$\beta(p)$	.727	.529	.352	.236	.135	.017

Many of these values of β values are unacceptably large, due in part to the relatively small sample size.

The proposed test procedure is still reasonable for testing the more mathematically correct null hypothesis that $p \geq .2$. In this case, there is no longer a single α, but instead there is an α for each p that is at least .2: $\alpha(.2)$, $\alpha(.25)$, $\alpha(.314)$, $\alpha(.325)$, and so on. It is easily verified, though, that $\alpha(p) < \alpha(.2) = .234$ for all $p > .2$. That is, the largest value of α occurs for the boundary value .2 between H_0 and H_a. Thus whatever the probability α is for the simplified null hypothesis, it will be no larger for the more realistic H_0.   ∎

Example 9.6 (Example 9.3 continued) Let $X_1, \ldots, X_{10}$ denote the highway fuel efficiencies (mpg) of 10 randomly selected trucks with the new, modified engine. (If $n = 10$ seems small, bear in mind that vehicles used for testing often cannot then be sold to customers.) Under the assumptions of Example 9.3, $X_1, \ldots, X_{10}$ is a random sample of size 10 from a normal distribution with mean value μ and standard deviation $\sigma = 3$. To test H_0: $\mu = 25$ versus H_a: $\mu > 25$, consider the following test procedure:

Test statistic: $\bar{X}$ = the sample mean fuel efficiency of the 10 randomly selected trucks

Rejection region: $R = [27.5, \infty)$; that is, reject H_0 if $\bar{x} \geq 27.5$,

where $\bar{x}$ is the observed value of the test statistic

Because the rejection region consists only of large values of the test statistic, the test is said to be *upper-tailed*.

The sample mean fuel efficiency $\bar{X}$ then has a normal distribution with $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 3/\sqrt{10} \approx .95$. Calculation of α and β now involves a routine standardization of $\bar{X}$ followed by reference to the standard normal probabilities of Appendix Table A.3:

$$\alpha = P(\text{type I error}) = P(H_0 \text{ is rejected when it is true})$$
$$= P(\bar{X} \geq 27.5 \text{ when } \bar{X} \sim \text{normal with } \mu_{\bar{X}} = 25, \sigma_{\bar{X}} = .95)$$
$$= 1 - \Phi\left(\frac{27.5 - 25}{.95}\right) = 1 - \Phi(2.63) = .0042$$
$$\beta(26.5) = P(\text{type II error when } \mu = 26.5)$$
$$= P(H_0 \text{ is not rejected when it is false because } \mu = 26.5)$$
$$= P(\bar{X} < 27.5 \text{ when } \bar{X} \sim \text{normal with } \mu_{\bar{X}} = 26.5, \sigma_{\bar{X}} = .95)$$
$$= \Phi\left(\frac{27.5 - 26.5}{.95}\right) = \Phi(1.05) = .8531$$
$$\beta(28) = \Phi\left(\frac{27.5 - 28}{.95}\right) = .2993 \qquad \beta(29) = .0571$$

For the specified test procedure, only 0.4% of all experiments carried out as described will result in H_0 being rejected when it is actually true. However, the chance of a type II error is very large when $\mu = 26.5$ (only a small departure from H_0), somewhat less when $\mu = 28$, and quite small when $\mu = 29$ (a rather large departure from H_0). These error probabilities are illustrated in Figure 9.1. Notice that α is computed using the probability distribution of the test statistic when H_0 is true, whereas determination of β requires knowing the test statistic's distribution when H_0 is false.
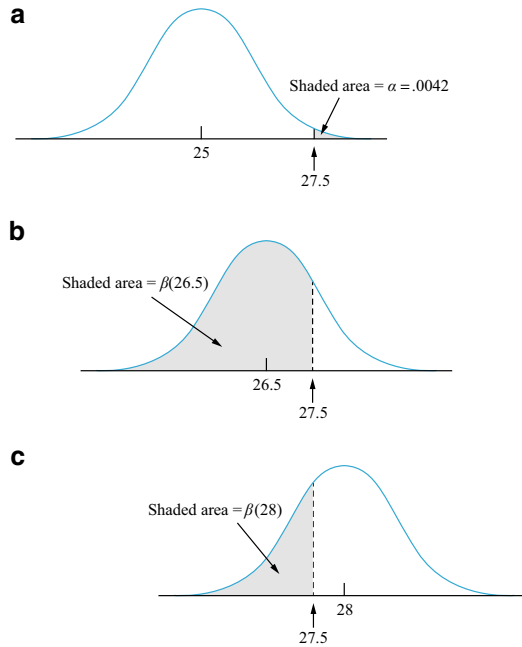
Figure 9.1 α and β illustrated for Example 9.6: (**a**) the distribution of $\bar{X}$ when $\mu = 25$ (H_0 true); (**b**) the distribution of $\bar{X}$ when $\mu = 26.5$ (H_0 false); (**c**) the distribution of $\bar{X}$ when $\mu = 28$ (H_0 false)

As in Example 9.5, if the more realistic null hypothesis $\mu \leq 25$ is considered, there is an α for each parameter value for which H_0 is true: $\alpha(25)$, $\alpha(24.2)$, $\alpha(23.6)$, and so on. It is easily verified, though, that $\alpha(25)$ is the largest of all these type I error probabilities. Focusing on the boundary value amounts to working explicitly with the "worst case." ∎

Selecting the Rejection Region

The specification of a cutoff value for the rejection region in the examples just considered was fairly arbitrary. Use of the rejection region $R_3 = \{0, 1, 2, 3\}$ in Example 9.5 resulted in $\alpha = .234$, $\beta(.10) = .236$, and $\beta(.15) = .529$. Many would think these error probabilities intolerably large. Perhaps they can be decreased by changing the cutoff value.

Example 9.7 (Example 9.5 continued) Let us use the same survey plan and test statistic X as previously described in the screwtop wine problem but now consider the rejection region $R_2 = \{0, 1, 2\}$. Since X still has a binomial distribution with parameters $n = 25$ and p,

$$\alpha = P(H_0 \text{ is rejected when } p = .2)$$
$$= P(X \leq 2 \text{ when } X \sim \text{Bin}(25, .2)) = B(2; \ 25, \ .2) = .098$$

The type I error probability has been decreased by using the new rejection region. However, a price has been paid for this decrease:

$$\beta(.1) = P(H_0 \text{ is not rejected when } p = .1)$$
$$= P(X > 2 \text{ when } X \sim \text{Bin}(25, .1)) = 1 - B(2; 25, .1) = .463$$
$$\beta(.15) = 1 - B(2; 25, .15) = .746$$

Both these β's are larger than the corresponding error probabilities .236 and .529 for the region R_3. In retrospect, this is not surprising: α is computed by summing over probabilities of test statistic values *in the rejection region*, whereas β is the probability that X falls in the *complement* of the rejection region. Making the rejection region smaller must therefore decrease α while increasing β for any fixed alternative value of the parameter. ∎

A similar trade-off between α and β would occur if we changed the rejection region cutoff in Example 9.6. Looking at Figure 9.1, it's clear that if we shifted the cutoff $c = 27.5$ to the left (e.g., to $c = 27$), the two β's illustrated would decrease (less cumulative area under the normal curves) but α would increase (greater upper-tail area than before). The results of these examples can be generalized in the following manner.

PROPOSITION Suppose a study and a sample size are fixed and a test statistic is chosen. Then decreasing the size of the rejection region to obtain a smaller value of α results in a larger value of β for any particular parameter value consistent with H_a, and vice versa.

This proposition says that once the test statistic and n are fixed, there is no rejection region that will simultaneously make both α and all β's small. A region must be chosen to effect a compromise between α and β. The approach adhered to by most statistical practitioners is to specify the largest value of α that can be tolerated and find a rejection region having that value of α. This makes β as small as possible subject to the bound on α. The resulting value of α is often referred to as the **significance level** of the test. Traditional levels of significance are .10, .05, and .01, although the level in any particular problem will depend on the seriousness of a type I error—the more serious this error, the smaller should be the significance level. The corresponding test procedure is called a **level α test** (e.g., a level .05 test or a level .01 test). A test with significance level α is one for which the type I error probability is controlled at the specified level.

Example 9.8 (Example 9.6 continued) For the truck engine scenario, suppose a hypothesis test with significance level $\alpha = .05$ is desired. The rejection region will still have the form $\bar{x} \geq c$, but the value of c is determined by α:

$$.05 = P(\text{type I error}) = P(\bar{X} \geq c \text{ when } H_0 \text{ is true})$$
$$= P(\bar{X} \geq c \text{ when } \bar{X} \sim N(25, .95))$$
$$= 1 - \Phi\left(\frac{c - 25}{.95}\right) \Rightarrow \Phi\left(\frac{c - 25}{.95}\right) = .95$$

The last expression above implies that $(c - 25)/.95$ is the 95th percentile of the standard normal distribution, $z_{.05}$. Either from Section 4.3 or directly from Appendix Table A.3, $z_{.05} = 1.645$, from which the desired rejection region cutoff is $c = 25 + (1.645)(.95) \approx 26.56$ mpg.

So, a level .05 test of H_0: $\mu = 25$ versus H_a: $\mu > 25$ in this scenario involves rejecting H_0 if and only if $\bar{x} \geq 26.56$. Then β is the probability that $\bar{X} < 26.56$ and can be calculated for any $\mu > 25$. ∎

Power

Many statistical software packages will calculate type II error probabilities for a variety of test procedures, including those presented in this and subsequent chapters. This is typically expressed in terms of **power**, defined as the probability that the test procedure will reject H_0. For parameter values consistent with H_a, this is simply $1 - \beta$. As the name is meant to imply, greater power is better: lower values of β (i.e., less chance of a type II error) correspond to higher power values.

Like β, there is not a single value for the power of a test procedure, but rather a different value for each possible value of the parameter. As a result, though power can be calculated for a single parameter value, it is more common to see a **power curve**, where the horizontal axis represents possible values of the parameter and the vertical axis displays power.

Example 9.9 (Example 9.8 continued) Figure 9.2 shows the power curve for the test procedure that rejects H_0 when $\bar{x} \geq 26.56$. For each value of μ consistent with H_a: $\mu > 25$, the power of the test procedure is simply $P(\bar{X} \geq 26.56)$. Note that the power of the test at $\mu = 25$ is $\alpha = .05$ by the definition of power. The power increases as the value of the parameter moves further from the null value—a large departure from H_0 is more likely to be detected than a small departure.
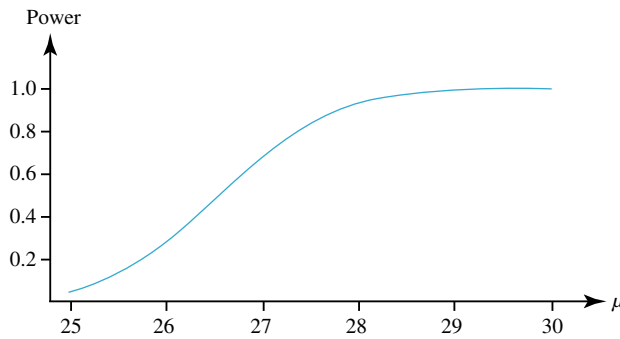
Figure 9.2 Power curve for the test procedure of Examples 9.8–9.9 ■

One final, but important, note: the probabilities α, β, and power are all functions of the selected *test procedure*, not of any sample data. They reflect the chance that certain outcomes of the test procedure *will happen* in the future, when a random sample of the specified size n is selected.

Exercises: Section 9.1 (1–14)

1. For each of the following assertions, state whether it is a legitimate statistical hypothesis and why:

   a. $H: \sigma > 100$
   b. $H: \tilde{x} = 45$
   c. $H: s \leq .20$
   d. $H: \sigma_1/\sigma_2 < 1$
   e. $H: \bar{X} - \bar{Y} = 5$
   f. $H: \lambda \leq .01$, where λ is the parameter of an exponential distribution used to model component lifetime

2. For the following pairs of assertions, indicate which do not comply with our rules for setting up hypotheses and why (the subscripts 1 and 2 differentiate between quantities for two different populations or samples):

   a. H_0: $\mu = 100$, H_a: $\mu > 100$
   b. H_0: $\sigma = 20$, H_a: $\sigma \leq 20$
   c. H_0: $p \neq .25$, H_a: $p = .25$
   d. H_0: $\mu_1 - \mu_2 = 25$, H_a: $\mu_1 - \mu_2 > 100$
   e. H_0: $S_1^2 = S_2^2$, H_a: $S_1^2 \neq S_2^2$
   f. H_0: $\mu = 120$, H_a: $\mu = 150$

g. $H_0: \sigma_1/\sigma_2 = 1$, $H_a: \sigma_1/\sigma_2 \neq 1$

h. $H_0: p_1 - p_2 = -.1$, $H_a: p_1 - p_2 < -.1$

3. To determine whether the girder welds in a new performing arts center meet specifications, a random sample of welds is selected, and tests are conducted on each weld in the sample. Weld strength is measured as the force required to break the weld. Suppose the specifications state that mean strength of welds should exceed 100 lb/in^2; the inspection team decides to test $H_0: \mu = 100$ versus $H_a: \mu > 100$. Explain why it might be preferable to use this H_a rather than $\mu < 100$.

4. Let μ denote the true average radioactivity level (picocuries per liter). The value 5 pCi/L is considered the dividing line between safe and unsafe water. Would you recommend testing $H_0: \mu = 5$ versus $H_a: \mu > 5$ or $H_0: \mu = 5$ versus $H_a: \mu < 5$? Explain your reasoning. [*Hint*: Think about the consequences of a type I and type II error for each possibility.]

5. Before agreeing to purchase a large order of polyethylene sheaths for a particular type of high-pressure oil-filled submarine power cable, a company wants to see conclusive evidence that the true standard deviation of sheath thickness is < .05 mm. What hypotheses should be tested, and why? In this context, what are the type I and type II errors?

6. Many older homes have electrical systems that use fuses rather than circuit breakers. A manufacturer of 40-amp fuses wants to make sure that the mean amperage at which its fuses burn out is in fact 40. If the mean amperage is lower than 40, customers will complain because the fuses require replacement too often. If the mean amperage is higher than 40, the manufacturer might be liable for damage to an electrical system due to fuse malfunction. To verify the amperage of the fuses, a sample of fuses is to be selected and inspected. If a hypothesis test were to be performed on the resulting data, what null and alternative hypotheses would be of interest to the manufacturer? Describe type I and type II errors in the context of this problem situation.

7. Water samples are taken from water used for cooling as it is being discharged from a power plant into a river. It has been determined that as long as the mean temperature of the discharged water is at most 150 °F, there will be no negative effects on the river's ecosystem. To investigate whether the plant is in compliance with regulations that prohibit a mean discharge-water temperature above 150°, 50 water samples will be taken at randomly selected times, and the temperature of each sample recorded. The resulting data will be used to test the hypotheses $H_0: \mu = 150°$ versus $H_a: \mu > 150°$. In the context of this situation, describe type I and type II errors. Which type of error would you consider more serious? Explain.

8. A regular type of laminate is currently being used by a manufacturer of circuit boards. A special laminate has been developed to reduce warpage. The regular laminate will be used on one sample of specimens and the special laminate on another sample, and the amount of warpage will then be determined for each specimen. The manufacturer will then switch to the special laminate only if it can be demonstrated that the true average amount of warpage for that laminate is less than for the regular laminate. State the relevant hypotheses, and describe the type I and type II errors in the context of this situation.

9. Two different companies have applied to provide internet service in a region. Let p denote the proportion of all potential subscribers who favor the first company over the second. Consider testing $H_0: p = .5$ versus $H_a: p \neq .5$ based on a random sample of 25 individuals. Let X denote the number in the sample who favor the first company and x represent the observed value of X.

a. Which of the following rejection regions is most appropriate and why?

$R_1 = \{x : x \leq 7 \text{ or } x \geq 18\}$,
$R_2 = \{x : x \leq 8\}, R_3 = \{x : x \geq 17\}$

b. Using the selected rejection region, what would you conclude if 6 of the 25 queried favored company 1?

c. In the context of this problem situation, describe what type I and type II errors are.

d. What is the probability distribution of the test statistic X when H_0 is true? Use it to compute the probability of a type I error.

e. For the region selected in part (a), compute the probability of a type II error and the power when $p = .3, .4, .6$, and $.7$.

10. For healthy individuals the level of prothrombin in the blood is approximately normally distributed with mean 20 mg/dL and standard deviation 4 mg/dL. Low levels indicate low clotting ability. In studying the effect of gallstones on prothrombin, the level of each patient in a sample is measured to see if there is a deficiency. Let μ be the true average level of prothrombin for gallstone patients (and assume $\sigma = 4$).

a. What are the appropriate null and alternative hypotheses?

b. Let $\bar{X}$ denote the sample average level of prothrombin in a sample of $n = 20$ randomly selected gallstone patients. Consider the test procedure with test statistic $\bar{X}$ and rejection region $\bar{x} \leq 17.92$. What is the probability distribution of the test statistic when H_0 is true? What is the probability of a type I error for the test procedure?

c. What is the probability distribution of the test statistic when $\mu = 16.7$? Using the test procedure of part (b), what is the probability that gallstone patients will be judged not deficient in prothrombin, when in fact $\mu = 16.7$ (a type II error)?

d. How would you change the test procedure of part (b) to obtain a test with significance level .05? What impact would this change have on the error probability of part (c)?

e. Consider the standardized test statistic $Z = (\bar{X} - 20)/(\sigma/\sqrt{n}) = (\bar{X} - 20)/.8944$. What are the values of Z corresponding to the rejection region of part (b)?

11. The calibration of a scale is to be checked by weighing a 10-kg test specimen 25 times. Suppose that the results of different weighings are independent of one another and that the weight on each trial is normally distributed with $\sigma = .200$ kg. Let μ denote the true average weight reading on the scale.

a. What hypotheses should be tested?

b. Suppose the scale is to be recalibrated if either $\bar{x} \geq 10.1032$ or $\bar{x} \leq 9.8968$. What is the probability that recalibration is carried out when it is actually unnecessary?

c. What is the probability that recalibration is judged unnecessary when in fact $\mu = 10.1$? When $\mu = 9.8$?

d. Let $z = (\bar{x} - 10)/(\sigma/\sqrt{n})$. For what value c is the rejection region of part (b) equivalent to the "two-tailed" region either $z \geq c$ or $z \leq -c$?

e. If the sample size were only 10 rather than 25, how should the procedure of part (d) be altered so that $\alpha = .05$?

f. Using the test of part (e), what would you conclude from the following sample data?

| 9.981 | 10.006 | 9.857 | 10.107 | 9.888 |
| 9.728 | 10.439 | 10.214 | 10.190 | 9.793 |

12. A new design for the braking system on a certain type of car has been proposed. For the current system, the true average braking distance at 40 mph under specified conditions is known to be 120 ft. It is proposed that the new design be implemented only if sample data strongly indicates a reduction in true average braking distance for the new design.

a. Define the parameter of interest and state the relevant hypotheses.

b. Suppose braking distance for the new system is normally distributed with $\sigma = 10$. Let $\bar{X}$ denote the sample average braking distance for a random sample of 36 observations. Which of the following

rejection regions is appropriate: $R_1 = \{\bar{x} : \bar{x} \geq 124.80\}$, $R_2 = \{\bar{x} : \bar{x} \leq 115.20\}$, $R_3 = \{\bar{x} : \text{ either } \bar{x} \geq 125.13 \text{ or } \bar{x} \leq 114.87\}$?

c. What is the significance level for the appropriate region of part (b)? How would you change the region to obtain a test with $\alpha = .001$?

d. What is the probability that the new design is not implemented when its true average braking distance is actually 115 ft and the appropriate region from part (b) is used?

e. Let $Z = (\bar{X} - 120)/(\sigma/\sqrt{n})$. What is the significance level for the rejection region $\{z: z \leq -2.33\}$? For the region $\{z: z \leq -2.88\}$?

13. Let $X_1, \ldots, X_n$ denote a random sample from a normal population distribution with a known value of σ.

a. For testing the hypotheses $H_0: \mu = \mu_0$ versus $H_a: \mu > \mu_0$ (where μ_0 is a fixed number), show that the test with test statistic $\bar{X}$ and rejection region $\bar{x} \geq \mu_0 + 2.33\sigma/\sqrt{n}$ has significance level .01.

b. Let $d = \mu - \mu_0$, the difference between the true and hypothesized values of the population mean. Graph the power function of the test procedure in part (a) as a function of d.

c. Suppose the procedure of part (a) is used to test $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$. If $\mu_0 = 100$, $n = 25$, and $\sigma = 5$, what is the probability of committing a type I error when $\mu = 99$? When $\mu = 98$? In general, what can be said about the probability of a type I error when the actual value of μ is less than μ_0? Verify your assertion.

14. Reconsider the situation of Exercise 11 and suppose the rejection region is $\{\bar{x} : \bar{x} \geq 10.1004 \text{ or } \bar{x} \leq 9.8940\} = \{z : z \geq 2.51 \text{ or } z \leq -2.65\}$.

a. What is α for this procedure?

b. What is β when $\mu = 10.1$? When $\mu = 9.9$? Is this desirable?

c. Graph the power function for this test procedure as a function of the unknown μ.

9.2   Tests About a Population Mean

In Sections 8.1–8.2, confidence intervals for a population mean μ were developed in two stages: first, for the (unrealistic) scenario when the population standard deviation σ is known, then for cases when both μ and σ are unknown. We now develop test procedures for these same two cases. Later in this section, we provide some practical advice on the implementation of hypothesis tests for μ.

Tests About μ for Normal Data with Known σ
Throughout this subsection, we assume that

1. The population distribution is normal.
2. The value of the population standard deviation σ is known.

Although the assumption that the value of σ is known is rarely met in practice, this case provides a good starting point because of the ease with which general procedures and their properties can be developed. Let $X_1, \ldots, X_n$ represent a random sample of size n from the normal population. Then the sample mean $\bar{X}$ has a normal distribution with expected value $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. The null hypothesis is $H_0: \mu = \mu_0$, so μ_0 is the *null value* of the parameter. When H_0 is true, $\mu_{\bar{X}} = \mu_0$. Consider now the statistic Z obtained by standardizing $\bar{X}$ under the assumption that H_0 is true:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \tag{9.1}$$

Substitution of the computed sample mean $\bar{x}$ into (9.1) gives z, the distance between $\bar{x}$ and μ_0 expressed in "standard deviation units." For example, if the null hypothesis is $H_0: \mu = 100$, $\sigma_{\bar{X}} = 2$ and $\bar{x} = 103$, then the test statistic value is given by $z = (103 - 100)/2 = 1.5$. That is, the observed value of $\bar{x}$ is 1.5 standard deviations (of $\bar{X}$) above what we expect it to be when H_0 is true. The statistic Z is a natural measure of the distance between $\bar{X}$, the estimator of μ, and its expected value when H_0 is true. If this distance is too great in a direction consistent with H_a, the null hypothesis should be rejected.

Suppose first that the alternative hypothesis has the form $H_a: \mu > \mu_0$. Then an $\bar{x}$ value less than μ_0 certainly does not provide support for H_a. Such an $\bar{x}$ corresponds to a negative value of z, since $\bar{x} - \mu_0$ is negative and the divisor $\sigma/\sqrt{n}$ is positive. Similarly, an $\bar{x}$ value that exceeds μ_0 by only a small amount (corresponding to z which is positive but small) does not suggest that H_0 should be rejected in favor of H_a. The rejection of H_0 is appropriate only when $\bar{x}$ considerably exceeds μ_0—that is, when the z value is positive and large. In summary, the appropriate rejection region has the form $z \geq c$ for some relatively large positive constant c.

As discussed in Section 9.1, the cutoff value c should be chosen to control the probability of a type I error at the desired level α. This is easily accomplished because the distribution of the test statistic Z when H_0 is true is the standard normal distribution (that's why μ_0 was subtracted in standardizing):

$$\alpha = P(\text{type I error}) = P(H_0 \text{ is rejected when } H_0 \text{ is true})$$
$$= P(Z \geq c \text{ when } Z \sim N(0, 1)) = 1 - \Phi(c) \Rightarrow$$
$$\Phi(c) = 1 - \alpha \Rightarrow c = \Phi^{-1}(1 - \alpha) = z_\alpha$$

That is, the rejection region $z \geq z_\alpha$ has type I error probability α. For instance, if a level .01 test is desired, then H_0 should be rejected if $z \geq c = z_{.01} = 2.33$. This test procedure is *upper-tailed* because the rejection region consists only of large values of the test statistic.

Analogous reasoning for the alternative hypothesis $H_a: \mu < \mu_0$ suggests a rejection region of the form $z \leq c$, where c is a suitably chosen negative number ($\bar{x}$ is far below μ_0 if and only if z is quite negative). Because Z has a standard normal distribution when H_0 is true, taking $c = -z_\alpha$ results in $P(\text{type I error}) = \alpha$. This is a *lower-tailed* test. For example, $z_{.10} = 1.28$ implies that the rejection region $z \leq -1.28$ specifies a test with significance level .10.

Finally, when the alternative hypothesis is $H_a: \mu \neq \mu_0$, H_0 should be rejected if $\bar{x}$ is too far to either side of μ_0. This is equivalent to rejecting H_0 if either $z \geq c$ or $z \leq -c$. Suppose we desire $\alpha = .05$. Then,

$$.05 = P(Z \geq c \text{ or } Z \leq -c \text{ when } Z \sim N(0, 1))$$
$$= \Phi(-c) + 1 - \Phi(c) = 2[1 - \Phi(c)]$$

Thus c is such that $1 - \Phi(c)$, the area under the standard normal curve to the right of c, is .025 (and not .05!). From Section 4.3 or Appendix Table A.3, $c = 1.96$, and the rejection region is $\{z \geq 1.96$ or $z \leq -1.96\}$. For any α, the *two-tailed* rejection region $z \geq z_{\alpha/2}$ *or* $z \leq -z_{\alpha/2}$ has type I error probability α (since area $\alpha/2$ is captured under each of the two tails of the z curve). Again, the key reason for using the standardized test statistic Z is that because Z has a known distribution when H_0 is true (standard normal), a rejection region with desired type I error probability is easily obtained by using an appropriate critical value.

The foregoing test procedures are summarized in the accompanying box, and the corresponding rejection regions are illustrated in Figure 9.3.

z curve (probability distribution of test statistic Z when H_0 is true)
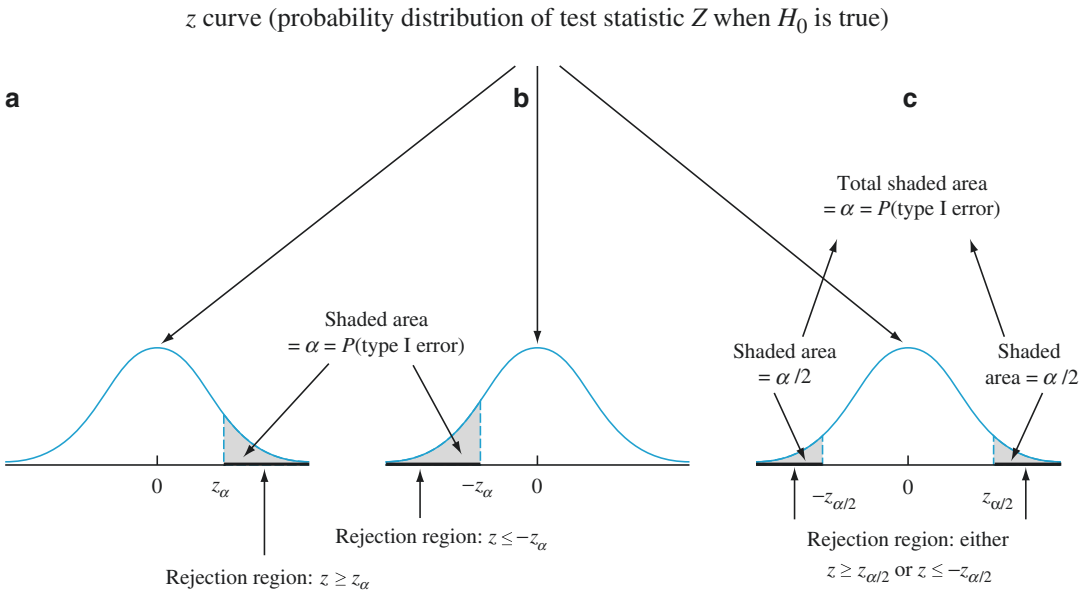
Figure 9.3 Rejection regions for z tests: (**a**) upper-tailed test; (**b**) lower-tailed test; (**c**) two-tailed test

| **THE ONE-SAMPLE** **z TEST** | Null hypothesis: H_0: $\mu = \mu_0$ |
| | Test statistic value: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ |

Alternative Hypothesis	**Rejection Region for Level α Test**
H_a: $\mu > \mu_0$	$z \geq z_\alpha$ (upper-tailed test)
H_a: $\mu < \mu_0$	$z \leq -z_\alpha$ (lower-tailed test)
H_a: $\mu \neq \mu_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)

Use of the following sequence of steps is recommended when testing hypotheses about a parameter; these steps will be repeated frequently throughout the remainder of the book. The formulation of hypotheses (steps 1 and 2) should be done *before* examining the data.

1. Identify the parameter of interest and describe it in the context of the problem situation.
2. Determine the null value, and state the appropriate null and alternative hypotheses.
3. Check the plausibility of any assumptions or requirements for the test procedure under consideration to be valid.
4. Give the formula for the computed value of the test statistic (substituting the null value and the known values of any other parameters, but *not* those of any sample-based quantities).
5. State the rejection region for the selected significance level α.
6. Compute any necessary sample quantities, substitute into the formula for the test statistic value, and compute that value.
7. Decide whether H_0 should be rejected and state this conclusion in the problem context.

Example 9.10 If the activation temperature of an automated sprinkler system used for fire protection in an office building is too high, a fire could do substantial damage before water is dispersed. On the other hand, activation at too low a temperature could cause water damage when there is little fire threat. A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°. A sample of $n = 9$ systems, when tested, yields a sample average activation temperature of 131.08 °F. If the distribution of activation times is normal with standard deviation 1.5 °F, does the data contradict the manufacturer's claim at significance level $\alpha = .01$?

1. Parameter of interest: μ = true average activation temperature
2. Hypotheses: H_0: $\mu = 130$ (null value = $\mu_0 = 130$)
   H_a: $\mu \neq 130$ (a departure from the claimed value in *either* direction is of concern)
3. Assumptions/requirements: We have assumed an underlying normal population distribution of activation temperatures with a known population standard deviation.
4. Test statistic value:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 130}{1.5/\sqrt{n}}$$

5. Rejection region: The form of H_a implies use of a two-tailed test with rejection region *either* $z \geq z_{.005}$ *or* $z \leq -z_{.005}$. From Section 4.3 or Appendix Table A.3, $z_{.005} = 2.576$, so we reject H_0 if either $z \geq 2.576$ or $z \leq -2.576$.
6. Substituting $n = 9$ and $\bar{x} = 131.08$,

$$z = \frac{131.08 - 130}{1.5/\sqrt{9}} = \frac{1.08}{.5} = 2.16$$

   That is, the observed sample mean is a bit more than 2 standard deviations above what would have been expected were H_0 true.
7. The computed value $z = 2.16$ does *not* fall in the rejection region, so H_0 cannot be rejected at significance level .01. The data does not give sufficient evidence to conclude that the true average differs from the design value of 130. ∎

Power, β, and Sample Size Determination for the One-Sample z Test

The one-sample z tests are among the few in statistics for which there are simple formulas available for β, the probability of a type II error. Consider first the upper-tailed test with rejection region $z \geq z_\alpha$. This is equivalent to $\bar{x} \geq \mu_0 + z_\alpha \cdot \sigma/\sqrt{n}$, so H_0 will not be rejected if $\bar{x} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n}$. Now let μ' denote a particular value of μ that exceeds the null value μ_0. Then,

$$\beta(\mu') = P(H_0 \text{ is not rejected when } \mu = \mu')$$
$$= P(\bar{X} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n} \text{ when } \mu = \mu')$$
$$= P\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} < z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ when } \mu = \mu'\right)$$
$$= \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$$

The power of the upper-tailed one-sample z test is then $1 - \beta(\mu')$. As μ' increases, $\mu_0 - \mu'$ becomes more negative, so $\beta(\mu')$ will be small and power will be large when μ' greatly exceeds μ_0 (because the value at which Φ is evaluated will then be quite negative). Power and β for the lower-tailed and two-tailed tests are derived in an analogous manner.

In addition to specifying the α level, investigators may prescribe a desired power level at an alternative value μ' that is of particular concern. In the sprinkler example, company officials might view $\mu' = 132$ as a very substantial departure from H_0: $\mu = 130$ and therefore wish to have a 90% chance of rejecting H_0 (power = .90) at that temperature—that is, $\beta(132) = .10$—in addition to, say, $\alpha = .01$. More generally, consider the two restrictions $P(\text{type I error}) = \alpha$ and $\beta(\mu') = \beta$ for specified α, μ', and β. Then for an upper-tailed test, the sample size n should be chosen to satisfy

$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \beta$$

This implies that

$$z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} = \Phi^{-1}(\beta) = -z_\beta$$

It is easy to solve this equation for the desired n. A parallel argument yields the necessary sample size for lower- and two-tailed tests as summarized in the next box.

Alternative Hypothesis   Type II Error Probability $\beta(\mu')$ for a Level α Test

H_a: $\mu > \mu_0$
$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$$

H_a: $\mu < \mu_0$
$$1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$$

H_a: $\mu \neq \mu_0$
$$\Phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$$

where $\Phi(z)$ = the standard normal cdf. For each case, power = $1 - \beta(\mu')$.
The sample size n for which a level α test also has $\beta(\mu') = \beta$ at the alternative value μ' is

$$n = \begin{cases} \left[\dfrac{\sigma\left(z_\alpha + z_\beta\right)}{\mu_0 - \mu'}\right]^2 & \begin{array}{l}\text{for a one-tailed} \\ \text{(upper or lower) test}\end{array} \\[4mm] \left[\dfrac{\sigma\left(z_{\alpha/2} + z_\beta\right)}{\mu_0 - \mu'}\right]^2 & \begin{array}{l}\text{for a two-tailed test} \\ \text{(an approximate solution)}\end{array} \end{cases}$$

Example 9.11 Let μ denote the true average tread life of a type of tire. Consider testing the hypotheses H_0: $\mu = 30{,}000$ versus H_a: $\mu > 30{,}000$ based on a sample of size $n = 16$ from a normal population distribution with $\sigma = 1500$. A test with $\alpha = .01$ requires $z_\alpha = z_{.01} = 2.33$. The probability of making a type II error when $\mu = 31{,}000$ is

$$\beta(31{,}000) = \Phi\left(2.33 + \frac{30{,}000 - 31{,}000}{1500/\sqrt{16}}\right) = \Phi(-.34) = .3669$$

The probability of rejecting H_0 when $\mu = 31{,}000$, i.e., the power, is $1 - .3669 = .6331$.

Since $z_{.1} = 1.28$, the requirement that the level .01 test also have $\beta(31{,}000) = .1$ necessitates

$$n = \left[\frac{1500(2.33 + 1.28)}{30{,}000 - 31{,}000}\right]^2 = (-5.42)^2 = 29.32$$

The sample size must be an integer, so $n = 30$ tires should be used. ∎

The One-Sample t Test

We now modify the one-sample z test to accommodate the more realistic situation when σ is unknown, following a path similar to what was outlined in Section 8.2. Consider the test statistic obtained by replacing σ in (9.1) by the sample standard deviation, S:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \tag{9.2}$$

Assuming $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution, the rv (9.2) follows a t_{n-1} distribution when $H_0: \mu = \mu_0$ is true. Knowledge of the test statistic's distribution when H_0 is true (the "null distribution") allows us to construct a rejection region for which the type I error probability is controlled at the desired level. For instance, consider testing $H_0: \mu = \mu_0$ against $H_a: \mu > \mu_0$ using (9.2). Use of the upper-tail t critical value $t_{\alpha,n-1}$ to specify the rejection region $t \geq t_{\alpha,n-1}$ implies that

$$\begin{aligned}
P(\text{type I error}) &= P(H_0 \text{ is rejected when it is true}) \\
&= P(T \geq t_{\alpha,n-1} \text{ when } T \text{ has a } t \text{ distribution with } n - 1 \text{ df}) \\
&= \alpha
\end{aligned}$$

The rejection region for the t test differs from that of the z test only in that a t critical value $t_{\alpha,n-1}$ replaces the z critical value z_α. Similar comments apply to alternative hypotheses for which a lower-tailed or two-tailed test is appropriate.

THE ONE-SAMPLE t TEST	Null hypothesis: $H_0: \mu = \mu_0$ Test statistic value: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Alternative Hypothesis	Rejection Region for a Level α Test
$H_a: \mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (upper-tailed)
$H_a: \mu < \mu_0$	$t \leq -t_{\alpha,n-1}$ (lower-tailed)
$H_a: \mu \neq \mu_0$	either $t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$ (two-tailed)

Graphs of these rejection regions are essentially the same as those in Figure 9.3; simply replace the z curves and z critical values with appropriate t curves and t critical values.

Example 9.12 Particulate matter from roads contributes to pollution when those particles are washed into nearby waterways by rain. The size of the particles can impact the effectiveness of various stormwater control measures. The authors of the article "Characterizing Runoff from Roads: Particle Size Distributions, Nutrients, and Gross Solids" (*J. Environ. Engr.* 2016) took roadside measurements at several sites in North Carolina. For each assay they recorded d_{50}, the median size of particles in the assay (a standard measure of particle size in such studies). Here are the d_{50} values (microns) for $n = 9$ assays performed off I-40 near Black Mountain:

$$82.9 \quad 56.8 \quad 66.5 \quad 49.4 \quad 105.4 \quad 79.5 \quad 82.5 \quad 50.7 \quad 43.0$$

Previous studies indicated that the typical d_{50} value alongside roads of this type is 44 microns. Does the sample data provide convincing statistical evidence that the true mean d_{50} value differs from 44 microns? Let's carry out a test using a significance level of $\alpha = .01$.

1. $\mu =$ true average d_{50} value (microns) for particulate matter assays at the Black Mountain site
2. H_0: $\mu = 44$
   H_a: $\mu \neq 44$
3. A normal probability plot (not shown) indicates that the population distribution could plausibly be normal, so the one-sample t test will be used.
4. $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{\bar{x} - 44}{s/\sqrt{n}}$
5. From Appendix Table A.6, $t_{\alpha/2, n-1} = t_{.005,8} = 3.355$. So we reject H_0 if either $t \geq 3.355$ or $t \leq -3.355$.
6. From the data provided, $\bar{x} = 68.52$ and $s = 20.49$. Substituting,

$$t = \frac{68.52 - 44}{20.49/\sqrt{9}} = \frac{24.52}{6.83} = 3.59$$

7. Because the computed value $t = 3.59$ falls in the rejection region ($3.59 \geq 3.355$), H_0 is rejected at the .01 level. Even this small sample of data provides convincing statistical evidence that the true mean d_{50} value for roadside particulates at the Black Mountain site differs from the "typical" value of 44 microns seen in other studies. ∎

Some Practical Advice

The validity of the one-sample t test rests on Gosset's Theorem, which assumes a normally distributed population. The plausibility of this assumption can be checked with a normal probability plot. But as we noted in Chapter 8, the t distributions are "robust" against violations of normality when the sample size n is reasonably large. That is, when using data from a large sample (say, $n > 40$), the results of applying the one-sample t test procedure should be reasonably accurate even if the underlying population distribution is not normal.

We have also seen that, for n large, the z and t_{n-1} distributions are quite similar, so that using a z distribution to determine rejection region cutoffs gives very similar results to the one-sample t test procedure. In current practice, researchers typically use the one-sample t test even for large samples, except in the extremely rare case where σ is known.

The one situation in which inferences for μ cannot be based on a t procedure is when the sample size is small *and* the data strongly suggests a nonnormal population. Methods to address this situation are considered at the end of this chapter and in Chapter 14.

Example 9.13 A sample of bills for meals was obtained at a restaurant (by Erich Brandt). For each of 70 bills the tip was found as a percentage of the raw bill (before taxes). Does it appear that the population mean tip percentage for this restaurant exceeds the standard 15%? Here are the 70 tip percentages:

14.21	20.24	20.10	14.94	15.69	15.04	12.04	20.16	17.85	16.35
19.12	20.37	15.29	18.39	27.55	16.01	10.94	13.52	17.42	14.48
29.87	17.92	19.74	22.73	14.56	15.16	16.09	16.42	19.07	13.74
13.46	16.79	19.03	19.19	19.23	12.39	16.89	18.93	13.56	17.70
11.48	13.96	21.58	11.94	19.02	17.73	20.07	40.09	19.88	22.79
15.23	16.09	19.19	11.91	18.21	15.37	16.31	16.03	48.77	12.31
21.53	12.76	18.07	14.11	15.86	20.67	15.66	18.54	27.88	13.81

Figure 9.4 shows a descriptive summary obtained from Minitab. The *sample* mean tip percentage is 17.986, which obviously is greater than 15.
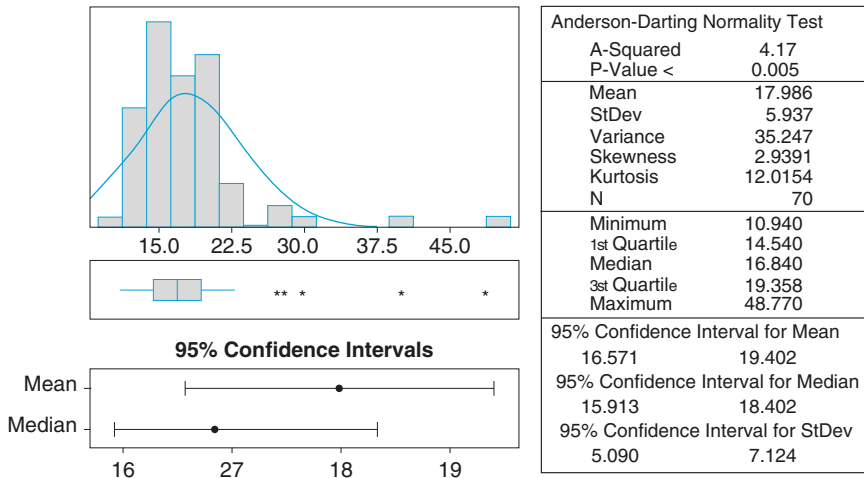
Figure 9.4 Minitab descriptive summary for the tip data of Example 9.13

1. μ = true average tip percentage
2. H_0: $\mu = 15$
   H_a: $\mu > 15$
3. The distribution is positively skewed because there are some very large tips (and a normal probability plot therefore would not exhibit a linear pattern). But the large sample size ($n = 70 > 40$) means that the one-sample t test does not require a normal population distribution.
4. $t = \dfrac{\bar{x} - 15}{s/\sqrt{n}}$
5. Using a test with a significance level .05, H_0 will be rejected if $t \geq t_{.05,70-1} \approx 1.667$ (an upper-tailed test).
6. With $n = 70$, $\bar{x} = 17.986$, and $s = 5.937$,

$$t = \frac{17.986 - 15}{5.937/\sqrt{70}} = \frac{2.986}{.7096} = 4.21$$

7. Since $4.21 > 1.667$, H_0 is rejected. There is convincing statistical evidence that the population mean tip percentage exceeds 15%. ∎

Power, β, and Sample Size Determination for the One-Sample t Test

When the sample size is large (as in Example 9.13), power and sample size calculations for the one-sample t test can be approximated by the formulas provided earlier in this section. Notice that a plausible value of σ must be specified; the sample standard deviation s may be used for this purpose, although power and sample size calculations are often performed prior to collecting any data.

Alternatively, μ' values of interest are sometimes expressed as a certain number of standard deviations from the null value. For example, researchers may be interested in a one-quarter sd increase from the null value, in which case $\mu' = \mu_0 + 0.25\sigma$. Re-expressing μ' in the form $\mu_0 + d\sigma$, where the value d can be positive or negative, simplifies the expressions for β and n presented earlier in this section so that they no longer depend explicitly on the unknown σ. For instance, the formula for β in an upper-tailed one-sample z test under this substitution becomes

$$\beta(\mu') = \beta(\mu_0 + d\sigma) = \Phi\left(z_\alpha + \frac{\mu_0 - (\mu_0 + d\sigma)}{\sigma/\sqrt{n}}\right) = \Phi\left(z_\alpha - d\sqrt{n}\right)$$

The other formulas simplify in a similar fashion.

Exact calculations of power and $\beta(\mu')$ for the one-sample t test (i.e., not using the normal approximations) are much less straightforward. This is because the test statistic $T = (\bar{X} - \mu_0)/(S/\sqrt{n})$ in (9.2) does *not* have a t distribution when H_0 is false. Rather, when the true value of μ is anything other than μ_0, T has a much more complicated distribution, related to the following definition.

DEFINITION   Let $Z \sim N(0, 1)$ and $Y \sim \chi_v^2$ be independent random variables. For any real number δ, the random variable

$$\frac{Z + \delta}{\sqrt{Y/v}} \tag{9.3}$$

has a **noncentral t distribution** with v degrees of freedom and **noncentrality parameter** δ. Note that when $\delta = 0$, the rv (9.3) matches the definition of the t distribution from Section 6.3 and thus has a t_v distribution.

There is no closed-form expression for the noncentral t pdf when $\delta \neq 0$, and so software is essential for calculations based on it. It can be shown (Exercise 38) that when $\mu = \mu'$, the one-sample t test statistic (9.2) has a noncentral t distribution with $n - 1$ df and noncentrality parameter

$$\delta = \frac{\mu' - \mu_0}{\sigma/\sqrt{n}} \tag{9.4}$$

Now consider determining the power of a lower-tailed one-sample t test; the upper-tailed and two-tailed calculations proceed similarly. Let $F(x; v, \delta)$ denote the cdf of the noncentral t distribution. Then

$$\begin{aligned}
\text{power} &= P(T \leq -t_{\alpha,n-1} \text{ when } \mu = \mu' \text{ rather than } \mu_0) \\
&= P\left(T \leq -t_{\alpha,n-1} \text{ when } T \sim \text{noncentral } t, \text{ df} = n - 1, \delta = \frac{\mu' - \mu_0}{\sigma/\sqrt{n}}\right) \\
&= F\left(-t_{\alpha,n-1}; n - 1, \frac{\mu' - \mu_0}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

Many statistical software packages can calculate this quantity once all the inputs are specified. As in the previous discussion, substitutions of the form $\mu' = \mu_0 + d\sigma$ simplify power and β expressions and do not require knowledge (or even an estimate) of σ.

Example 9.14   The true average voltage drop from collector to emitter of insulated gate bipolar transistors of a certain type is supposed to be at most 2.5 V. An investigator selects a sample of $n = 10$ such transistors and uses the resulting voltages as a basis for testing H_0: $\mu = 2.5$ versus H_a: $\mu > 2.5$ using a t test with significance level $\alpha = .05$. If the standard deviation of the voltage distribution is $\sigma = .100$, how likely is it that H_0 will be (correctly) rejected when $\mu = 2.6$?

For the values specified, the t critical value is $t_{.05,10-1} = 1.833$ and, using (9.4), the noncentrality parameter is $\delta = (2.6 - 2.5)/(.100/\sqrt{10}) = 3.162$. For this upper-tailed test,

$$
\begin{aligned}
\text{power} &= P(T \geq t_{.05,10-1} \text{ when } \mu = 2.6 \text{ rather than } 2.5) \\
&= P(T \geq 1.833 \text{ when } T \sim \text{noncentral } t, \ \text{df} = 9, \ \delta = 3.162) \\
&= 1 - F(1.833; 9, 3.162)
\end{aligned}
$$

The R command `pt(1.833, df = 9, ncp = 3.162)` reveals that $F(1.833; 9, 3.162) = .1025$, so the power under these circumstances is $1 - .1025 = .8975$. The value .1025 itself is $\beta(2.6)$.

Rather than compute one power value at a time, software can be instructed to create one or more power curves. Figure 9.5 shows Minitab power curves using the setting of this example for three different sample sizes: $n = 5$, 10, and 20. The horizontal axis, labeled Difference, represents the quantity $\mu' - \mu_0$. The previous power value of .8975 corresponds to the height of the $n = 10$ curve in Figure 9.5 at horizontal value $\mu' - \mu_0 = 2.6 - 2.5 = .1$.
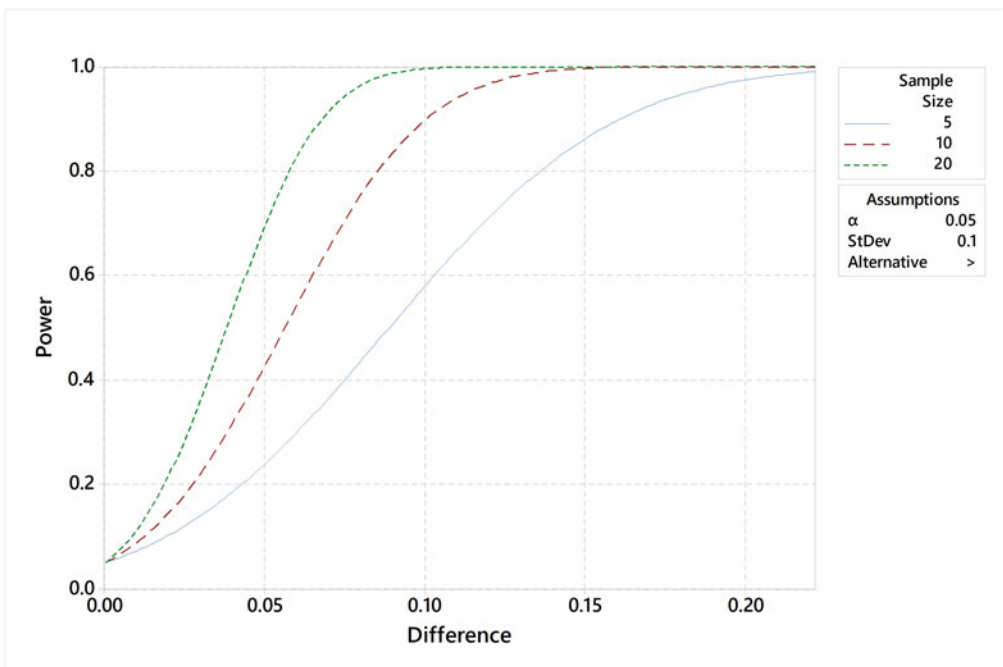
Figure 9.5   Power curves for Example 9.14

Figure 9.5 reveals two intuitive features of the power of this t test. First, for any fixed difference $\mu' - \mu_0$, power increases with sample size: the $n = 20$ power curve lies above the curves for the smaller sample sizes. That is, for any fixed departure from H_0, a larger sample size will increase the likelihood of correctly detecting that H_0 is false and H_a is true. Second, for any fixed sample size, power increases as the "Difference" increases, i.e., as the distance between μ' and μ_0 grows. We are more likely to reject H_0: $\mu = 2.5$ in favor of H_a: $\mu > 2.5$ if the true value of μ is 2.6, say, than if $\mu = \mu' = 2.51$, since the latter represents a very small departure from H_0 and is thus much more difficult to detect.

Software can also provide the sample size necessary to obtain a certain power or β at a specified alternative value μ'. For example, how large must n be to increase the power at $\mu' = 2.6$ to 95% (equivalently, reduce the chance of a type II error to $\beta(2.6) = .05$)? Figure 9.6 shows the result of making the appropriate request to Minitab, from which the answer $n = 13$ is obtained.

```
Power and Sample Size

1-Sample t Test

Testing mean = null (versus > null)
Calculating power for mean = null + 0.1
Alpha = 0.05 Sigma = 0.1

   Sample Size              Target Power              Actual Power
   13                       0.9500                    0.9597
```

Figure 9.6   Minitab sample size output for Example 9.14   ∎

Exercises: Section 9.2 (15–38)

15. Let the test statistic Z have a standard normal distribution when H_0 is true. Give the significance level for each of the following situations:

    a. H_a: $\mu > \mu_0$, rejection region $z \geq 1.88$
    b. H_a: $\mu < \mu_0$, rejection region $z \leq -2.75$
    c. H_a: $\mu \neq \mu_0$, rejection region $z \geq 2.88$ or $z \leq -2.88$

16. Let the test statistic T have a t distribution when H_0 is true. Give the significance level for each of the following situations:

    a. H_a: $\mu > \mu_0$, df $= 15$, rejection region $t \geq 3.733$
    b. H_a: $\mu < \mu_0$, $n = 24$, rejection region $t \leq -2.500$
    c. H_a: $\mu \neq \mu_0$, $n = 31$, rejection region $t \geq 1.697$ or $t \leq -1.697$

17. The true average diameter of ball bearings of a certain type is supposed to be .5 in. A one-sample t test will be carried out to see whether this is the case. What

conclusion is appropriate in each of the following situations?

    a. $n = 13$, $t = 1.6$, $\alpha = .05$
    b. $n = 13$, $t = -1.6$, $\alpha = .05$
    c. $n = 25$, $t = -2.6$, $\alpha = .01$
    d. $n = 25$, $t = -3.9$

18. The drying time (min) of a particular paint on a test board under controlled conditions is known to be normally distributed with $\mu = 75$ and $\sigma = 9$. A new additive has been developed for the purpose of improving drying time. The hypotheses H_0: $\mu = 75$ versus H_a: $\mu < 75$ are to be tested using a random sample of $n = 25$ observations. Assume drying times are still normally distributed with $\sigma = 9$.

    a. How many standard deviations (of $\bar{X}$) below the null value is $\bar{x} = 72.3$?
    b. If $\bar{x} = 72.3$, what is the conclusion using $\alpha = .01$?
    c. What is α for the test procedure that rejects H_0 when $z \leq -2.88$?

d. For the test procedure of part (c), what is $\beta(70)$?

e. If the test procedure of part (c) is used, what n is necessary to ensure that $\beta(70) = .01$?

f. If a level .01 test is used with $n = 100$, what is the probability of a type II error when $\mu = 76$?

19. The melting point of each of 16 samples of a brand of hydrogenated vegetable oil was determined, resulting in $\bar{x} = 94.32$. Assume that the distribution of melting point is normal with $\sigma = 1.20$.

   a. Test H_0: $\mu = 95$ versus H_a: $\mu \neq 95$ using a two-tailed level .01 test.

   b. If a level .01 test is used, what is $\beta(94)$, the probability of a type II error when $\mu = 94$?

   c. What value of n is necessary to ensure that $\beta(94) = .1$ when $\alpha = .01$?

20. Answer the following questions for the tire problem in Example 9.11.

   a. If $\bar{x} = 30{,}960$ and a level $\alpha = .01$ test is used, what is the decision?

   b. If a level .01 test is used, what is $\beta(30{,}500)$? What is the power at $\mu' = 30{,}500$ miles?

   c. If a level .01 test is used and it is also required that $\beta(30{,}500) = .05$, what sample size n is necessary?

   d. If $\bar{x} = 30{,}960$, what is the smallest α at which H_0 can be rejected (based on $n = 16$)?

21. Lightbulbs of a certain type are advertised as having an average lifetime of 750 h. The price of these bulbs is very favorable, so a potential customer has decided to go ahead with a purchase arrangement unless it can be conclusively demonstrated that the true average lifetime is smaller than what is advertised. A random sample of 50 bulbs was selected, the lifetime of each bulb determined, and the appropriate hypotheses were tested, resulting in the accompanying output.

```
Variable  N    Mean   StDev SEMean   Z    P-Value
Lifetime  50 738.44  38.20   5.40  −2.14   0.016
```

What conclusion would be appropriate for a significance level of .05? A significance level of .01? What significance level and conclusion would you recommend?

22. The industry standard for the amount of alcohol poured into many types of drinks (e.g., gin for a gin and tonic, whiskey on the rocks) is 1.5 oz. Each individual in a sample of 8 bartenders with at least 5 years of experience was asked to pour rum for a rum and coke into a short, wide (tumbler) glass, resulting in the following data:

2.00   1.78   2.16   1.91   1.70   1.67   1.83   1.48

(Summary quantities agree with those given in the article "Bottoms Up! The Influence of Elongation on Pouring and Consumption Volume," *J. Consumer Res.* 2003: 455–463.)

   a. What does a boxplot suggest about the distribution of the amount poured?

   b. Carry out a test of hypotheses to decide whether there is strong evidence for concluding that the true average amount poured differs from the industry standard.

   c. Does the validity of the test you carried out in (b) depend on any assumptions about the population distribution? If so, check the plausibility of such assumptions.

   d. Suppose the actual standard deviation of the amount poured is .20 oz. Determine the probability of a type II error for the test of (b) when the true average amount poured is actually (1) 1.6, (2) 1.7, (3) 1.8.

23. Exercise 46 in Chapter 1 gave $n = 26$ observations on escape time (sec) for oil workers in a simulated exercise, from which the sample mean and sample standard deviation are 370.69 and 24.36, respectively. Suppose the investigators had believed a priori that true average escape time would be at most 6 min. Does the data contradict this prior belief? Assuming normality, test the appropriate hypotheses using a significance level of .05.

24. Although the U.S. Food and Drug Administration recommends against using kitchen utensils to dose liquid medicines, many people still do so, resulting in dosing errors and even pediatric poisonings. The letter "Spoons Systematically Bias Dosing of Liquid Medicine" (*Annals of Internal Med.* 2010: 66–67) reported on an experiment involving a sample of 195 individuals. Each individual was asked to pour exactly 5 mL of a liquid medication into a medium-sized tablespoon whose capacity was 15 mL. The sample mean amount poured was 4.58 mL and the sample standard deviation was 2.55 mL. Does this data indicate that the true average amount poured is different from the desired dose? Test at the .05 level.

25. Consider the following core wood density measurements (g/mm^3) from a sample of 25 canopy trees in western Thailand ("Radial Variation of Wood Functional Traits Reflect Size-Related Adaptations of Tree Mechanics and Hydraulics," *Functional Ecology* 2017: 260–272)

391.2	431.0	447.1	375.3	470.7
543.7	592.7	546.7	601.8	598.8
492.3	454.4	548.7	494.9	585.6
647.8	639.2	700.4	640.1	620.5
755.2	668.7	644.6	717.7	663.0

a. Perform a hypothesis test at the .05 level to determine if the true mean core wood density differs from 600 g/mm^3.
b. This data appeared in Example 8.11, where a 95% CI for μ was computed to be (528.0, 613.8). Explain how the results of your hypothesis test in part (a) are consistent with this confidence interval.

26. The article "Development of Novel Industrial Laminated Planks from Sweetgum Lumber" (*J. Bridge Engr.* 2008: 64–66) provides the following data on the modulus of rupture (psi) for a sample of planks:

6807.99	7637.06	6663.28	6165.03	6991.41	6992.23
6981.46	7569.75	7437.88	6872.39	7663.18	6032.28
6906.04	6617.17	6984.12	7093.71	7659.50	7378.61
7295.54	6702.76	7440.17	8053.26	8284.75	7347.95
7422.69	7886.87	6316.67	7713.65	7503.33	7674.99

a. Perform a hypothesis test at the .01 level to determine if the true modulus of rupture for this type of plank differs from 7500 psi.
b. A 99% confidence interval for μ is (6929.7, 7476.7); this was calculated using the one-sample t interval of Chapter 8. Explain how the results of your hypothesis test in part (a) are consistent with this confidence interval.

27. On the label, Pepperidge Farm bagels are said to weigh four ounces each (113 g). A random sample of six bagels resulted in the following weights (in grams):

117.6 109.5 111.6 109.2 119.1 110.8

a. Based on this sample, is there any reason to doubt that the population mean is at least 113 g?
b. Suppose that the population mean is actually 110 g and that the distribution is normal with standard deviation 4 g. Based on a z test of H_0: $\mu = 113$ against H_a: $\mu < 113$ with $\alpha = .05$, find the probability of rejecting H_0 with six observations.
c. Under the conditions of part (b) with $\alpha = .05$, how many more observations would be needed in order for the power to be at least .95?

28. The target thickness for silicon wafers used in a type of integrated circuit is 245 μm. A sample of 50 wafers is obtained and the thickness of each one is determined, resulting in a sample mean thickness of 246.18 μm and a sample standard deviation of 3.60 μm. Does this data suggest that true average wafer thickness is something other than the target value? Test at the .10 level.

29. A well-designed and safe workplace can contribute greatly to increased productivity. It is especially important that workers not be

asked to perform tasks, such as lifting, that exceed their capabilities. The accompanying data on maximum weight of lift (MAWL, in kg) for a frequency of four lifts/min was reported in the article "The Effects of Speed, Frequency, and Load on Measured Hand Forces for a Floor-to-Knuckle Lifting Task" (*Ergonomics* 1992: 833–843); subjects were randomly selected from the population of healthy males age 18–30. Assuming that MAWL is normally distributed, does the following data suggest that the population mean MAWL exceeds 25? Test using a significance level of .05.

25.8   36.6   26.3   21.8   27.2

30. The article "The Foreman's View of Quality Control" (*Quality Engr* 1990: 257–280) described an investigation into the coating weights for large pipes resulting from a galvanized coating process. Production standards call for a true average weight of 200 lb per pipe. The accompanying descriptive summary and boxplot are from Minitab.

```
Variable N      Mean   Median TrMean StDev SE
                                           Mean
ctg wt   30     206.73 206.00 206.81 6.35  1.16
Variable Min    Max    Q1     Q3
ctg wt   193.00 218.00 202.75 212.00
```


Coating weight

190   200   210   220
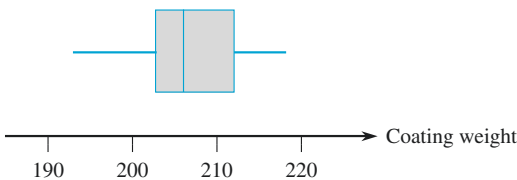
a. What does the boxplot suggest about the status of the specification for true average coating weight?

b. A normal probability plot of the data was quite straight. Use the descriptive output to test the appropriate hypotheses.

31. The amount of shaft wear (.0001 in.) after a fixed mileage was determined for each of $n = 8$ internal combustion engines having copper lead as a bearing material, resulting in $\bar{x} = 3.72$ and $s = 1.25$.

a. Assuming that the distribution of shaft wear is normal with mean μ, use the t test at level .05 to test $H_0: \mu = 3.50$ versus $H_a: \mu > 3.50$.

b. Using $\sigma = 1.25$, what is the type II error probability $\beta(\mu')$ of the test for the alternative $\mu' = 4.00$?

32. The recommended daily dietary allowance for zinc among males older than age 50 years is 15 mg/day. The article "Nutrient Intakes and Dietary Patterns of Older Americans: A National Study" (*J. Gerontol.* 1992: M145–150) reports the following summary data on intake for a sample of males age 65–74 years: $n = 115$, $\bar{x} = 11.3$, and $s = 6.43$. Does this data indicate that average daily zinc intake in the population of all males age 65–74 falls below the recommended allowance?

33. In an experiment designed to measure the time necessary for an inspector's eyes to become used to the reduced amount of light necessary for penetrant inspection, the sample average time for $n = 9$ inspectors was 6.32 s and the sample standard deviation was 1.65 s. It has previously been assumed that the average adaptation time was at least 7 s. Assuming adaptation time to be normally distributed, does the data contradict prior belief? Use the t test with $\alpha = .1$.

34. A sample of 12 radon detectors of a certain type was selected, and each was exposed to 100 pCi/L of radon. The resulting readings were as follows:

105.6   90.9   91.2   96.9   96.5   91.3
100.1   105.0   99.6   107.7   103.3   92.4

a. Does this data suggest that the population mean reading under these conditions differs from 100? State and test the appropriate hypotheses using $\alpha = .05$.

b. Suppose that prior to the experiment, a value of $\sigma = 7.5$ had been assumed. How many determinations would then have been appropriate to obtain $\beta = .10$ for the alternative $\mu = 95$? [*Note:* Software required.]

35. Show that for any $\Delta > 0$, when the population distribution is normal and σ is known, the two-tailed test satisfies $\beta(\mu_0 - \Delta) = \beta(\mu_0 + \Delta)$, so that $\beta(\mu')$ is symmetric about μ_0.

36. For a fixed alternative value μ', show that $\beta(\mu') \to 0$ as $n \to \infty$ for either a one-tailed or a two-tailed z test in the case of a normal population distribution with known σ.

37. Let $F(x; v, \delta)$ denote the cdf of the noncentral t distribution.

   a. Determine the power function of an upper-tailed one-sample t test in terms of

F. [*Hint:* Imitate the steps shown for the lower-tailed case in this section.]

   b. Repeat part (a) for the two-tailed one-sample t test.

38. Show that when $\mu = \mu'$, the one-sample t statistic (9.2) has a noncentral t distribution with $n - 1$ df and noncentrality parameter δ given by (9.4). [*Hint:* $(\bar{X} - \mu')/(\sigma/\sqrt{n})$ has a standard normal distribution. Re-write (9.2) and follow the steps in Section 6.4 that showed why $(\bar{X} - \mu)/(S/\sqrt{n})$ has a t_{n-1} distribution.]

9.3   Tests About a Population Proportion

Let p denote the proportion of individuals or objects in a population who possess a specified property (e.g., students who graduate college debt-free or former smokers who now vape). If an individual or object with the property is labeled a success (S), then p is the population proportion of successes. Tests concerning p will be based on a random sample of size n from the population. Provided that n is small relative to the population size, the rv $X =$ the number of S's in the sample has at least approximately a binomial distribution. Furthermore, if n itself is large, both X and the estimator $\hat{P} = X/n$ are approximately normally distributed. We first consider large-sample tests based on this latter fact and then turn to the small-sample case that directly uses the binomial distribution.

Large-Sample Tests

The estimator $\hat{P} = X/n$ is unbiased [$E(\hat{P}) = p$], has approximately a normal distribution, and its standard deviation is $\sigma_{\hat{P}} = \sqrt{p(1 - p)/n}$. These facts were used in Section 8.3 to obtain a confidence interval for p. When $H_0: p = p_0$ is true, $E(\hat{P}) = p_0$ and $\sigma_{\hat{P}} = \sqrt{p_0(1 - p_0)/n}$. It then follows that when n is large and H_0 is true, the test statistic

$$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1 - p_0)/n}} \tag{9.5}$$

has approximately a standard normal distribution.

Test procedures based on (9.5) can then be developed in a fashion similar to those of the first half of Section 9.2. For instance, if the alternative hypothesis is $H_a: p > p_0$ and the upper-tailed rejection region $z \geq z_\alpha$ is used, then

$$P(\text{type I error}) = P(H_0 \text{ is rejected when it is true})$$
$$= P(Z \geq z_\alpha \text{ when } Z \text{ has approximately a standard normal distribution}) \approx \alpha$$

Thus the desired level of significance α is attained by using the critical value that captures area α in the upper tail of the z curve. Rejection regions for the other two alternative hypotheses, lower-tailed for $H_a: p < p_0$ and two-tailed for $H_a: p \neq p_0$, are justified in an analogous manner.

THE ONE-PROPOR- TION z TEST	Null hypothesis: H_0: $p = p_0$

Test statistic value: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
H_a: $p > p_0$	$z \geq z_\alpha$ (upper-tailed)
H_a: $p < p_0$	$z \leq -z_\alpha$ (lower-tailed)
H_a: $p \neq p_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed)

These test procedures are valid provided that both $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

Example 9.15 Obesity is an increasing problem in America among all age groups. The article "Factors Affecting Obesity and Waist Circumference Among U.S. Adults" (*Prevention of Chronic Diseases* 2019) reported that 686 individuals in a sample of 2014 adult men were found to be obese (a body mass index exceeding 30; this index is a measure of weight relative to height). An earlier survey based on people's own assessment revealed that 20% of adult Americans considered themselves obese. Does the recent data suggest that the true proportion of men who are obese is more than 1.5 times the percentage from the self-assessment survey? Let's carry out a test of hypotheses using a significance level of .10.

1. $p =$ the proportion of all American men who are obese.
2. Saying that the current percentage is 1.5 times the self-assessment percentage is equivalent to the assertion that the current percentage is 30%, from which we have the null hypothesis H_0: $p = .30$. The phrase "more than" in the question implies that the alternative hypothesis is H_a: $p > .30$.
3. Since $np_0 = 2014(.3) \geq 10$ and $nq_0 = 2014(.7) \geq 10$, the large-sample z test can certainly be used.
4. The test statistic value is

$$z = (\hat{p} - .3)/\sqrt{(.3)(.7)/n}$$

5. The form of H_a implies that an upper-tailed test is appropriate: Reject H_0 if $z \geq z_{.10} = 1.28$.
6. $\hat{p} = 686/2014 = .341$, from which $z = (.341 - .3)/\sqrt{(.3)(.7)/2014} = 3.98$.
7. Since 3.98 exceeds the critical value 1.28, z lies in the rejection region. This justifies rejecting the null hypothesis. Using a significance level of .10, it does appear that more than 30% of American adult men are obese. ∎

Power, β, and Sample Size Determination for the One-Proportion z Test

When H_0 is true, the test statistic Z has approximately a standard normal distribution. Now suppose that H_0 is *not* true and that $p = p'$. Then Z still has approximately a normal distribution (because it is a linear function of $\hat{P}$), but its mean value and variance are no longer 0 and 1, respectively. Instead,

$$E(Z) = \frac{p' - p_0}{\sqrt{p_0(1 - p_0)/n}} \qquad V(Z) = \frac{p'(1 - p')/n}{p_0(1 - p_0)/n}$$

The power for an upper-tailed test is $P(Z \geq z_\alpha$ when $p = p')$, whereas the chance of a type II error is $\beta(p') = P(Z < z_\alpha$ when $p = p')$. These can be computed by using the given mean and variance to standardize and then referring to the standard normal cdf. In addition, if it is desired that the level α

test also have $\beta(p') = \beta$ for a specified value of β, this equation can be solved for the necessary n as in Section 9.2. General expressions for $\beta(p')$ and n are given in the accompanying box.

Alternative Hypothesis	$\beta(p')$
$H_a: p > p_0$	$\Phi\left(\dfrac{p_0 - p' + z_\alpha\sqrt{p_0 q_0/n}}{\sqrt{p'q'/n}}\right)$
$H_a: p < p_0$	$1 - \Phi\left(\dfrac{p_0 - p' - z_\alpha\sqrt{p_0 q_0/n}}{\sqrt{p'q'/n}}\right)$
$H_a: p \neq p_0$	$\Phi\left(\dfrac{p_0 - p' + z_{\alpha/2}\sqrt{p_0 q_0/n}}{\sqrt{p'q'/n}}\right) - \Phi\left(\dfrac{p_0 - p' - z_{\alpha/2}\sqrt{p_0 q_0/n}}{\sqrt{p'q'/n}}\right)$

where $q_0 = 1-p_0$, $q' = 1-p'$, and $\Phi(z) =$ the standard normal cdf. For each case, power $= 1 - \beta(p')$.

The sample size n for which the level α test also satisfies $\beta(p') = \beta$ is

$$
n = \begin{cases}
\left[\dfrac{z_\alpha\sqrt{p_0 q_0} + z_\beta\sqrt{p'q'}}{p' - p_0}\right]^2 & \text{one-tailed test} \\[3ex]
\left[\dfrac{z_{\alpha/2}\sqrt{p_0 q_0} + z_\beta\sqrt{p'q'}}{p' - p_0}\right]^2 & \text{two-tailed test (an approximate solution)}
\end{cases}
$$

Example 9.16 A package-delivery service advertises that at least 90% of all packages brought to its office by 9 a.m. for delivery in the same city are delivered by noon that day. Let p denote the true proportion of such packages that are delivered as advertised and consider the hypotheses $H_0: p = .9$ versus $H_a: p < .9$. If only 80% of all packages are delivered as advertised, how likely is it that a level .01 test based on $n = 225$ packages will detect such a departure from H_0? With $\alpha = .01$, $p_0 = .9$, $p' = .8$, and $n = 225$,

$$
\beta(.8) = 1 - \Phi\left[\frac{.9 - .8 - 2.33\sqrt{(.9)(.1)/225}}{\sqrt{(.8)(.2)/225}}\right] = 1 - \Phi(2.00) = .0228
$$

Thus the probability that H_0 will be rejected using the test when $p = .8$—the power of the test procedure—is $1 - .0228 = .9772$. Roughly 98% of all samples of size 225 will result in correct rejection of H_0.

What should the sample size be to ensure 99% power when p is actually .8? The 99% power requirement is equivalent to $\beta(.8) = .01$. Using $z_\alpha = z_\beta = z_{.01} = 2.33$ in the sample size formula yields

$$
n = \left[\frac{2.33\sqrt{(.9)(.1)} + 2.33\sqrt{(.8)(.2)}}{.8 - .9}\right]^2 \approx 266 \qquad\blacksquare
$$

Small-Sample Tests

Test procedures when the sample size n is small are based directly on the binomial distribution rather than the normal approximation. Consider the alternative hypothesis $H_a: p > p_0$ and again let X be the number of successes in the sample. Then X is the test statistic, and the upper-tailed rejection region has the form $x \geq c$. When H_0 is true, X has a binomial distribution with parameters n and p_0, so

$$P(\text{type I error}) = P(H_0 \text{ is rejected when it is true})$$
$$= P(X \geq c \text{ when } X \sim \text{Bin}(n, p_0))$$
$$= 1 - P(X \leq c - 1 \text{ when } X \sim \text{Bin}(n, p_0))$$
$$= 1 - B(c - 1; \ n, \ p_0)$$

As the critical value c decreases, more x values are included in the rejection region and $P(\text{type I error})$ increases. Because X has a discrete probability distribution, it is usually not possible to find a value of c for which $P(\text{type I error})$ is exactly the desired significance level α (e.g., .05 or .01). Instead, the largest rejection region of the form $\{c, c + 1, \ldots, n\}$ satisfying $1 - B(c - 1; n, p_0) \leq \alpha$ is used.

Let p' denote a value of p consistent with the alternative hypothesis (so $p' > p_0$). When $p = p'$, $X \sim \text{Bin}(n, p')$, so

$$\beta(p') = P(\text{type II error when } p = p') = P(X < c \text{ when } X \sim \text{Bin}(n, p'))$$
$$= B(c - 1; \ n, \ p')$$

and power $= 1 - \beta(p')$. Both of these are straightforward binomial probability calculations. On the other hand, the sample size n necessary to ensure that a level α test also has specified β at a particular alternative value p' must be determined by trial and error using the binomial cdf.

Test procedures for H_a: $p < p_0$ and for H_a: $p \neq p_0$ are constructed in a similar manner. In the former case, the appropriate rejection region has the form $x \leq c$ (a lower-tailed test). The critical value c is the largest number satisfying $B(c; n, p_0) \leq \alpha$. The rejection region when the alternative hypothesis is H_a: $p \neq p_0$ consists of both large and small x values.

Example 9.17 A plastics manufacturer has developed a new type of plastic trash can and proposes to sell them with an unconditional 6-year warranty. To see whether this is economically feasible, 20 prototype cans are subjected to an accelerated life test to simulate 6 years of use. The proposed warranty will be modified only if the sample data strongly suggests that fewer than 90% of such cans would survive the 6-year period. Let p denote the proportion of all cans that would survive the accelerated test. The relevant hypotheses are then H_0: $p = .9$ versus H_a: $p < .9$. A decision will be based on the test statistic X, the number among the 20 that survive. If the desired significance level is $\alpha = .05$, then c must satisfy $B(c; 20, .9) \leq .05$. From Appendix Table A.1, $B(15; 20, .9) = .043$ and $B(16; 20, .9) = .133$. The appropriate rejection region is therefore $x \leq 15$. If the accelerated test results in $x = 14$, H_0 would be rejected in favor of H_a, necessitating a modification of the proposed warranty. The probability of a type II error for the alternative value $p' = .8$ is

$$\beta(.8) = P(H_0 \text{ is not rejected when } X \sim \text{Bin}(20, \ .8))$$
$$= P(X > 15 \text{ when } X \sim \text{Bin}(20, \ .8))$$
$$= 1 - B(15; \ 20, \ .8) = 1 - .370 = .630$$

That is, when $p = .8$, 63% of all samples consisting of $n = 20$ cans would result in H_0 being incorrectly not rejected; the power of this test procedure is just 37%. This error probability is high because 20 is a small sample size and $p' = .8$ is close to the null value $p_0 = .9$. ∎

Exercises: Section 9.3 (39–48)

39. State DMV records indicate that of all vehicles undergoing emissions testing during the previous year, 70% passed on the first try. A random sample of 200 cars tested in a particular county during the current year yields 124 that passed on the initial test. Does this suggest that the true proportion for this county during the current year differs from the previous statewide proportion? Test the relevant hypotheses using $\alpha = .05$.

40. Natural cork in wine bottles is subject to deterioration, and as a result wine in such bottles may experience contamination. The article "Effects of Bottle Closure Type on Consumer Perceptions of Wine Quality" (*Amer. J. Enology Viticulture* 2007: 182–191) reported that in a tasting of commercial chardonnays, 16 of 91 bottles were considered spoiled to some extent by cork-associated characteristics. Does this data provide strong evidence for concluding that more than 15% of all such bottles are contaminated in this way? Carry out a test of hypotheses using a significance level of .10.

41. A manufacturer of nickel–hydrogen batteries randomly selects 100 nickel plates for test cells, cycles them a specified number of times, and determines that 14 of the plates have blistered.
    a. Does this provide compelling evidence for concluding that more than 10% of all plates blister under such circumstances? State and test the appropriate hypotheses using a significance level of .05. In reaching your conclusion, what type of error might you have committed?
    b. If it is really the case that 15% of all plates blister under these circumstances and a sample size of 100 is used, how likely is it that the null hypothesis of part (a) will not be rejected by the level .05 test? Answer this question for a sample size of 200.

    c. How many plates would have to be tested to have $\beta(.15) = .10$ for the test of part (a)?

42. A random sample of 150 recent donations at a blood bank reveals that 82 were type A blood. Does this suggest that the actual percentage of type A donations differs from 40%, the percentage of the population having type A blood? Carry out a test of the appropriate hypotheses using a significance level of .01. Would your conclusion have been different if a significance level of .05 had been used?

43. A university library ordinarily has a complete shelf inventory done once every year. Because of new shelving rules instituted the previous year, the head librarian believes it may be possible to save money by postponing the inventory. The librarian decides to select at random 1000 books from the library's collection and have them searched in a preliminary manner. If evidence indicates strongly that the true proportion of misshelved or unlocatable books is <.02, then the inventory will be postponed.
    a. Among the 1000 books searched, 15 were misshelved or unlocatable. Test the relevant hypotheses and advise the librarian what to do (use $\alpha = .05$).
    b. If the true proportion of misshelved and lost books is actually .01, what is the probability that the inventory will be (unnecessarily) taken?
    c. If the true proportion is .05, what is the probability that the inventory will be postponed?

44. The authors of the article "Luck of the Draw: Creating Chinese Brand Names" (*J. of Advertising Res.* 2008: 523–530) counted the number of "strokes" in the characters for the names of 1202 Chinese brand names. Certain totals for the number of strokes are considered lucky in Chinese culture, and the researchers hypothesized that a majority of Chinese brand names

would have a "lucky" number of strokes. Among the 1202 names sampled, 715 had a "lucky" number of strokes. Test the researchers' hypothesis at the $\alpha = .01$ significance level.

45. A plan for an executive traveler's club has been developed by an airline on the premise that 5% of its current customers would qualify for membership. A random sample of 500 customers yielded 40 who would qualify.

  a. Using this data, test at level .01 the null hypothesis that the company's premise is correct against the alternative that it is not correct.

  b. What is the probability that when the test of part (a) is used, the company's premise will be judged correct when in fact 10% of all current customers qualify?

46. Each of a group of 20 intermediate tennis players is given two rackets, one having nylon strings and the other synthetic gut strings. After several weeks of playing with the two rackets, each player will be asked to state a preference for one of the two types of strings. Let p denote the proportion of all such players who would prefer gut to nylon, and let X be the number of players in the sample who prefer gut. Because gut strings are more expensive, consider the null hypothesis that at most 50% of all such players prefer gut. We simplify this to H_0: $p = .5$, planning to reject H_0 only if sample evidence strongly favors gut strings.

  a. Which of the rejection regions $\{15, 16, 17, 18, 19, 20\}$, $\{0, 1, 2, 3, 4, 5\}$, or $\{0, 1, 2, 3, 17, 18, 19, 20\}$ is most appropriate, and why are the other two not appropriate?

  b. What is the probability of a type I error for the chosen region of part (a)? Does the region specify a level .05 test? Is it the best level .05 test?

  c. If 60% of all enthusiasts prefer gut, calculate the probability of a type II error using the appropriate region from part (a). Repeat if 80% of all enthusiasts prefer gut.

  d. If 13 out of the 20 players prefer gut, should H_0 be rejected using a significance level of .10?

47. A manufacturer of plumbing fixtures has developed a new type of washerless faucet. Let $p = P$(a randomly selected faucet of this type will develop a leak within 2 years under normal use). The manufacturer has decided to proceed with production unless it can be determined that p is too large; the borderline acceptable value of p is specified as .10. The manufacturer decides to subject n of these faucets to accelerated testing (approximating 2 years of normal use). With X = the number among the n faucets that leak before the test concludes, production will commence unless the observed X is too large. It is decided that if $p = .10$, the probability of not proceeding should be at most .10, whereas if $p = .30$ the probability of proceeding should be at most .10. Can $n = 10$ be used? $n = 20$? $n = 25$? What is the appropriate rejection region for the chosen n, and what are the actual error probabilities when this region is used?

48. Scientists have recently become concerned about the safety of Teflon cookware and various food containers because perfluorooctanoic acid (PFOA) is used in the manufacturing process. An article in the July 27, 2005, *New York Times* reported that of 600 children tested, 96% had PFOA in their blood. According to the FDA, 90% of all Americans have PFOA in their blood.

  a. Does the data on PFOA incidence among children suggest that the percentage of all children who have PFOA in their blood exceeds the FDA percentage for all Americans? Carry out an appropriate test of hypotheses.

  b. If 95% of all children have PFOA in their blood, how likely is it that the null hypothesis tested in (a) will be rejected when a significance level of .01 is employed?

  c. Referring back to (b), what sample size would be necessary for the relevant probability to be .10?

9.4   *P*-Values

Using the rejection region method to test hypotheses entails first selecting a significance level α. Then after computing the value of the test statistic, the null hypothesis H_0 is rejected if the value falls in the rejection region and is otherwise not rejected. We now consider another way of reaching a conclusion in a hypothesis-testing analysis. This alternative approach is based on calculation of a certain probability called a *P-value*. One advantage is that the *P*-value provides an intuitive measure of the strength of evidence in the data against H_0.

DEFINITION   The **P-value** is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample.

The definition is quite a mouthful! Here are some key points:

- The *P*-value is a probability.
- This probability is calculated assuming that H_0 is true.
- The *P*-value is a function of the sample data.
- To determine the *P*-value, we must decide which values of the test statistic are "at least as contradictory to H_0" as the value obtained from our sample.

Example 9.18 Urban storm water can be contaminated by many sources, including discarded batteries. When ruptured, these batteries release metals of environmental significance. The paper "Urban Battery Litter" (*J. Environ. Engr.* 2009: 46–57) presented summary data for characteristics of a variety of batteries found in urban areas around Cleveland. A sample of 51 Panasonic AAA batteries gave a sample mean zinc mass of 2.06 g and a sample standard deviation of .141 g. Does this data provide compelling evidence for concluding that the population mean zinc mass exceeds 2.0 g?

   With μ denoting the true average zinc mass (g) for such batteries, the relevant hypotheses are H_0: $\mu = 2.0$ versus H_a: $\mu > 2.0$. The sample size is large enough so that the one-sample t test can be used without making any specific assumption about the shape of the population distribution. The test statistic value is

$$t = \frac{\bar{x} - 2.0}{s/\sqrt{n}} = \frac{2.06 - 2.0}{.141/\sqrt{51}} = 3.04$$

Now we must decide which values of t are "at least as contradictory to H_0." Let's first consider an easier task: Which values of $\bar{x}$ are at least as contradictory to the null hypothesis as 2.06 g, the mean of the observations in our sample? Because $>$ appears in H_a, it should be clear that 2.10 g is at least as contradictory to H_0 as is 2.06, so is 2.25, and so in fact is *any* $\bar{x}$ value that exceeds 2.06. An $\bar{x}$ value that exceeds 2.06 g corresponds to a value of t that exceeds 3.04. Thus the *P*-value is

$$P\text{-value} = P(T \geq 3.04 \text{ when } \mu = 2.0)$$

Since the test statistic T was created by subtracting the null value 2.0 in the numerator, when $\mu = 2.0$ (i.e., when H_0 is true) T has approximately a t distribution with $51 - 1 = 50$ df. As a result,

$$P\text{-value} = P(T \geq 3.04 \text{ when } \mu = 2.0)$$
$$\approx \text{ area under the } t_{50} \text{ curve to the right of } 3.04$$
$$\approx .0019$$

The area under the *t* curve was determined using software.   ■

We will shortly illustrate how to determine the *P*-value for any *z* or *t* test; that is, any test where the reference distribution is the standard normal or some *t* distribution. For the moment, though, let's focus on reaching a conclusion once the *P*-value is available. Because it is a probability, the *P*-value must be between 0 and 1. What kinds of *P*-values provide evidence against the null hypothesis? Consider two specific instances:

- *P*-value = .250: In this case, fully 25% of all possible test statistic values are more contradictory to H_0 than the one that came out of our sample. So our data is not all that contradictory to the null hypothesis: even if H_0 is true, we'd see "more extreme" data than ours one-quarter of the time.
- *P*-value = .0019: Here, only .19% of all possible test statistic values are at least as contradictory to H_0 as what we obtained. Thus the sample appears to be highly contradictory to the null hypothesis.

More generally, *the smaller the P-value, the more evidence there is in the sample data against the null hypothesis and for the alternative hypothesis*. That is, H_0 should be rejected in favor of H_a when the *P*-value is sufficiently small. So what constitutes "sufficiently small"?

Whatever rule we use, it should not result in decisions that contradict the rejection region procedures we have seen previously. Consider, for instance, an upper-tailed *z* test at the $\alpha = .01$ level, for which the *z* critical value is $z_{.01} = 2.33$. Using precisely the logic of the previous example, the *P*-value of the hypothesis test should be the area under the standard normal curve to the right of the observed test statistic value *z*. Two such possible *P*-values are illustrated in Figure 9.7. But the rejection region already prescribes that we should reject H_0 if $z \geq 2.33$ and fail to reject H_0 if $z < 2.33$. Figure 9.7a shows that for any *z* value in the rejection region, the resulting *P*-value will be $\leq .01$; conversely, as seen in Figure 9.7b, the *P*-value will be $> .01$ precisely when $z < 2.33$, instructing us to not reject H_0.
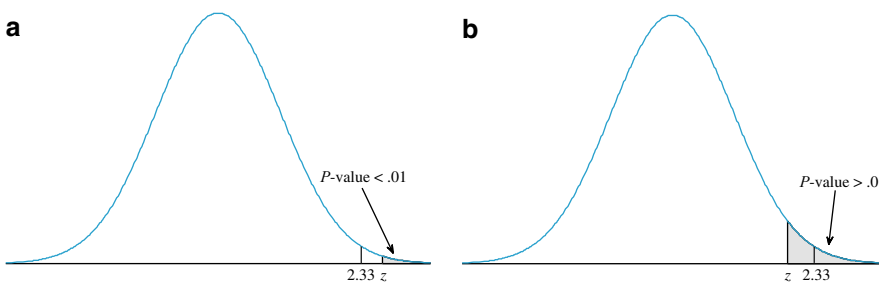
Figure 9.7   *P*-values for an upper-tailed *z* test: (**a**) *P*-value $\leq$ .01 if $z \geq 2.33$ (reject H_0); (**b**) *P*-value $> .01$ if $z < 2.33$ (do not reject H_0)

The preceding illustration generalizes to other tests (lower- and two-tailed, *t* as well as *z*) and other significance levels, leading to the following decision rule.

DECISION RULE BASED ON THE *P*-VALUE	Select a significance level α (as before, the desired type I error probability). Then reject H_0 if P-value $\leq \alpha$; do not reject H_0 if P-value $> \alpha$.

Figure 9.8 provides an easy way to visualize the decision rule. The calculation of the P-value depends on whether the test is upper-, lower-, or two-tailed. However, once it has been calculated, the comparison with α does not depend on which type of test was used.
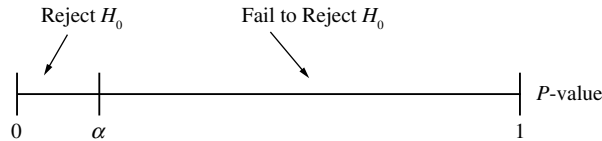
Figure 9.8 Comparing α and the P-value

   In Example 9.18, we calculated P-value = .0019. Then using a significance level of .01, we would reject the null hypothesis in favor of the alternative hypothesis because .0019 $\leq$ .01. However, suppose we had selected a significance level of .001, which requires more substantial evidence from the data before H_0 can be rejected. In this case we would not reject H_0 because .0019 > .001. Note that α should be specified *before* data is collected and the P-value calculated. It would be unethical to compute the P-value first and then select a significance level that would guarantee the desired outcome (e.g., deliberately choosing α greater than the P-value so that H_0 is rejected).

Example 9.19 The true average time to initial relief of pain for a best-selling pain reliever is known to be 10 min. Let μ denote the true average time to relief for a company's newly developed reliever. Suppose that when data from an experiment involving the new pain reliever was analyzed, the P-value for testing H_0: $\mu = 10$ versus H_a: $\mu < 10$ was calculated as .0384. Since the P-value is less than $\alpha = .05$, H_0 would be rejected by anyone carrying out the test at level .05. However, at level .01, H_0 would *not* be rejected because .0384 > .01. Again, α should be specified in advance of analyzing the data. ∎

   The most widely used statistical computer packages automatically include a P-value when a hypothesis-testing analysis is performed. A conclusion can then be drawn directly from the output, without reference to a table of critical values. With the P-value in hand, an investigator can see at a quick glance whether H_0 should be rejected at the prescribed α level. In addition, knowing the P-value allows a decision maker to distinguish between a close call (e.g., $\alpha = .05$, P-value = .0498) and a very clear-cut conclusion (e.g., $\alpha = .05$, P-value = .0003), something that would not be possible just from the statement "H_0 can be rejected at significance level .05."

P-Values for *z* Tests

The P-value for a z test (i.e., one based on a test statistic whose distribution when H_0 is true is at least approximately standard normal) is easily determined from the information in Appendix Table A.3. Consider an upper-tailed test and let z denote the computed value of the test statistic Z. As illustrated in Figure 9.7, the P-value is just the area to the right of the computed value z under the standard normal curve. The corresponding cumulative area is $\Phi(z)$, so in this case P-value $= 1 - \Phi(z)$. An analogous argument for a lower-tailed test shows that the P-value is the area captured by the computed value z in the lower tail of the standard normal curve, $\Phi(z)$.

More care must be exercised in the case of a two-tailed test. Suppose first that z is positive. We know to reject H_0 if and only if $z \geq z_{\alpha/2}$, which occurs precisely when $1 - \Phi(z) \leq \alpha/2$, or $2[1 - \Phi(z)] \leq \alpha$. Comparing this to the earlier decision rule, we infer that the P-value is precisely the quantity $2[1 - \Phi(z)]$. If z is negative, a similar argument leads to P-value $= 2[1 - \Phi(-z)]$. Since $-z = |z|$ when z is negative, the P-value $= 2[1 - \Phi(|z|)]$ for either positive or negative z.

$$P\text{-value} = \begin{cases} 1 - \Phi(z) & \text{for an upper-tailed test} \\ \Phi(z) & \text{for a lower-tailed test} \\ 2[1 - \Phi(|z|)] & \text{for a two-tailed test} \end{cases}$$

Each of these is the probability of getting a value at least as extreme as what was obtained (assuming H_0 true). The three cases are illustrated in Figure 9.9.
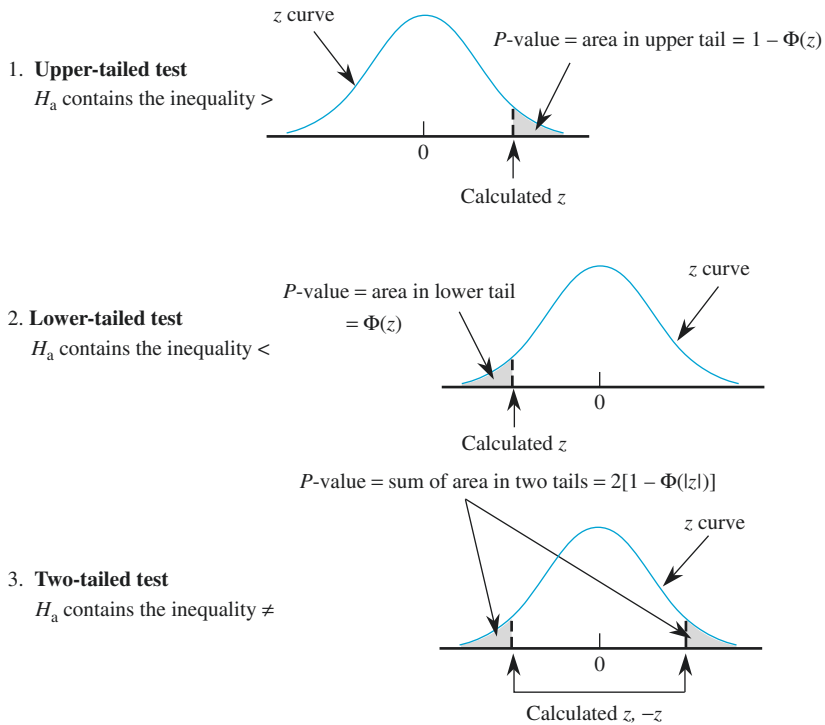
Figure 9.9 Determination of the P-value for a z test

The next example illustrates the use of the P-value approach to hypothesis testing by means of a sequence of steps modified from our previously recommended sequence.

Example 9.20 A Gallup poll (reported July 15, 2019) found that 29% of 1018 U.S. adults support statehood for the District of Columbia. Thirty years prior, 31% of U.S. adults held this opinion. Does the 2019 sample provide convincing statistical evidence at the $\alpha = .10$ level that the proportion of U.S. adults supporting DC statehood changed over those thirty years?

1. Parameter of interest: $p =$ proportion of all U.S. adults in 2019 who support statehood for the District of Columbia

2. Null hypothesis: H_0: $p = .31$ (no change since 1989)
   Alternative hypothesis: H_a: $p \neq .31$

3. Assuming H_0 is true, $np = np_0 = 1018(.31) \geq 10$ and $nq = nq_0 = 1018(1 - .31) \geq 10$. Thus, a one-proportion z test may be applied.

4. Formula for test statistic value: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \dfrac{\hat{p} - .31}{\sqrt{(.31)(.69)/n}}$

5. Calculation of test statistic value: $z = \dfrac{.29 - .31}{\sqrt{(.31)(.69)/1018}} = -1.38$

6. Determination of P-value: Because the test is two-tailed,

$$P\text{-value} = 2[1 - \Phi(|-1.38|)] = .1676$$

7. Conclusion: Using a significance level of .10, H_0 would not be rejected since $.1676 > .10$. At this significance level, there is insufficient evidence to conclude that the proportion of U.S. adults who support DC becoming a state changed over thirty years. ∎

P-Values for *t* Tests

Just as the P-value for a z test is a z curve area, the P-value for a t test will be a t curve area. Figure 9.9 illustrates the three possible cases: simply replace each z value or z curve with a t value or t curve. The number of df for the one-sample t test is $n - 1$.

The table of t critical values used previously for confidence and prediction intervals doesn't contain enough information about any particular t distribution to allow for accurate determination of desired areas, so we have included another t table in Appendix Table A.7, one that contains a tabulation of upper-tail t curve areas. Each different column of the table is for a different number of df, and the rows are for calculated values of the test statistic t ranging from 0.0 to 4.0 in increments of .1. For example, the number .074 appears at the intersection of the 1.6 row and the 8 df column, so the area under the 8 df curve to the right of 1.6 (an upper-tail area) is .074. Because t curves are symmetric, .074 is also the area under the 8 df curve to the left of -1.6 (a lower-tail area).

Suppose, for example, that a test of H_0: $\mu = 100$ versus H_a: $\mu > 100$ is based on the 8 df t distribution. If the calculated value of the test statistic is $t = 1.6$, then the P-value for this upper-tailed test is .074. Because .074 exceeds .05, we would not be able to reject H_0 at a significance level of .05. If the alternative hypothesis is H_a: $\mu < 100$ and a test based on 20 df yields $t = -3.2$, then Appendix Table A.7 shows that the P-value is the captured lower-tail area .002. The null hypothesis can be rejected at either level .05 or .01. Finally, for H_a: $\mu \neq 100$ if a t test is based on 20 df and $t = 3.2$, then the P-value for this two-tailed test is $2(.002) = .004$. This would also be the P-value for $t = -3.2$. The tail area is doubled because values both larger than 3.2 and smaller than -3.2 are more contradictory to H_0 than what was calculated (values farther out in *either* tail of the t curve; see the bottom graph in Figure 9.9).

Example 9.21 The recommended daily intake of calcium for adults ages 18–30 is 1000 mg/day. The article "Dietary and Total Calcium Intakes Are Associated with Lower Percentage Total Body and Truncal Fat in Young, Healthy Adults" (*J. Amer. College of Nutr.* 2011: 484–490) reported the following summary data for a sample of 76 healthy Caucasian males from southwestern Ontario, Canada: $n = 76$, $\bar{x} = 1093$, $s = 477$. Let's carry out a test at significance level .01 to see whether the population mean daily intake exceeds the recommended value.

1. μ = the mean daily calcium intake for this population (healthy Caucasian males from southwestern Ontario)

2. H_0: $\mu = 1000$
   H_a: $\mu > 1000$

3. Since $n = 76 > 40$, the one-sample t test is valid here (even if the calcium intake distribution is not normally distributed).

4. $t = \dfrac{\bar{x} - 1000}{s/\sqrt{n}}$

5. $t = \dfrac{1093 - 1000}{477/\sqrt{76}} = 1.70$

6. The *P*-value is the area under the t_{75} curve to the right of 1.70 (the inequality in H_a implies that the test is upper-tailed). From Table A.7, this area is between .047 (the upper-tail area at 60 df) and .046 (the upper-tail area at 120 df). Software gives a *P*-value of .0467.

7. Because this *P*-value is larger than .01, H_0 cannot be rejected. There is not compelling evidence to conclude at significance level .01 that the population mean daily intake exceeds the recommended value (even though the sample mean does so). Note that the opposite conclusion would result from using a significance level of .05. But the smaller α that we used requires more persuasive evidence from the data before rejecting H_0. ∎

More on Interpreting *P*-Values

The *P*-value resulting from carrying out a test on a selected sample is *not* the probability that H_0 is true, nor is it the probability of rejecting the null hypothesis. Once again, it is the probability, calculated assuming that H_0 is true, of obtaining a test statistic value at least as contradictory to the null hypothesis as the value that actually resulted. For example, consider testing H_0: $\mu = 50$ against H_0: $\mu < 50$ using a lower-tailed z test. If the calculated value of the test statistic is $z = -2.00$, then

$$P\text{-value} = \Phi(z) = \Phi(-2.00) = .0228$$

But if a second sample is selected, the resulting value of z will almost surely be different from -2.00, so the corresponding *P*-value will also likely differ from .0228. Because the test statistic value itself varies from one sample to another, the *P*-value will also vary from one sample to another. That is, the test statistic is a random variable, and so *the P-value will also be a random variable*. A first sample may give a *P*-value of .0228, a second sample result in a *P*-value of .1175, a third yield .0606 as the *P*-value, and so on.

   If H_0 is false, we hope the *P*-value will be close to 0 so that the null hypothesis can be rejected. On the other hand, when H_0 is true, we'd like the *P*-value to exceed the selected significance level so that the correct decision to not reject H_0 is made. The next example presents simulations to show how the *P*-value behaves both when the null hypothesis is true and when it is false.

Example 9.22 The fuel efficiency (mpg) of any particular new vehicle under specified driving conditions may not be identical to the EPA figure that appears on the vehicle's sticker. Suppose that four different vehicles of a particular type are to be selected and driven over a certain course, after which the fuel efficiency of each one is to be determined. Let μ denote the true average fuel efficiency under these conditions.

   Consider testing H_0: $\mu = 30$ versus H_0: $\mu > 30$ using the one-sample t test based on the resulting sample. Since the test is based on $n - 1 = 3$ degrees of freedom, the *P*-value for an upper-tailed test is the area under the t curve with 3 df to the right of the calculated t.

   Let's first suppose that H_0 is true. We used software to generate 10,000 different samples, each containing 4 observations, from a normal population distribution with mean value $\mu = 30$ and standard deviation $\sigma = 2$. The first sample and resulting summary quantities were

$$x_1 = 30.830, \ x_2 = 32.232, \ x_3 = 30.276, \ x_4 = 27.718 \Rightarrow$$
$$\bar{x} = 30.264 \quad s = 1.8864 \quad t = \frac{30.264 - 30}{1.8864/\sqrt{4}} = .2799$$

The P-value is the area under the t_3 curve to the right of .2799, which according to software is .3989. Using a significance level of .05, the null hypothesis would of course not be rejected. The values of t for the next four samples were -1.7591, .6082, $-.7020$, and 3.1053, with corresponding P-values .912, .293, .733, and .0265.

Figure 9.10a (p. 575) shows a histogram of the 10,000 P-values from this simulation experiment. About 4.5% of these P-values are in the first class interval from 0 to .05. Thus when using a significance level of .05, the null hypothesis is rejected in roughly 4.5% of these 10,000 tests. If we continue to generate samples and carry out the test for each one at significance level .05, in the long run 5% of the P-values will be in the first class interval—because when H_0 is true and a test with significance level .05 is used, by definition the probability of rejecting H_0 (i.e., of committing a type I error) is .05.

Looking at the histogram, it appears that the distribution of P-values is relatively flat. In fact, it can be shown that when H_0 is true, the probability distribution of the P-value is a uniform [0, 1] distribution. Since $P(U \leq .05) = .05$ for a Uniform [0, 1] rv, we again have that the probability of rejecting H_0 when it is true is .05, the chosen significance level.

Now consider what happens when H_0 is false because $\mu = 31$. We again generated 10,000 different samples of size 4, but now each from a normal distribution with $\mu = 31$ and $\sigma = 2$. The t statistic and P-value were calculated as before for each sample, and Figure 9.10b gives a histogram of the 10,000 resulting P-values. The shape of this histogram is quite different from that of Figure 9.10a: there is a much greater tendency for the P-value to be small (closer to 0) when $\mu = 31$ than when $\mu = 30$. Again H_0 is rejected at significance level .05 whenever the P-value is at most .05 (in the first class interval). Unfortunately this is the case for only about 19% of the 10,000 P-values. So only about 19% of the 10,000 tests correctly reject the null hypothesis (an estimate of the test's power); for the other 81%, a type II error is committed. The difficulty is that the sample size is extremely small and 31 is not very different from the value asserted by the null hypothesis.

Figure 9.10c illustrates what happens to the P-value when H_0 is false because $\mu = 32$ (still with $n = 4$ and $\sigma = 2$). The histogram is even more concentrated toward values close to 0 than was the case when $\mu = 31$. In general, as μ moves further to the right of the null value 30, the distribution of the P-values will become more and more concentrated on values close to 0. Even here, a bit fewer than 50% of the 10,000 P-values are smaller than .05. So it is still slightly more likely than not that the null hypothesis is incorrectly not rejected, principally because n is so small.

The big idea of this example is that because the value of any test statistic is random, the P-value will also be a random variable and thus have a distribution. The farther the actual value of the parameter is from the value specified by the null hypothesis, the more the distribution of the P-value will be concentrated on values close to 0 and the greater the chance that the test will correctly reject H_0 (corresponding to smaller β).

a $H_0: \mu = 30$ is true.
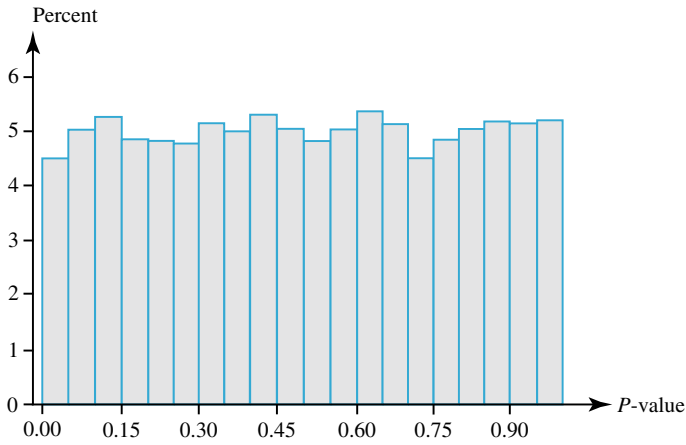
b H_0 is false because $\mu = 31$.
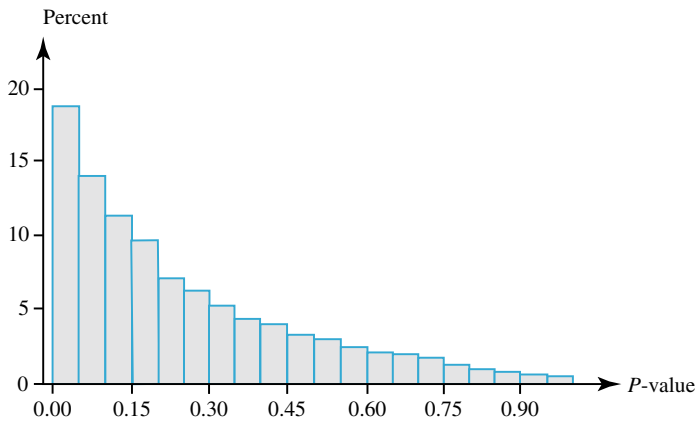
c H_0 is false because $\mu = 32$.
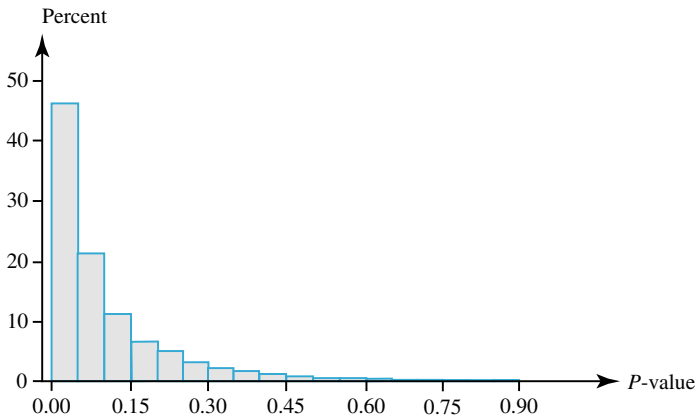
Figure 9.10 *P*-value simulation results for Example 9.22

Exercises: Section 9.4 (49–63)

49. For which of the given P-values would the null hypothesis be rejected when performing a level .05 test?
    a. 001   b. .021   c. .078   d. .047   e. .148

50. Pairs of P-values and significance levels, α, are given. For each pair, state whether the observed P-value would lead to rejection of H_0 at the given significance level.

    a. P-value = .084, α = .05
    b. P-value = .003, α = .001
    c. P-value = .498, α = .05
    d. P-value = .084, α = .10
    e. P-value = .039, α = .01
    f. P-value = .218, α = .10

51. Let μ denote the mean reaction time to a certain stimulus. For a one-sample z test of H_0: $\mu = 5$ versus H_a: $\mu > 5$ (i.e., assuming σ is known), find the P-value associated with each of the given values of the z test statistic.

    a. 1.42   b. .90   c. 1.96   d. 2.48   e. −.11

52. Newly purchased tires of a certain type are supposed to be filled to a pressure of 30 lb/in^2. Let μ denote the true average pressure. Find the P-value associated with each given one-sample z statistic value for testing H_0: $\mu = 30$ versus H_a: $\mu \neq 30$.

    a. 2.10   b. −1.75   c. −.55   d. 1.41   e. −5.3

53. Give as much information as you can about the P-value of a t test in each of the following situations:
    a. Upper-tailed test, df = 8, $t = 2.0$
    b. Lower-tailed test, df = 11, $t = -2.4$
    c. Two-tailed test, df = 15, $t = -1.6$
    d. Upper-tailed test, df = 19, $t = -.4$
    e. Upper-tailed test, df = 5, $t = 5.0$
    f. Two-tailed test, df = 40, $t = -4.8$

54. The paint used to make lines on roads must reflect enough light to be clearly visible at night. Let μ denote the true average reflectometer reading for a new type of paint under consideration. A test of H_0: $\mu = 20$ versus H_a: $\mu > 20$ will be based on a random sample of size n from a normal population distribution. What conclusion is appropriate in each of the following situations?
    a. $n = 15$, $t = 3.2$, $\alpha = .05$
    b. $n = 9$, $t = 1.8$, $\alpha = .01$
    c. $n = 24$, $t = -.2$

55. Let μ denote true average serum receptor concentration for all pregnant women. The average for all women is known to be 5.63. The article "Serum Transferrin Receptor for the Detection of Iron Deficiency in Pregnancy" (*Amer. J. Clin. Nutrit.* 1991: 1077–1081) reports that P-value > .10 for a test of H_0: $\mu = 5.63$ versus H_a: $\mu \neq 5.63$ based on $n = 176$ pregnant women. Using a significance level of .01, what would you conclude?

56. An aspirin manufacturer fills bottles by weight rather than by count. Since each bottle should contain 100 tablets, the average weight per tablet should be 5 grains. Each of 100 tablets taken from a very large lot is weighed, resulting in a sample average weight per tablet of 4.87 grains and a sample standard deviation of .35 grain. Does this information provide strong evidence for concluding that the company is not filling its bottles as advertised? Test the appropriate hypotheses using $\alpha = .01$ by first computing the P-value and then comparing it to the specified significance level.

57. Because of variability in the manufacturing process, the actual yielding point of a sample of mild steel subjected to increasing stress will usually differ from the theoretical yielding point. Let p denote the true proportion of samples that yield before their theoretical yielding point. If on the basis of

a sample it can be concluded that more than 20% of all specimens yield before the theoretical point, the production process will have to be modified.

  a. If 15 of 60 specimens yield before the theoretical point, what is the *P*-value when the appropriate test is used, and what would you advise the company to do?

  b. If the true percentage of "early yields" is actually 50% (so that the theoretical point is the median of the yield distribution) and a level .01 test is used, what is the probability that the company concludes a modification of the process is necessary?

58. Standard-size boxes for a particular brand of cereal indicate a net weight of 14 oz. A consumer group purchases a random sample of 50 such cereal boxes and weighs their contents. If the average of these 50 weights is 13.8 oz with a standard deviation of 1.1 oz, does the consumer group have sufficient evidence to conclude that the cereal company is under-filling its packages? Test at the $\alpha = .05$ level using the *P*-value method.

59. A random sample of soil specimens was obtained, and the amount of organic matter (%) in the soil was determined for each specimen, resulting in the accompanying data (from "Engineering Properties of Soil," *Soil Sci.* 1998: 93–102).

1.10	5.09	0.97	1.59	4.60	0.32	0.55	1.45
0.14	4.47	1.20	3.50	5.02	4.67	5.22	2.69
3.98	3.17	3.03	2.21	0.69	4.47	3.31	1.17
0.76	1.17	1.57	2.62	1.66	2.05		

The values of the sample mean and sample standard deviation are 2.481 and 1.616, respectively. Does this data suggest that the true average percentage of organic matter in such soil is something other than 3%? Carry out a test of the appropriate hypotheses at significance level .10 by first determining the *P*-value. Would your conclusion be different if $\alpha = .05$ had been used? [*Note*: A normal probability plot of the data shows an acceptable pattern in light of the reasonably large sample size.]

60. Repeat the analysis of Exercise 40 using the *P*-value method. Do you arrive at the same conclusion?

61. A pen has been designed so that true average writing lifetime under controlled conditions (involving the use of a writing machine) is at least 10 h. A random sample of 18 pens is selected, the writing lifetime of each is determined, and a normal probability plot of the resulting data supports the use of a one-sample *t* test.

  a. What hypotheses should be tested if the investigators believe a priori that the design specification has been satisfied?

  b. What conclusion is appropriate if the hypotheses of part (a) are tested, $t = -2.3$, and $\alpha = .05$?

  c. What conclusion is appropriate if the hypotheses of part (a) are tested, $t = -1.8$, and $\alpha = .01$?

  d. What should be concluded if the hypotheses of part (a) are tested and $t = -3.6$?

62. A spectrophotometer used for measuring CO concentration [ppm (parts per million) by volume] is checked for accuracy by taking readings on a manufactured gas (called span gas) in which the CO concentration is very precisely controlled at 70 ppm. If the readings suggest that the spectrophotometer is not working properly, it will have to be recalibrated. Assume that if it is properly calibrated, measured concentration for span gas samples is normally distributed. On the basis of the six readings —85, 77, 82, 68, 72, and 69—is recalibration necessary? Carry out a test of the relevant hypotheses using the *P*-value approach with $\alpha = .05$.

63. The relative conductivity of a semiconductor device is determined by the amount of impurity "doped" into the device during its manufacture. A silicon diode to be used for a specific purpose requires an average cut-on voltage of .60 V, and if this is not achieved, the amount of impurity must be adjusted. A sample of diodes was selected and the cut-on voltage was determined. The accompanying SAS output resulted from a request to test the appropriate hypotheses.

```
N     Mean      Std Dev      T     Prob >|T|
15  0.6453333  0.0899100  1.9527887  0.0711
```

[*Note*: By default, SAS's *P*-value is for a two-tailed test.] What would be concluded for a significance level of .01? .05? .10?

9.5   The Neyman–Pearson Lemma and Likelihood Ratio Tests

The test procedures presented thus far are (hopefully) intuitively reasonable, but have not been shown to be "best" in any sense. How can an optimal test be obtained, one for which the type II error probability is as small as possible, subject to controlling the type I error probability at the desired level?

Simple Hypotheses

Our starting point here will be a rather unrealistic situation from a practical viewpoint: testing a *simple* null hypothesis against a *simple* alternative hypothesis. A **simple hypothesis** is one which, when true, completely specifies the distribution of the sample X_i's. Suppose, for example, that $X_1, \ldots, X_n$ form a random sample from an exponential distribution with parameter λ. Then the hypothesis $H: \lambda = 5$ is simple, since when H is true each X_i has an exponential distribution with parameter $\lambda = 5$. We might then consider $H_0: \lambda = 5$ versus $H_a: \lambda = 10$, both of which are simple hypotheses. The hypothesis $H_a: \lambda < 5$ is *not* simple, because when H_a is true, the distribution of each X_i might be exponential with $\lambda = 4$ or with $\lambda = 2.8$ or ….

Similarly, if the X_i's constitute a random sample from a normal distribution with *known* σ, then $H: \mu = 100$ is a simple hypothesis. But if the value of σ is unknown, this hypothesis is not simple because the distribution of each X_i is not completely specified; it could be $N(100, 15)$ or $N(100, 12)$ or $N(100, \sigma)$ for any other positive value of σ. For a hypothesis to be simple, the value of *every* parameter in the pmf or pdf of the X_i's must be specified.

Throughout this chapter we have always employed **composite** (that is, not simple) alternative hypotheses. In practice, a pair of simple hypotheses such as $H_0: \lambda = 5$ versus $H_a: \lambda = 10$ are almost never tested, since they imply that no other λ value is possible (what if both are false because $\lambda = 7.6$?). However, when hypothesis testing was developed about a century ago, early statistical pioneers developed optimal methods for a pair of simple hypotheses and then built up from that foundation.

The Neyman–Pearson Lemma

The next result was a milestone in the theory of hypothesis testing—a method for constructing a best test for a pair of simple hypotheses. Let $f(x_1, \ldots, x_n; \theta)$ be the joint pmf or pdf of the X_i's. Our simple null hypothesis will assert that $\theta = \theta_0$ and the simple alternative hypothesis will claim that $\theta = \theta_a$. The result carries over to the case of more than one parameter as long as the value of each parameter is completely specified in both H_0 and H_a.

THE NEYMAN–
PEARSON LEMMA

For testing a simple null hypothesis H_0: $\theta = \theta_0$ versus a simple alternative hypothesis H_a: $\theta = \theta_a$, let k be a fixed positive number and form the rejection region

$$R^* = \left\{ (x_1, \ldots, x_n) : \frac{f(x_1, \ldots, x_n; \theta_a)}{f(x_1, \ldots, x_n; \theta_0)} \geq k \right\} \tag{9.6}$$

Let $\alpha^* = P((X_1, \ldots, X_n) \in R^*$ when $\theta = \theta_0)$, the probability of a type I error using R^*, and let β^* denote the type II error probability (i.e., the probability that the X_i's lie in the complement of R^* when $\theta = \theta_a$).

Then for any other test procedure with type I error probability α satisfying $\alpha \leq \alpha^*$, the probability of a type II error must satisfy $\beta \geq \beta^*$. That is, the test with rejection region R^* has the smallest type II error probability among *all* tests for which the type I error probability is at most α^*.

The test statistic value in (9.6) is called a *likelihood ratio*—it's the ratio of the alternative likelihood to the null likelihood. We'll explore likelihood ratio tests more deeply later in this section. As in previous sections of this chapter, the constant k in the rejection region is tied to the type I error probability α^*. In the continuous case, k can be selected to give one of the traditional significance levels (.05, .01, and so on), whereas in the discrete case $\alpha^* = .057$ or $.039$ may be as close as one can get to .05.

Roughly speaking, the Neyman–Pearson Lemma prescribes, subject to a given significance level, the test procedure that minimizes the chance of committing a type II error. Equivalently, it maximizes the power of the hypothesis test—that is, R^* in (9.6) defines the *most powerful* test of the simple hypotheses H_0: $\theta = \theta_0$ versus H_a: $\theta = \theta_a$ at its level of significance.

Example 9.23   As part of quality control at a semiconductor plant, consider randomly selecting $n = 5$ newly-made integrated circuits of a certain type and determining the number of defects on each one. Let X_i denote the number of such defects for the ith selected circuit ($i = 1, \ldots, 5$), and suppose that the X_i's form a random sample from a Poisson distribution with parameter μ. Let's find the best test for testing H_0: $\mu = 1$ versus H_a: $\mu = 2$. The Poisson likelihood is $f(x_1, \ldots, x_5; \mu) = e^{-5\mu} \mu^{\Sigma x_i} / \Pi x_i!$. Substituting first $\mu = 2$, then $\mu = 1$, and then taking the ratio of these two likelihoods as in (9.6) gives the rejection region

$$R^* = \left\{ (x_1, \ldots, x_5) : e^{-5} 2^{\Sigma x_i} \geq k \right\}$$

Multiplying both sides of the inequality by e^5 and taking a logarithm allows us to re-write the rejection region as $\sum x_i \geq c$, where $c = \ln(ke^5)/\ln(2)$.

This latter rejection region is completely equivalent to R^*: for any particular value k there will be a corresponding value c, and vice versa. But it is much easier to express the rejection region in this latter form and then select c to obtain a desired significance level than it is to determine an appropriate value of k for the likelihood ratio. In particular, the rv $Y = \sum X_i$ has a Poisson distribution with parameter 5μ (via a moment generating function argument), so when H_0 is true $Y \sim$ Poisson(5). If we use $c = 10$ in the rejection region, then from Table A.2

$$\alpha^* = P(Y \geq 10 \text{ when } Y \sim \text{Poisson}(5)) = 1 - .968 = .032$$

Choosing instead $c = 9$ gives $\alpha^* = .068$. If we insist that the significance level be at most .05, then the optimal rejection region is $R^* = \{(x_1, \ldots, x_5) : \sum x_i \geq 10\}$, and $\alpha^* = P(\text{type I error}) = .032$.

When H_a is true, the test statistic Y has a Poisson distribution with parameter $5(2) = 10$. Thus

$$\beta^* = P(H_0 \text{ is } not \text{ rejected when } H_a \text{ is true})$$
$$= P(Y < 10 \text{ when } Y \sim \text{Poisson}(10)) = .458$$

The Neyman–Pearson Lemma guarantees that *any* other test procedure based on these 5 observations, provided its type I error probability is $\leq .032$, must necessarily have a type II error probability greater than or equal to .458. Equivalently, every test in this situation with $\alpha \leq .032$ has power no better than $1 - .458 = .542$; to increase power here would require increasing α^*.

Obviously the type II error probability here is quite large (and the power rather low). This is because the sample size $n = 5$ is too small to allow for effective discrimination between $\mu = 1$ and $\mu = 2$. For a sample size of 10, the best test having significance level at most .05 uses $c = 16$, for which $\alpha^* = .049$ (Poisson parameter = 10) and $\beta^* = .157$ (Poisson parameter = 20).

Finally, returning to a sample size of $n = 5$, $c = 10$ implies that $10 = \ln(ke^5)/\ln(2)$, from which $k = 2^{10}/e^5 \approx 6.9$. For the best test to have a significance level of at most .05, the null hypothesis should be rejected only when the likelihood for the alternative value of μ is more than about 7 times what it is for the null value. ∎

Example 9.24 Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance 1; the argument to be presented will work for any other *known* value of σ. Consider testing $H_0: \mu = \mu_0$ versus $H_a: \mu = \mu_a$ where $\mu_a > \mu_0$. The likelihood ratio in (9.6) is

$$\frac{\left(\frac{1}{2\pi}\right)^{n/2} e^{-(1/2)\Sigma(x_i - \mu_a)^2}}{\left(\frac{1}{2\pi}\right)^{n/2} e^{-(1/2)\Sigma(x_i - \mu_0)^2}} = e^{(\mu_a - \mu_0)\Sigma x_i - (n/2)\left(\mu_a^2 - \mu_0^2\right)}$$
$$= \left[e^{-n\left(\mu_a^2 - \mu_0^2\right)/2} \right] \cdot \left[e^{(\mu_a - \mu_0)\Sigma x_i} \right]$$

The term in the first set of brackets is a numerical constant. Then $\mu_a - \mu_0 > 0$ implies that the likelihood ratio will be at least k if and only if $\sum x_i \geq k'$ for some k', that is, if and only if $\bar{x} \geq k''$ for some k'', which means if and only if

$$z = \frac{\bar{x} - \mu_0}{1/\sqrt{n}} \geq c$$

for some c. When H_0 is true, the rv Z has a standard normal distribution (because $\sigma = 1$; again, this argument works for any σ). If we now let $c = z_{.01} = 2.33$, then $\alpha^* = P(Z \geq c) = .01$. By the Neyman–Pearson Lemma, our old friend the one-sample z test has minimum β among all tests for which $\alpha \leq .01$. ∎

Proof of the Neyman–Pearson Lemma We shall consider the case in which the X_i's have a discrete distribution, so that type I and type II error probabilities are obtained by summation. In the continuous case, integration replaces summation. Let R denote the rejection region of any test procedure based on the X_i's, so that

$$\alpha = P((X_1, \ldots, X_n) \in R \text{ when } \theta = \theta_0) = \sum_R f(x_1, \ldots, x_n; \theta_0)$$
$$\beta = P((X_1, \ldots, X_n) \in R' \text{ when } \theta = \theta_a) = 1 - \sum_R f(x_1, \ldots, x_n; \theta_a)$$

(β is the probability *outside* the rejection region R, and the complement rule has been applied). Next, let $k > 0$ be any constant, and consider the linear combination $k\alpha + \beta$:

$$k\alpha + \beta = k \sum_R f(x_1, \ldots, x_n; \theta_0) + 1 - \sum_R f(x_1, \ldots, x_n; \theta_a)$$

$$= 1 + \sum_R [k \cdot f(x_1, \ldots, x_n; \theta_0) - f(x_1, \ldots, x_n; \theta_a)]$$

The expression in brackets can be positive or negative. Now comes the clever part: among all possible test procedures, $k\alpha + \beta$ is minimized by choosing R to be exactly the set where the expression in brackets is negative (or zero). That is, $k\alpha + \beta$ is minimized by using the rejection region

$$\{(x_1, \ldots, x_n): k \cdot f(x_1, \ldots, x_n; \theta_0) - f(x_1, \ldots, x_n; \theta_a) \leq 0\} = \left\{ (x_1, \ldots, x_n): \frac{f(x_1, \ldots, x_n; \theta_a)}{f(x_1, \ldots, x_n; \theta_0)} \geq k \right\},$$

which is precisely R^* from (9.6).

   With α^* and β^* defined as in the statement of the Neyman–Pearson Lemma, what we have established is that using R^* minimizes $k\alpha + \beta$, i.e., that $k\alpha^* + \beta^* \leq k\alpha + \beta$ for all *other* choices of rejection region. In particular, for all test procedures satisfying $\alpha \leq \alpha^*$, $\alpha - \alpha^* \leq 0$, and so for these test procedures

$$k\alpha^* + \beta^* \leq k\alpha + \beta \Rightarrow \beta^* \leq \beta + k(\alpha - \alpha^*) \leq \beta + 0 = \beta$$

Thus we have shown that $\beta^* \leq \beta$ for all such procedures, as desired.   ∎

   An essentially identical argument shows that the same rejection region (9.6) can be used to minimize the chance of a type I error, subject to a constraint on the type II error probability. That is, with the same notation as above, the chance of a type I error is $\geq \alpha^*$ for all test procedures for which $P(\text{type II error}) \leq \beta^*$.

Power and Uniformly Most Powerful Tests

The Neyman–Pearson Lemma identifies the most powerful test procedure when both hypotheses are simple. Next consider the more realistic scenario where one or both of the hypotheses are composite. In previous sections, the term *power* was primarily used when H_a was true (the chance of *correctly* rejecting H_0). The following definition generalizes this idea.

DEFINITION   Let Ω_0 and Ω_a be two disjoint sets of possible values of θ, and consider testing $H_0\colon \theta \in \Omega_0$ versus $H_a\colon \theta \in \Omega_a$ using a test with rejection region R. Then the **power function** of the test, denoted by $\pi(\cdot)$, is the probability of rejecting H_0 considered as a function of θ:

$$\pi(\theta') = P((X_1, \ldots, X_n) \in R \text{ when } \theta = \theta')$$

The power function is easily related to the type I and type II error probabilities:

$$\pi(\theta') = \begin{cases} P(\text{type I error when } \theta = \theta') = \alpha(\theta') & \text{when } \theta' \in \Omega_0 \\ 1 - P(\text{type II error when } \theta = \theta') = 1 - \beta(\theta') & \text{when } \theta' \in \Omega_a \end{cases}$$

Since we don't want to reject the null hypothesis when $\theta \in \Omega_0$ and do want to reject it when $\theta \in \Omega_a$, we desire a test for which the power function is close to 0 whenever θ' is in Ω_0 and close to 1 whenever θ' is in Ω_a. The **ideal power function**, though not achievable in practice, is

$$\pi(\theta') = \begin{cases} 0 & \text{when } \theta' \in \Omega_0 \\ 1 & \text{when } \theta' \in \Omega_a \end{cases}$$

Example 9.25 The drying time (min) of a particular paint on a test board under controlled conditions is known to be normally distributed with $\mu = 75$ and $\sigma = 9$. A new additive has been developed for the purpose of improving drying time. Assume that drying time with the additive is still normally distributed with the same standard deviation, and consider testing H_0: $\mu \geq 75$ versus H_a: $\mu < 75$ based on a sample of size $n = 100$. A test with significance level .10 rejects H_0 if $z \leq -z_{.10} = -1.28$, where $z = (\bar{x} - 75)/(9/\sqrt{100}) = (\bar{x} - 75)/.9$. Manipulating the inequality in the rejection region to isolate $\bar{x}$ gives the equivalent rejection region $\bar{x} \leq 73.848$.

   If $\mu = \mu'$, then $\bar{X}$ has a normal distribution with mean μ' and standard deviation $\sigma/\sqrt{n} = .9$. Thus the power function of the test is

$$\pi(\mu') = P(\bar{X} \leq 73.848 \text{ when } \mu = \mu') = \Phi\left(\frac{73.848 - \mu'}{.9}\right)$$

The ideal power function for these hypotheses equals 0 for $\mu \geq 75$ (H_0 is true) and equals 1 for $\mu < 75$ (H_a is true). Figure 9.11 shows both the actual power function $\pi(\mu')$ and the ideal function. The maximum power for $\mu \geq 75$ (i.e., in Ω_0) occurs at $\mu = 75$, on the boundary between Ω_0 and Ω_a; specifically, $\pi(75) = .10 = \alpha$ by design. Because the power function is continuous, there are values of μ smaller than 75 for which the power is quite small (barely above .10). Even with a large sample size, it is difficult to detect a very small departure from the null hypothesis. But as n increases, the actual power function will approach the ideal.
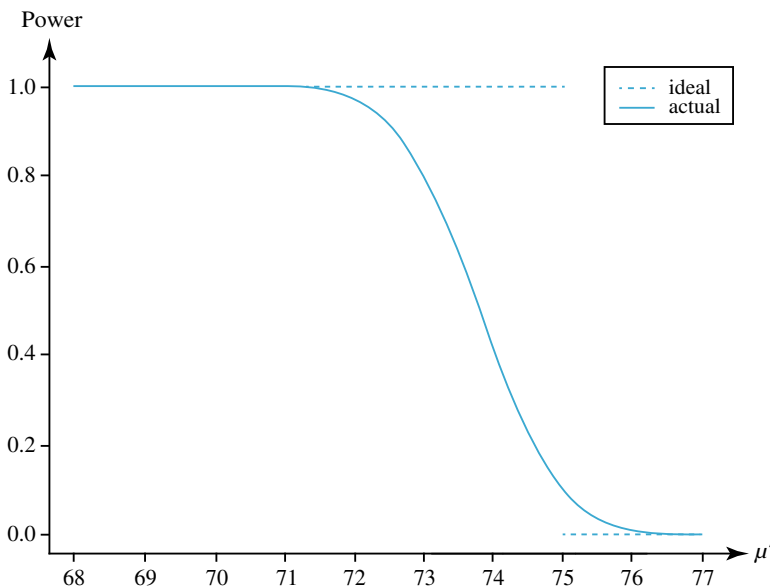
Figure 9.11 Graphs of power functions for Example 9.25   ∎

The Neyman–Pearson lemma says that when Ω_0 consists of a single value θ_0 and Ω_a also consists of a single value θ_a, the rejection region R^* in (9.6) specifies a test for which the power $\pi(\theta_a)$ at the alternative value θ_a is maximized subject to $\pi(\theta_0) \leq \alpha$ for some specified value of α. That is, R^* specifies a most powerful test subject to the restriction on the power when the null hypothesis is true. What about best tests when at least one of the two hypotheses is composite?

Example 9.26 (Example 9.23 continued) Consider again a random sample of size $n = 5$ from a Poisson distribution, and suppose we now wish to test $H_0: \mu \leq 1$ versus $H_a: \mu > 1$. Both of these hypotheses are composite. Arguing as in Example 9.23, for any value μ_a exceeding 1, the most powerful test of $H_0: \mu = 1$ versus $H_a: \mu = \mu_a$ with significance level equal to .032 (i.e., $\pi(1) = .032$) rejects the null hypothesis when $\sum x_i \geq 10$. Furthermore, it is easily verified that the $\pi(\mu') < .032$ for $\mu' < 1$.

Thus the test that rejects $H_0: \mu \leq 1$ in favor of $H_0: \mu > 1$ when $\sum x_i \geq 10$ has maximum power for *any* $\mu' = \mu_a > 1$, subject to the condition that $\pi(\mu') \leq \pi(1) = .032$ whenever $\mu' \leq 1$. This test is *uniformly most powerful*. ∎

More generally, a **uniformly most powerful (UMP) level α test** is one for which $\pi(\theta')$ is maximized for every $\theta' \in \Omega_a$ subject to $\pi(\theta') \leq \alpha$ for $\theta' \in \Omega_0$. Unfortunately UMP tests are fairly rare, especially in commonly encountered situations when H_0 and H_a are assertions about a single parameter θ while the distribution of the X_i's involves at least one other "nuisance parameter." For example, when the population distribution is normal with values of both μ and σ unknown, σ is a nuisance parameter when testing $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$. Be careful here—the null hypothesis is *not* simple, because Ω_0 consists of all pairs (μ, σ) for which $\mu = \mu_0$ and $\sigma > 0$, and there is certainly more than one such pair. In this situation, the one-sample t test is not UMP.

However, suppose we restrict attention to **unbiased tests**, those for which the smallest value of $\pi(\theta')$ for $\theta' \in \Omega_a$ is at least as large as the largest value of $\pi(\theta')$ for $\theta' \in \Omega_0$. Unbiasedness simply says that we are at least as likely to reject the null hypothesis when H_0 is false as we are to reject it when H_0 is true. The test proposed in Example 9.25 involving paint drying times is unbiased because, as Figure 9.11 shows, the power function at or to the right of 75 is smaller than it is to the left of 75. It can be shown that the one-sample t test is *UMP unbiased*; that is, it is uniformly most powerful among all tests that are unbiased. Several other commonly used tests also have this property. Please consult the references by Casella and Berger or DeGroot and Schervish for more details on UMP tests.

Likelihood Ratio Tests

The *likelihood ratio principle*, described below, is a frequently used method for finding an appropriate test statistic in a new situation. As before, denote the joint pmf or pdf of $X_1, \ldots, X_n$ by $f(x_1, \ldots, x_n; \theta)$. In the case of a random sample, it will be a product $f(x_1; \theta) \ldots f(x_n; \theta)$. As in the development of maximum likelihood estimates, when $f(x_1, \ldots, x_n; \theta)$ is regarded as a function of θ, it is called the likelihood function and is sometimes denoted $L(\theta)$.

Again consider testing $H_0: \theta \in \Omega_0$ versus $H_a: \theta \in \Omega_a$, where Ω_0 and Ω_a are disjoint sets, and let $\Omega = \Omega_0 \cup \Omega_a$. The set Ω is called the **parameter space**, since it represents all possible values of the parameter θ under consideration. In the Neyman–Pearson Lemma, the test statistic is the ratio of the likelihood when $\theta \in \Omega_a = \{\theta_a\}$ to the likelihood when $\theta \in \Omega_0 = \{\theta_0\}$, rejecting H_0 when the value of the ratio is "sufficiently large." For one or more composite hypotheses, we instead consider the ratio of the likelihood when $\theta \in \Omega_0$ to the likelihood when $\theta \in \Omega$; the latter effectively puts no constraints on the value of θ. A very *small* value of this ratio argues against the null hypothesis, since a small value arises when the data is much more consistent with H_a than with H_0. More formally,

1. Find the largest value of $L(\theta)$ for $\theta \in \Omega$ by finding the maximum likelihood estimate of θ; denote this estimate by $\hat{\theta}_{mle}$. Substitute this mle into the likelihood function to obtain $L(\hat{\theta}_{mle})$.
2. Find the largest value of $L(\theta)$ for $\theta \in \Omega_0$ by finding the maximum likelihood estimate of θ *within* Ω_0; denote this estimate by $\hat{\theta}_0$. Substitute this *restricted* mle into the likelihood function to obtain $L(\hat{\theta}_0)$.

   Because Ω_0 is a subset of Ω, this restricted likelihood $L(\hat{\theta}_0)$ can't be any larger than the likelihood $L(\hat{\theta}_{mle})$ obtained in the first step, and will be much smaller when the data is much more consistent with H_a than with H_0.
3. Form the **likelihood ratio test statistic**

$$\Lambda = \frac{L(\hat{\theta}_0)}{L(\hat{\theta}_{mle})} = \frac{f(x_1, \ldots, x_n; \hat{\theta}_0)}{f(x_1, \ldots, x_n; \hat{\theta}_{mle})}$$

and reject the null hypothesis in favor of the alternative when this ratio is $\leq k$. The critical value k is chosen to give a test with the desired significance level. In practice, the inequality $\Lambda \leq k$ is often re-expressed in terms of a more convenient statistic (such as the sum or mean of the observations) whose distribution is known or can be derived.

The above prescription, called a **likelihood ratio test**, remains valid if the single parameter θ is replaced by several parameters $\theta_1, \ldots, \theta_m$. The mles of all parameters must be obtained in both steps 1 and 2 and substituted back into the likelihood function.

Example 9.27 Consider a random sample from a normal distribution with the values of both parameters unknown. We wish to test H_0: $\mu = \mu_0$ versus H_a: $\mu \neq \mu_0$. Here Ω consists of all values of μ and σ for which $-\infty < \mu < \infty$ and $\sigma > 0$, and the likelihood function is

$$L(\mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-1/(2\sigma^2) \sum (x_i - \mu)^2}$$

In Section 7.2 we obtained the mles as $\hat{\mu}_{mle} = \bar{x}$, $\hat{\sigma}^2_{mle} = \sum (x_i - \bar{x})^2 / n$. Substituting these estimates back into the likelihood function gives

$$L(\hat{\mu}_{mle}, \hat{\sigma}_{mle}) = \cdots = \left(\frac{1}{2\pi \sum (x_i - \bar{x})^2 / n}\right)^{n/2} e^{-n/2}$$

Within Ω_0, μ in the foregoing likelihood is replaced by μ_0, so that only σ must be estimated. More precisely, the mle of μ subject to the constraint $\mu = \mu_0$ is trivially $\hat{\mu}_0 = \mu_0$. It is easily verified that the other mle under Ω_0 is $\hat{\sigma}^2_0 = \sum (x_i - \mu_0)^2 / n$. Substitution of this estimate in the likelihood function yields

$$L(\hat{\mu}_0, \hat{\sigma}_0) = \cdots = \left(\frac{1}{2\pi \sum (x_i - \mu_0)^2 / n}\right)^{n/2} e^{-n/2}$$

Thus we reject H_0 in favor of H_a when

$$\Lambda = \frac{L(\hat{\mu}_0, \hat{\sigma}_0)}{L(\hat{\mu}_{\mathrm{mle}}, \hat{\sigma}_{\mathrm{mle}})} = \cdots = \left(\frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \mu_0)^2}\right)^{n/2} \leq k$$

Raising both sides of this inequality to the power $2/n$, we reject H_0 whenever

$$\frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \mu_0)^2} \leq k^{2/n} = k'$$

This is intuitively reasonable: the value μ_0 is implausible for μ if the sum of squared deviations about the sample mean is much smaller than the sum of squared deviations about μ_0.

The denominator of this latter ratio can be expressed as

$$\sum [(x_i - \bar{x}) + (\bar{x} - \mu_0)]^2 = \sum (x_i - \bar{x})^2 + 2 \sum (\bar{x} - \mu_0)(x_i - \bar{x}) + n(\bar{x} - \mu_0)^2$$

The middle (i.e., cross-product) term in this expression is 0, because the constant $\bar{x} - \mu_0$ can be moved outside the summation, and then the sum of deviations from the sample mean is 0. Thus we should reject H_0 when

$$\frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2} = \frac{1}{1 + n(\bar{x} - \mu_0)^2 / \sum (x_i - \bar{x})^2} \leq k'$$

This latter ratio will be small when the second term in the denominator is large, so the condition for rejection becomes

$$\frac{n(\bar{x} - \mu_0)^2}{\sum (x_i - \bar{x})^2} \geq k''$$

Dividing both sides by $n - 1$ and taking square roots gives the rejection region

$$\text{either } \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \geq c \quad \text{or} \quad \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \leq -c$$

If we now let $c = t_{\alpha/2, n-1}$, we have exactly the two-tailed one-sample t test!

The bottom line is that when testing H_0: $\mu = \mu_0$ against the two-sided ($\neq$) alternative, the one-sample t test *is* the likelihood ratio test. This is also true of the upper-tailed version of the t test when the alternative is H_a: $\mu > \mu_0$ and of the lower-tailed test when the alternative is H_a: $\mu < \mu_0$. We could trace back through the argument to recover the critical constant k from c, but there is no point in doing this; the rejection region in terms of t is much more convenient than the rejection region in terms of the original likelihood ratio.   ∎

A number of tests discussed subsequently in this book, including the "pooled" t test from the next chapter and various tests from ANOVA (the analysis of variance) and regression analysis, can be derived by the likelihood ratio principle.

In many situations, the inequality for the rejection region of a likelihood ratio test cannot be manipulated to express the test procedure in terms of a simple statistic whose distribution can be ascertained. The following large-sample result, valid under fairly general conditions, can then be used.

THEOREM   If the sample size n is sufficiently large, then the statistic $-2\ln(\Lambda)$ has approximately a chi-squared distribution with v degrees of freedom when H_0 is true, where v is the difference between the number of "freely varying" parameters in Ω and the number of such parameters in Ω_0.

For example, if the distribution sampled is bivariate normal with the 5 parameters μ_1, μ_2 σ_1 σ_2, and ρ and the null hypothesis asserts that $\mu_1 = \mu_2$ and $\sigma_1 = \sigma_2$, then $v = 5 - 3 = 2$.

     By its definition $0 \le \Lambda \le 1$, and the likelihood ratio test rejects H_0 when this likelihood ratio is much less than 1. This is equivalent to rejecting H_0 when $-2\ln(\Lambda)$ is large and positive. The large-sample version of the test described in the theorem is thus upper-tailed: H_0 should be rejected if $-2\ln(\Lambda) \ge \chi^2_{\alpha,v}$, an upper-tail critical value extracted from Table A.5.

Example 9.28   Suppose a scientist makes n measurements of some physical characteristic, such as the specific gravity of a liquid. Let $X_1, \ldots, X_n$ denote the resulting measurement errors. Assume that these X_i's are independent and identically distributed according to the double exponential (Laplace) distribution: $f(x) = .5e^{-|x-\theta|}$ for $-\infty < x < \infty$. This pdf is symmetric about θ with somewhat heavier tails than the normal pdf. If $\theta = 0$ then the measurements are unbiased, so it is natural to test H_0: $\theta = 0$ versus H_a: $\theta \ne 0$. Here $v = 1 - 0 = 1$. The likelihood is

$$L(\theta) = (.5)^n e^{-\Sigma|x_i - \theta|}$$

Because of the minus sign preceding the summation, the likelihood is maximized when $\sum |x_i - \theta|$ is minimized. The absolute value function is not differentiable, and therefore differential calculus cannot be used. Instead, consider for a moment the case $n = 5$ and let $y_1 < \cdots < y_5$ denote the ordered values of the x_i's. For example, suppose a random sample of size 5 from the Laplace distribution with $\theta = 0$ is $-.24998, .75446, -.19053, 1.16237, .83229$, so $(y_1, \ldots, y_5) = (-.24998, -.19053, .75446, .83229, 1.16237)$. Then

$$\sum |x_i - \theta| = \sum |y_i - \theta| = \begin{cases} y_1 + y_2 + y_3 + y_4 + y_5 - 5\theta & \theta < y_1 \\ -y_1 + y_2 + y_3 + y_4 + y_5 - 3\theta & y_1 \le \theta < y_2 \\ -y_1 - y_2 + y_3 + y_4 + y_5 - \theta & y_2 \le \theta < y_3 \\ -y_1 - y_2 - y_3 + y_4 + y_5 + \theta & y_3 \le \theta < y_4 \\ -y_1 - y_2 - y_3 - y_4 + y_5 + 3\theta & y_4 \le \theta < y_5 \\ -y_1 - y_2 - y_3 - y_4 - y_5 + 5\theta & \theta \ge y_5 \end{cases}$$

The graph of this expression as a function of θ appears in Figure 9.12 (p. 551), from which it is apparent that the minimum occurs at $y_3 = \tilde{x} = .75446$, the sample median. (The situation is similar whenever n is odd. When n is even, the function achieves its minimum for any θ between $y_{n/2}$ and $y_{(n/2)+1}$; one such θ is $(y_{n/2} + y_{(n/2)+1})/2 = \tilde{x}$. In summary, the mle of θ is the sample median.)

     The likelihood ratio statistic for testing the relevant hypotheses is $\Lambda = (.5)^n e^{-\Sigma|x_i|}/[(.5)^n e^{-\Sigma|x_i-\tilde{x}|}]$. Simplifying and computing $-2\ln(\Lambda)$ gives the rejection region $2\sum|x_i| - 2\sum|x_i - \tilde{x}| \ge \chi^2_{\alpha,1}$ for the large-sample version of the likelihood ratio test.

     Suppose that a sample of $n = 30$ errors results in $\sum|x_i| = 38.6$ and $\sum|x_i - \tilde{x}| = 37.3$. Then

$$-2\ln(\Lambda) = 2\left(\sum|x_i| - \sum|x_i - \tilde{x}|\right) = 2.6$$
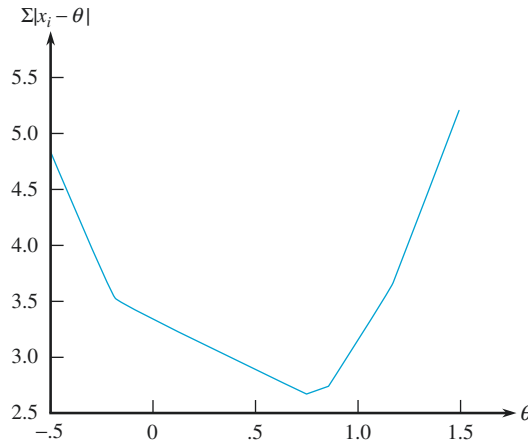
Figure 9.12 Determining the mle of the double exponential parameter by minimizing $\sum |x_i - \theta|$

Comparing this to $\chi^2_{.05,1} = 3.84$, we would not reject the null hypothesis at the 5% significance level. It is plausible that the measurement process indeed has mean/median 0, as desired. ■

Exercises: Section 9.5 (64–74)

64. For a random sample of n individuals taking a licensing exam, let $X_i = 1$ if the ith individual in the sample passes the exam and $X_i = 0$ otherwise ($i = 1, \ldots, n$).

    a. With p denoting the proportion of all exam-takers who pass, show that the most powerful test of H_0: $p = .5$ versus H_a: $p = .75$ rejects H_0 when $\sum x_i \geq c$.

    b. If $n = 20$ and you want $\alpha \leq .05$ for the test of (a), would you reject H_0 if 15 of the 20 individuals in the sample pass the exam?

    c. What is the power of the test you used in (b) when $p = .75$ [i.e., what is $\pi(.75)$]?

    d. Is the test derived in (a) UMP for testing the hypotheses H_0: $p = .5$ versus H_a: $p > .5$? Explain your reasoning.

    e. Graph the power function $\pi(p)$ of the test for the hypotheses of (d) when $n = 20$ and $\alpha \leq .05$.

    f. Return to the scenario of (a), and suppose the test is based on a sample size of 50. If the probability of a type II error is approximately .025, what is the

approximate significance level of the test (use a normal approximation)?

65. The error X in a measurement has a normal distribution with mean value 0 and variance σ^2. Consider testing H_0: $\sigma^2 = 2$ versus H_a: $\sigma^2 = 3$ based on a random sample $X_1, \ldots, X_n$ of errors.

    a. Show that a most powerful test rejects H_0 when $\sum x_i^2 \geq c$.

    b. For $n = 10$, find the value of c for the test in (a) that results in $\alpha = .05$.

    c. Is the test of (a) UMP for H_0: $\sigma^2 = 2$ versus H_a: $\sigma^2 > 2$? Justify your assertion.

66. Suppose that X, the fraction of a container that is filled, has pdf $f(x;\theta) = \theta x^{\theta-1}$ for $0 < x < 1$ (where $\theta > 0$), and let $X_1, \ldots, X_n$ be a random sample from this distribution.

    a. Show that the most powerful test for H_0: $\theta = 1$ versus H_a: $\theta = 2$ rejects the null hypothesis if $\sum \ln(x_i) \geq c$.

    b. Is the test of (a) UMP for testing H_0: $\theta = 1$ versus H_a: $\theta > 1$? Explain your reasoning.

c. If $n = 50$, what is the (approximate) value of c for which the test has significance level .05?

67. Consider a random sample of n component lifetimes, where the distribution of lifetime is exponential with parameter λ.

   a. Obtain a most powerful test for H_0: $\lambda = 1$ versus H_a: $\lambda = .5$, and express the rejection region in terms of a "simple" statistic.
   b. Is the test of (a) uniformly most powerful for H_0: $\lambda = 1$ versus H_a: $\lambda < 1$? Justify your answer.

68. Consider a random sample of size n from the "shifted exponential" distribution with pdf $f(x; \theta) = e^{-(x-\theta)}$ for $x > \theta$ and 0 otherwise (the graph is that of the ordinary exponential pdf with $\lambda = 1$ shifted so that it begins its descent at θ rather than at 0). Let Y_1 denote the smallest order statistic, and show that the likelihood ratio test of H_0: $\theta \leq 1$ versus H_a: $\theta > 1$ rejects the null hypothesis if y_1, the observed value of Y_1, is $\geq c$.

69. Suppose that each of n randomly selected individuals is classified according to his/her genotype with respect to a particular genetic characteristic and that the three possible genotypes are AA, Aa, and aa with long-run proportions (probabilities) θ^2, $2\theta(1 - \theta)$, and $(1 - \theta)^2$, respectively $(0 < \theta < 1)$. It is then straightforward to show that the likelihood is

$$\theta^{2x_1} \cdot [2\theta(1 - \theta)]^{x_2} \cdot (1 - \theta)^{2x_3}$$

where x_1, x_2, and x_3 are the number of individuals in the sample who have the AA, Aa, and aa genotypes, respectively. Show that the most powerful test for testing H_0: $\theta = .5$ versus H_a: $\theta = .8$ rejects the null hypothesis when $2x_1 + x_2 \geq c$. Is this test UMP for the alternative H_a: $\theta > .5$? Explain. [*Note*: The fact that the joint distribution of X_1, X_2, and X_3 is multinomial

can be used to obtain the value of c that yields a test with any desired significance level when n is large.]

70. The error in a measurement is normally distributed with mean μ and standard deviation 1. Consider a random sample of n errors, and show that the likelihood ratio test for H_0: $\mu = 0$ versus H_a: $\mu \neq 0$ rejects the null hypothesis when either $\bar{x} \geq c$ or $\bar{x} \leq -c$. What is c for a test with $\alpha = .05$? How does the test change if the standard deviation of an error is σ_0 (known) and the relevant hypotheses are H_0: $\mu = \mu_0$ versus H_a: $\mu \neq \mu_0$?

71. Measurement error in a particular situation is normally distributed with mean value μ and standard deviation 4. Consider testing H_0: $\mu = 0$ versus H_a: $\mu \neq 0$ based on a sample of $n = 16$ measurements.

   a. Verify that the usual test with significance level .05 rejects H_0 if either $\bar{x} \geq 1.96$ or $\bar{x} \leq -1.96$. [*Note*: That this test is unbiased follows from the fact that the way to capture the largest area under the z curve above an interval having width 3.92 is to center that interval at 0 (so it extends from -1.96 to 1.96).]
   b. Consider the test that rejects H_0 if either $\bar{x} \geq 2.17$ or $\bar{x} \leq -1.81$. What is α, that is, $\pi(0)$?
   c. What is the power of the test proposed in (b) when $\mu = .1$ and when $\mu = -.1$? (Note that .1 and $-.1$ are very close to the null value, so one would not expect large power for such values.) Is the test unbiased?
   d. Calculate the power of the usual test when $\mu = .1$ and when $\mu = -.1$. Is the usual test a most powerful test? [*Hint*: Refer to your calculations in (c).] [*Note*: It can be shown that the usual test is most powerful among all unbiased tests.]

72. A test of whether a coin is fair will be based on $n = 50$ tosses. Let X be the resulting number of heads. Consider two rejection regions: $R_1 = \{x: \text{either } x \leq 17 \text{ or } x \geq 33\}$ and $R_2 = \{x: \text{either } x \leq 18 \text{ or } x \geq 37\}$.

   a. Determine the significance level (type I error probability) for each rejection region.

   b. Determine the power of each test when $p = .49$. Is the test with rejection region R_1 a uniformly most powerful level .033 test? Explain.

   c. Is the test with rejection region R_2 unbiased? Explain.

   d. Sketch the power function for the test with rejection region R_1, and then do so for the test with the rejection region R_2. What does your intuition suggest about the desirability of using the rejection region R_2?

73. Reconsider the one-sample t test of Example 9.27.

   a. With $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$, show that the likelihood ratio is equal to $\Lambda = [1 + t^2/(n-1)]^{-n/2}$, and therefore the approximate chi-square statistic is $-2\ln(\Lambda) = n\ln[1 + t^2/(n-1)]$.

   b. Apply part (a) to test the hypotheses of Exercise 59, using the data given there. Compare your results with the answers found in Exercise 59.

74. The test statistic in the Neyman–Pearson Lemma and the likelihood ratio test statistic Λ are intimately related. Consider testing H_0: $\theta = \theta_0$ versus H_a: $\theta = \theta_a$, and let Λ^* denote the test statistic in (9.6). Show that

$$\Lambda = \begin{cases} 1/\Lambda^* & \text{if } L(\theta_0) \leq L(\theta_a) \\ 1 & \text{otherwise} \end{cases}$$

9.6   Further Aspects of Hypothesis Testing

We close this chapter by briefly considering several additional aspects of hypothesis testing, including the distinction between statistical significance (rejecting H_0 at a particular α) and the practical import of a departure from H_0, the relationship between tests and confidence intervals or bounds, and test procedures based on bootstrapping.

Statistical Versus Practical Significance

Although the process of reaching a decision by using the methodology of classical hypothesis testing involves selecting a level of significance and then rejecting or not rejecting H_0 at that level, simply reporting the α used and the decision reached conveys little of the information contained in the sample data. Especially when the results of an experiment are to be communicated to a large audience, rejection of H_0 at level .05 will be much more convincing if the observed value of the test statistic greatly exceeds the 5% critical value than if it barely exceeds that value. This is precisely what led to the notion of P-value as a way of reporting significance without imposing a particular α on others who might wish to draw their own conclusions. In fact, the editorial "Moving to a World Beyond '$p < 0.05$' " (*The American Statistician* 2019) calls for researchers to always report their actual P-values, rather than just whether hypotheses were rejected at the .05 level, and some research journals have begun adopting this policy.

Even if a P-value is included in a summary of results, however, there may be difficulty in interpreting this value and in making a decision. This is in part because a small P-value, which would ordinarily indicate *statistical significance* in that it would strongly suggest rejection of H_0 in favor of H_a, may be the result of a large sample size in combination with a departure from H_0 that has little *practical significance*. In many experimental situations, only departures from H_0 of large magnitude

would be worthy of detection, whereas a small departure from H_0 would have little practical importance. The editorial cited above also recommends the abolishment of the phrase "statistically significant" precisely because of this confusion.

Consider as an example testing $H_0: \mu = 100$ versus $H_a: \mu > 100$ where μ is the mean of a normal population with $\sigma = 10$. Suppose a true value of $\mu = 101$ would not represent a serious departure from H_0, in the sense that not rejecting H_0 when $\mu = 101$ would be a relatively inconsequential error; this would be the case, for example, if μ represented the average IQ score within some population. For a reasonably large sample size n this μ would lead to an $\bar{x}$ value near 101, so we would not want this sample evidence to argue strongly for rejection of H_0 when $\bar{x} = 101$ is observed. For various sample sizes, Table 9.1 records both the P-value when $\bar{x} = 101$ and also the probability of not rejecting H_0 at level .01 when $\mu = 101$.

Table 9.1 An illustration of the effect of sample size on P-values and β

n	P-value when $\bar{x} = 101$	$\beta(101)$ for level .01 test
25	.3085	.9664
100	.1587	.9082
400	.0228	.6293
900	.0013	.2514
1600	.0000335	.0475
2500	.000000297	.0038
10,000	7.69×10^{-24}	.0000

The second column in Table 9.1 shows that even for moderately large sample sizes, the P-value of $\bar{x} = 101$ argues very strongly for rejection of H_0 whereas the observed $\bar{x}$ itself suggests that in practical terms the true value of μ differs little from the null value $\mu_0 = 100$. The third column points out that even when there is little practical difference between the true μ (101) and the null value (100), for a fixed α a large sample size will almost always lead to rejection of the null hypothesis at that level. To summarize, one must be especially careful in interpreting evidence when the sample size is very large, since any small departure from H_0 will almost surely be detected by a test, yet such a departure may have little practical significance.

The Relationship Between Confidence Intervals and Hypothesis Tests

A confidence interval (Chapter 8) specifies a range of plausible values for an unknown population parameter. In contrast, the test procedures of this chapter focused on deciding whether a parameter equals a particular specified value. Not surprisingly, these two statistical inference methods are related and, in general, will yield consistent conclusions about a parameter when based on the same sample.

Consider again a hypothesis test for a population mean μ of the form $H_0: \mu = 100$ versus $H_a: \mu \neq 100$. Rather than following the techniques of this chapter, what if we constructed a confidence interval for μ instead? If 100 is within this confidence interval, then 100 is a plausible value of μ; hence, we should not reject the claim that μ equals 100 (i.e., don't reject H_0). Conversely, if 100 falls outside the confidence interval for μ, then 100 is *not* a plausible value for μ, and we should reject the hypothesis $\mu = 100$ in favor of the alternative $\mu \neq 100$. More generally, for $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$, *we reject H_0 if and only if μ_0 falls outside a confidence interval for μ.*

Two important mathematical connections need to be made here. First, in the preceding scenario, both the confidence interval and the alternative hypothesis were "two-sided." This is not coincidence: suppose instead that we wanted to decide between the claims H_0: $\mu = \mu_0$ and H_a: $\mu < \mu_0$. We would only decide in favor of H_a if the data provided convincing evidence that μ is lower than μ_0. This suggests computing an *upper confidence bound* for μ: if we can say with confidence that μ is at most some value B, and B is less than μ_0, then the data provides convincing evidence that μ is also less than μ_0. On the other hand, if μ_0 is less than B, then the confidence statement "$\mu < B$" doesn't tell us whether μ is lower than μ_0 or not. Hence, we would not be comfortable rejecting H_0: $\mu = \mu_0$ in favor of the alternative H_a: $\mu < \mu_0$. By the same reasoning, testing H_0: $\mu = \mu_0$ against the "upper-tailed" alternative H_a: $\mu > \mu_0$ is equivalent to computing a *lower* confidence bound for μ and observing whether μ_0 falls below that bound.

Second, any interval estimate carries with it an associated level of confidence (e.g., 95%), and every hypothesis test is carried out at a specified level of significance (e.g., 5%). *A hypothesis test at significance level α is equivalent to the appropriate confidence interval/bound at confidence level $100(1 - \alpha)\%$.* This should seem intuitively reasonable, but a mathematical demonstration can also be given (Exercise 80).

Example 9.29   Refer back to Example 9.12, in which data was provided on the d_{50} value (a measure of particulate matter size) for $n = 9$ roadside assays performed near Black Mountain, NC. Using the summary statistics $\bar{x} = 68.52$ microns, $s = 20.49$, and the t critical value $t_{.005,8} = 3.355$, a 99% CI for the true mean μ is

$$68.52 \pm 3.355 \cdot \frac{20.49}{\sqrt{9}} = 68.52 \pm 22.91 = (45.61, 91.43)$$

Because the interval does *not* include the value 44 microns, we can reject H_0: $\mu = 44$ in favor of H_a: $\mu \neq 44$ at the .01 level of significance. The significance level $\alpha = .01$ of the hypothesis test aligns with the selected confidence level: $99\% = 100(1 - .01)\%$.

If the researchers were instead interested in testing H_0: $\mu = 44$ versus H_a: $\mu > 44$, then a lower confidence bound for μ would be required, and H_0 would be rejected if that lower confidence bound exceeded 44 microns (since it would then follow that μ also exceeds 44).   ∎

Some caution must be taken when applying this notion of "duality" between intervals and tests to a population proportion p. This is because the standard deviation of $\hat{P}$ is estimated differently for confidence intervals and hypothesis tests: $\sqrt{\hat{p}\hat{q}/n}$ for the former, $\sqrt{p_0 q_0/n}$ for the latter. Hence, it is possible (though uncommon, especially for larger sample sizes) to get mutually contradictory conclusions about a hypothesized value p_0 when comparing a hypothesis test to the corresponding confidence interval.

General Large-Sample *z* Tests

The large-sample tests for p presented in Section 9.3 are a special case of more general large-sample procedures for a parameter θ. Let $\hat{\theta}$ be an estimator of θ that is at least approximately unbiased and has approximately a normal distribution. (Recall that, under very general conditions, maximum likelihood estimators have both of these properties.) The null hypothesis has the form H_0: $\theta = \theta_0$, where θ_0 denotes a number (the null value) appropriate to the problem context. A large-sample test statistic results from standardizing $\hat{\theta}$ under the assumption that H_0 is true [so that $E(\hat{\theta}) = \theta_0$]:

$$\text{Test statistic: } Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

If the alternative hypothesis is H_a: $\theta > \theta_0$, an upper-tailed test whose significance level is approximately α is specified by the rejection region $z \geq z_\alpha$. The other two alternatives, H_a: $\theta < \theta_0$ and H_a: $\theta \neq \theta_0$, are tested using a lower-tailed z test and a two-tailed z test, respectively.

In some cases, when H_0 is true the standard deviation of $\hat{\theta}$, $\sigma_{\hat{\theta}}$, involves no unknown parameters. For example, if $\theta = \mu$ and $\hat{\theta} = \bar{X}$, $\sigma_{\hat{\theta}} = \sigma_{\bar{X}} = \sigma/\sqrt{n}$, which involves no unknown parameters if the value of σ is known. In the case $\theta = p$, $\sigma_{\hat{\theta}} = \sigma_{\hat{P}} = \sqrt{p(1-p)/n}$, which involves the parameter of interest p itself. But $\sigma_{\hat{\theta}}$ does *not* involve any unknown parameters when H_0 is true, because we simply substitute $p = p_0$ into the standard error. When $\sigma_{\hat{\theta}}$ does involve unknown parameters, it is often possible to use an estimated standard deviation $S_{\hat{\theta}}$ in place of $\sigma_{\hat{\theta}}$ and still have Z approximately normally distributed when H_0 is true (because when n is large, $s_{\hat{\theta}} \approx \sigma_{\hat{\theta}}$ for most samples). The one-sample t test for large n furnishes an example of this: when σ is unknown, we use $S_{\hat{\theta}} = S_{\bar{X}} = S/\sqrt{n}$ in place of $\sigma/\sqrt{n}$ in the denominator of the test statistic (9.1), resulting in a t_{n-1}-distributed statistic. When n is large, the t_{n-1} and z distributions are virtually indistinguishable, and so the use of z-based rejection regions or P-values is not inappropriate in this situation.

Bootstrap Hypothesis Testing for μ

The bootstrap technique was introduced in Chapter 8 as a way of producing interval estimates for parameters without making additional assumptions about the population (e.g., normality). Analogous methodology exists for testing hypotheses about an unknown parameter (here, μ) when the one-sample t procedure described earlier is not applicable. Typically, this occurs when the sample size n is not large and the sample data is heavily skewed or otherwise indicate that population normality is not plausible.

The fundamental bootstrap concepts from Section 8.5 carry over to the hypothesis testing situation: first, a sample of data $x_1, \ldots, x_n$ is obtained. To approximate the sampling distribution of a statistic (here, $\bar{X}$), many resamples of size n are randomly selected with replacement from $x_1, \ldots, x_n$, and the statistic of interest is calculated for each resample. The distribution of those resample means $\bar{x}_1^*, \bar{x}_2^*, \ldots, \bar{x}_B^*$—the bootstrap distribution of $\bar{X}$—provides a reasonable approximation to the sampling distribution of $\bar{X}$. Inferences about the population mean μ can then be made.

Hypothesis testing introduces one wrinkle: we need information about the distribution of $\bar{X}$ *when the null hypothesis H_0: $\mu = \mu_0$ is true*. The linchpin of the basic bootstrap method is to treat the observed sample $x_1, \ldots, x_n$ as a population from which resamples will be drawn; however, this "population" does *not* have mean μ_0. The mean of the original sample is of course the observed sample mean $\bar{x}$, not μ_0. To address this issue, the sample data must be adjusted as follows: create new observations $w_1, \ldots, w_n$ by

$$w_i = x_i - \bar{x} + \mu_0, \quad i = 1, \ldots, n$$

This action simply relocates the original sample data in order to have mean μ_0; plots of the x_i's and the w_i's would be indistinguishable except for where they are centered. Now if we apply the basic bootstrap method to $w_1, \ldots, w_n$, the resulting resample means $\bar{w}_1^*, \bar{w}_2^*, \ldots, \bar{w}_B^*$ provide a semblance of what the distribution of $\bar{X}$ would look like if H_0 were true.

From this bootstrap distribution of $\bar{w}_i^*$'s, a **bootstrap P-value** can be obtained by determining what proportion of bootstrap means are at least as contradictory to H_0 as the observed value of the test statistic, $\bar{x}$. For example, if the alternative hypothesis is H_a: $\mu < \mu_0$, then the bootstrap P-value is the proportion of values among $\bar{w}_1^*, \bar{w}_2^*, \ldots, \bar{w}_B^*$ that are less than or equal to $\bar{x}$, the sample mean of the original data.

Example 9.30 As 3D printing increases in popularity, the accuracy and precision of 3D scanners have become ever more critical. The article "3D Scanning Automation for Die Casting Quality Control" (*Die Casting Engr.*, May, 2017: 16–18) describes a study in which a scanner was used on the same complicated object 12 times. For each run, the "flatness" (a sort of tolerance for surface smoothness) was recorded, resulting in the following measurements (microns):

$$23.50 \quad 22.73 \quad 23.63 \quad 23.50 \quad 23.16 \quad 23.61$$
$$23.54 \quad 22.64 \quad 23.55 \quad 23.41 \quad 23.49 \quad 23.18$$

Does the data provide convincing statistical evidence that the true mean flatness under these settings exceeds 23 microns? Let's test the hypotheses H_0: $\mu = 23$ versus H_a: $\mu > 23$. Figure 9.13 shows a normal probability plot of the data; its strongly nonlinear pattern indicates that the population distribution is very likely nonnormal. Since the sample size is small ($n = 12$), a one-sample t test would not be appropriate.
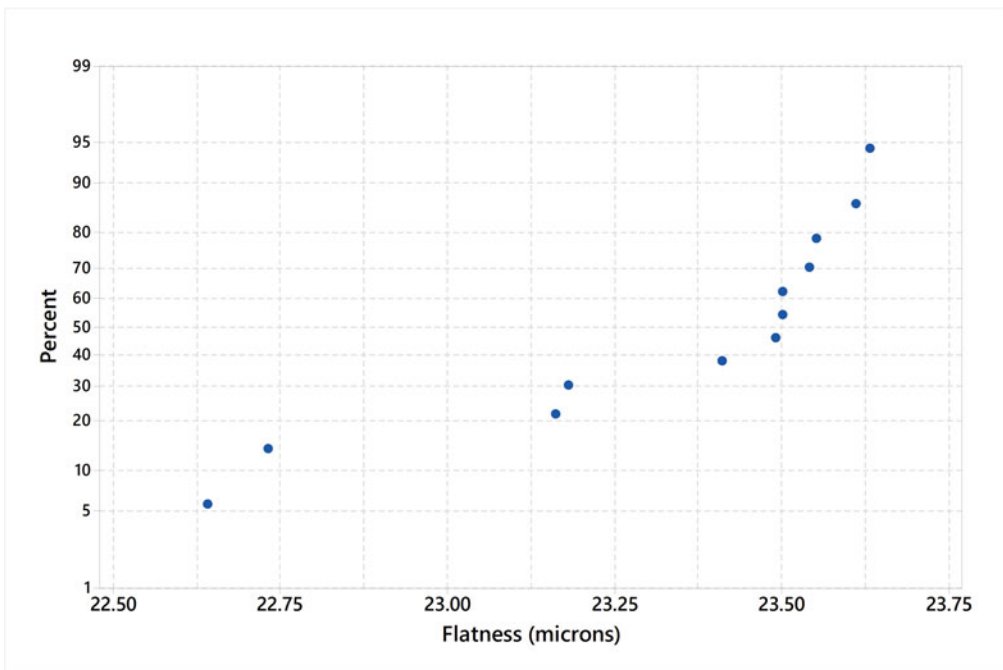
Figure 9.13 Normal probability plot of the flatness data in Example 9.30

Instead, we proceed with a bootstrap hypothesis test as described previously. The mean of the sample data is $\bar{x} = 23.3283$. An adjusted "population" $w_1, \ldots, w_{12}$ is created by subtracting $\bar{x}$ from each x_i and adding $\mu_0 = 23$:

$$w_1 = x_1 - \bar{x} + \mu_0 = 23.50 - 23.3283 + 23 = 23.1717, \quad w_2 = 22.73 - 23.3283 + 23 = 22.4017,$$

and so on. (A quick check confirms that the mean of the w_i's is $\mu_0 = 23$, as it should be.) Then bootstrap resampling is performed on the w_i's: take a sample of size 12 *with replacement* from $w_1, \ldots, w_{12}$, calculate the resample mean, and repeat. Figure 9.14 shows the result of $B = 10{,}000$ bootstrap resamples in R.
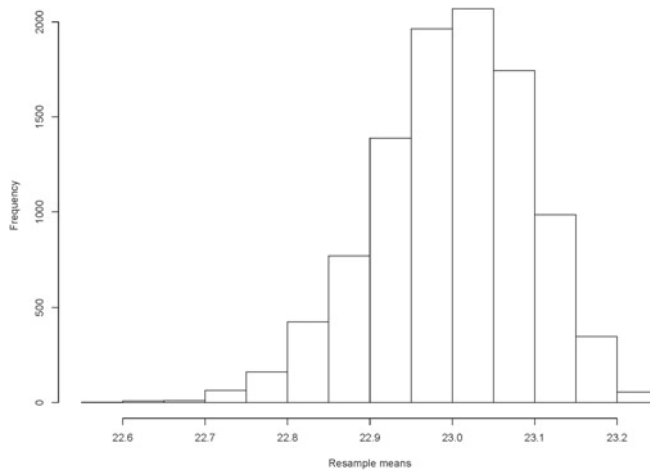
Figure 9.14   Bootstrap distribution for Example 9.30

The bootstrap distribution in Figure 9.14 is clearly skewed, validating the decision not to use a t test. The histogram in Figure 9.14 shows how the statistic $\bar{X}$ would be expected to behave across repeated samples of size $n = 12$ from the population *if* the null hypothesis is true and the population mean is 23. Because this is an upper-tailed test, the bootstrap P-value is the proportion of these bootstrap values that are at least as large as the real sample mean, $\bar{x} = 23.3283$. As is evident from the histogram, this is an extremely low probability—in fact, *zero* of the 10,000 resampled means were as large as $\bar{x}$. Thus our bootstrap P-value is 0, indicating that we should reject H_0 at any significance level. The data makes it clear that the population mean flatness of 3D scans under these settings is greater than 23 microns. ∎

Exercises: Section 9.6 (75–82)

75. Consider the large-sample level .01 test in Section 9.3 for testing H_0: $p = .2$ against H_a: $p > .2$.

    a. For the alternative value $p = .21$, compute $\beta(.21)$ for sample sizes $n = 100$, 2500, 10,000, 40,000, and 90,000.

    b. For $\hat{p} = x/n = .21$, determine the P-value when $n = 100$, 2500, 10,000, and 40,000.

    c. In most situations, would it be reasonable to use a level .01 test in conjunction with a sample size of 40,000? Why or why not?

76. Reconsider the paint-drying problem discussed in Example 9.25. The hypotheses were H_0: $\mu = 75$ versus H_a: $\mu < 75$, with σ assumed to have value 9. Consider the alternative value $\mu = 74$, which in the context of the problem would presumably not be a practically significant departure from H_0.

a. For a level .01 test, compute β at this alternative for sample sizes $n = 100$, 900, and 2500.

b. If the observed value of $\bar{X}$ is $\bar{x} = 74$, what can be said about the resulting P-value when $n = 2500$? Is the data statistically significant at any of the standard values of α?

c. Would you really want to use a sample size of 2500 along with a level .01 test (disregarding the cost of such an experiment)? Explain.

77. When $X_1, X_2, \ldots, X_n$ are independent $N(\mu, \sigma)$ variables and n is large, the sample variance S^2 has approximately a normal distribution with $E(S^2) = \sigma^2$ and $V(S^2) = 2\sigma^4/(n - 1)$.

a. Consider testing H_0: $\sigma = \sigma_0$. Use the mean and variance provided to construct a test statistic that has an approximately standard normal distribution when H_0 is true.

b. A manufacturer of exercise weights previously employed a process for which the standard deviation of the actual mass of its 10-lb. weights was .1 lb. After improving the process, the manufacturers wished to test H_0: $\sigma = .1$ versus H_a: $\sigma < .1$, where σ denotes the true standard deviation using the new process. A sample of 100 such weights has a sample standard deviation of .07 lb. Use this information and the test statistic in part (a) to determine whether H_0 should be rejected at the .05 level. [*Note:* Hypothesis testing for a population variance can also be based on the chi-squared distribution discussed in Section 8.4. See Exercises 98–99.]

78. When $X_1, X_2, \ldots, X_n$ are independent Poisson variables, each with parameter μ, and n is large, the sample mean $\bar{X}$ has approximately a normal distribution with $E(\bar{X}) = \mu$ and $V(\bar{X}) = \mu/n$. This implies that

$$Z = \frac{\bar{X} - \mu}{\sqrt{\mu/n}}$$

has approximately a standard normal distribution. For testing H_0: $\mu = \mu_0$, we can replace μ by μ_0 in the equation for Z to obtain a test statistic. This statistic is actually preferred to the one-sample t statistic with denominator $S/\sqrt{n}$ when the X_i's are Poisson because it is tailored explicitly to the Poisson assumption. If the number of requests for consulting received by a certain statistician during a 5-day work week has a Poisson distribution and the total number of consulting requests during a 36-week period is 160, does this suggest that the true average number of weekly requests exceeds 4.0? Test using $\alpha = .02$.

79. Consider the tip percentage data from Example 9.13.

a. Use the summary statistics $\bar{x} = 17.986$, $s = 5.937$, $n = 70$ and the t critical value $t_{.05,69} = 1.667$ to construct a 95% lower confidence bound for the population mean tip percentage μ.

b. Consider testing the hypotheses H_0: $\mu = 15$ versus H_a: $\mu > 15$. According to the bound in part (a), what is the rejection decision at the .05 level? Explain your reasoning.

c. Can the lower confidence bound in part (a) be used to test H_0: $\mu = 15$ versus H_a: $\mu \neq 15$ at the .05 level? Explain.

d. Return to the upper-tailed alternative H_a: $\mu > 15$. Does the lower confidence bound in part (a) prescribe a rejection decision at the .01 level? At the .10 level?

80. This exercise establishes the "duality" between confidence intervals/bounds and hypothesis tests for the one-sample t procedures. (Similar derivations apply to other inference methods.)

a. Consider the lower-tailed t test of $H_0: \mu = \mu_0$ versus $H_a: \mu < \mu_0$. Show that the test statistic $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$ falls in the level α rejection region if and only if μ_0 exceeds the one-sample t upper confidence bound for μ with confidence level $100(1 - \alpha)\%$.

b. Next, consider the upper-tailed alternative $H_a: \mu > \mu_0$. Show that the test statistic falls in the level α rejection region if and only if μ_0 is less than the lower $100(1 - \alpha)\%$ confidence bound for μ.

c. Finally, show the equivalency between the (two-sided) confidence interval for μ and the two-tailed one-sample t test of $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$.

81. Use the bootstrap hypothesis-testing method described in this section to test $H_0: \mu = 115$ versus $H_a: \mu < 115$ for the bagel data presented in Exercise 27.

82. Use the bootstrap hypothesis-testing method described in this section to test $H_0: \mu = 1.5$ versus $H_a: \mu \neq 1.5$ for the alcohol content data presented in Exercise 22.

Supplementary Exercises: (83–102)

83. When a drug is recalled for safety concerns (e.g., too many people having serious adverse reactions), the pharmaceutical company making the drug can only re-issue it by convincing the FDA that the reformulated version of the drug is safer than the original version.

a. In words, what are the null and alternative hypotheses for this situation? [*Hint:* the FDA will not allow re-issuance unless they see *convincing* evidence of a safety improvement.]

b. Describe the possible type I and type II errors in this scenario.

c. Which of the two possible errors is worse, and why? On that basis, how should the FDA determine the α level for testing whether the reformulated drug is safer?

84. A sample of 50 lenses used in eyeglasses yields a sample mean thickness of 3.05 mm and a sample standard deviation of .34 mm. The desired true average thickness of such lenses is 3.20 mm. Does the data strongly suggest that the true average thickness of such lenses is something other than what is desired? Test using $\alpha = .05$.

85. In the previous exercise, suppose the experimenter had believed before collecting the data that the value of σ was approximately .30. If the experimenter wished the probability of a type II error to be .05 when $\mu = 3.00$, was a sample size of 50 unnecessarily large?

86. It is specified that a certain type of iron should contain .85 g of silicon per 100 g of iron (.85%). The silicon content of each of 25 randomly selected iron specimens was determined, and the accompanying output resulted from a test of the appropriate hypotheses.

Variable	N	Mean	St Dev	SE Mean	T	P
sil cont	25	0.8880	0.1807	0.0361	1.05	0.30

a. What hypotheses were tested?

b. What conclusion would be reached for a significance level of .05, and why? Answer the same question for a significance level of .10.

87. A hot-tub manufacturer advertises that with its heating equipment, a temperature of 100 °F can be achieved in at most 15 min. A random sample of 32 tubs is selected, and the time necessary to achieve a 100 °F temperature is determined for each tub. The sample average time and sample standard deviation are 17.5 min and 2.2 min, respectively. Does this data cast doubt on the company's claim? Compute the P-value and use it to reach a conclusion at level .05 (assume that the heating-time distribution is approximately normal).

88. The true average breaking strength of ceramic insulators of a certain type is supposed to be at least 10 psi. They will be

used for a particular application unless sample data indicates conclusively that this specification has not been met. A test of hypotheses using $\alpha = .01$ is to be based on a random sample of ten insulators. Assume that the breaking-strength distribution is normal with unknown standard deviation. [*Note:* Software is required for this exercise.]

   a. If the true standard deviation is .80, how likely is it that insulators will be judged satisfactory when true average breaking strength is actually only 9.5? Only 9.0?
   b. What sample size would be necessary to have a 75% chance of detecting that H_0 is false when true average breaking strength is 9.5 when the true standard deviation is .80?

89. The article "Caffeine Knowledge, Attitudes, and Consumption in Adult Women" (*J. Nutrit. Ed.* 1992: 179–184) reports the following summary data on daily caffeine consumption for a sample of adult women: $n = 47$, $\bar{x} = 215$ mg, $s = 235$ mg, and range = 5 − 1176.

   a. Does it appear plausible that the population distribution of daily caffeine consumption is normal? Is it necessary to assume a normal population distribution to test hypotheses about the value of the population mean consumption? Explain your reasoning.
   b. Suppose it had previously been believed that mean consumption was at most 200 mg. Does the given data contradict this prior belief? Test the appropriate hypotheses at significance level .10 and include a *P*-value in your analysis.

90. The incidence of a certain type of chromosome defect in the U.S. adult male population is believed to be 1 in 75. A random sample of 800 individuals in U.S. penal institutions reveals 16 who have such defects. Can it be concluded that the incidence rate of this defect among prisoners differs from the presumed rate for the entire adult male population?

   a. State and test the relevant hypotheses using $\alpha = .05$. What type of error might you have made in reaching a conclusion?
   b. What *P*-value is associated with this test? Based on this *P*-value, could H_0 be rejected at significance level .20?

91. In an investigation of the toxin produced by a certain poisonous snake, a researcher prepared 26 different vials, each containing 1 g of the toxin, and then determined the amount of antitoxin needed to neutralize the toxin. The sample average amount of antitoxin necessary was found to be 1.89 mg, and the sample standard deviation was .42. Previous research had indicated that the true average neutralizing amount was 1.75 mg/g of toxin. Does the new data contradict the value suggested by prior research? Test the relevant hypotheses using the *P*-value approach. Does the validity of your analysis depend on any assumptions about the population distribution of neutralizing amount? Explain.

92. The sample average unrestrained compressive strength for 45 specimens of a particular type of brick was computed to be 3107 psi, and the sample standard deviation was 188. The distribution of unrestrained compressive strength may be somewhat skewed. Does the data strongly indicate that the true average unrestrained compressive strength is less than the design value of 3200? Test using $\alpha = .001$.

93. To test the ability of auto mechanics to identify simple engine problems, an automobile with a single such problem was taken in turn to 72 different car repair facilities. Only 42 of the 72 mechanics who worked on the car correctly identified the problem. Does this strongly indicate that the true proportion of mechanics who could identify this problem is less than .75? Compute the *P*-value and reach a conclusion accordingly.

94. Chapter 8 presented a CI for the variance σ^2 of a normal population distribution.

The key result there was that the rv $\chi^2 = (n-1)S^2/\sigma^2$ has a chi-squared distribution with $n - 1$ df. Consider the null hypothesis $H_0: \sigma^2 = \sigma_0^2$ (equivalently, $\sigma = \sigma_0$). Then when H_0 is true, the test statistic $\chi^2 = (n-1)S^2/\sigma_0^2$ has a chi-squared distribution with $n - 1$ df. If the relevant alternative is $H_a: \sigma^2 > \sigma_0^2$, rejecting H_0 if $\chi^2 \geq \chi_{\alpha,n-1}^2$ gives a test with significance level α. To ensure reasonably uniform characteristics for a particular application, it is desired that the true standard deviation of the softening point of a certain type of petroleum pitch be at most .50 °C. The softening points of ten different specimens were determined, yielding a sample standard deviation of .58 °C. Does this strongly contradict the uniformity specification? Test the appropriate hypotheses using $\alpha = .01$.

95. Referring to the previous exercise, suppose an investigator wishes to test $H_0: \sigma^2 = .04$ versus $H_a: \sigma^2 < .04$ based on $n = 21$ observations. The computed value of $20s^2/.04$ is 8.58. Place bounds on the P-value and then reach a conclusion at level .01.

96. When the population distribution is normal and n is large, the sample standard deviation S has approximately a normal distribution with $E(S) \approx \sigma$ and $V(S) \approx \sigma^2/(2n)$. We already know that in this case, for any n, $\bar{X}$ is normal with $E(\bar{X}) = \mu$ and $V(\bar{X}) = \sigma^2/n$.

a. Assuming that the underlying distribution is normal, what is an approximately unbiased estimator of the 99th percentile $\theta = \mu + 2.33\sigma$?

b. As discussed in Section 6.4, when the X_i's are normal $\bar{X}$ and S are independent rvs (one measures location whereas the other measures spread). Use this to compute $V(\hat{\theta})$ and $\sigma_{\hat{\theta}}$ for the estimator $\hat{\theta}$ of part (a). What is the estimated standard error $\hat{\sigma}_{\hat{\theta}}$?

c. Write a test statistic for testing $H_0: \theta = \theta_0$ that has approximately a standard normal distribution when H_0 is true. If soil pH is normally distributed in a certain region and 64 soil samples yield $\bar{x} = 6.33$, $s = .16$, does this provide strong evidence for concluding that at most 99% of all possible samples would have a pH of less than 6.75? Test using $\alpha = .01$.

97. Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with parameter λ. Then it can be shown that $2\lambda \sum X_i$ has a chi-squared distribution with $v = 2n$ (by first showing that $2\lambda X_i$ has a chi-squared distribution with $v = 2$).

a. Use this fact to obtain a test statistic and rejection region that together specify a level α test for $H_0: \mu = \mu_0$ versus each of the three commonly encountered alternatives. [*Hint*: $E(X_i) = \mu = 1/\lambda$, so $\mu = \mu_0$ is equivalent to $\lambda = 1/\mu_0$.]

b. Suppose that ten identical components, each having exponentially distributed time until failure, are tested. The resulting failure times are

95   16   11   3   42   71   225   64   87   123

Use the test procedure of part (a) to decide whether the data strongly suggests that the true average lifetime is less than the previously claimed value of 75.

98. Suppose the population distribution is normal with known σ. Let γ be such that $0 < \gamma < \alpha$. For testing $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$, consider the test that rejects H_0 if either $z \geq z_\gamma$ or $z \leq -z_{\alpha-\gamma}$, where the test statistic is $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$.

a. Show that $P(\text{type I error}) = \alpha$.

b. Derive an expression for $\beta(\mu')$. [*Hint*: Express the test in the form "reject H_0 if either $\bar{x} \geq c_1$ or $\leq c_2$."]

c. Let $\Delta > 0$. For what values of γ (relative to α) will $\beta(\mu_0 + \Delta) < \beta(\mu_0 - \Delta)$?

99. After a period of apprenticeship, an organization gives an exam that must be passed to be eligible for membership. Let $p = P$(randomly chosen apprentice passes). The organization wishes an exam that most but not all should be able to pass, so it decides that $p = .90$ is desirable. For a particular exam, the relevant hypotheses are H_0: $p = .90$ versus H_a: $p \neq .90$. Suppose ten people take the exam, and let $X =$ the number who pass.

   a. Does the lower-tailed region $\{0, 1, \ldots, 5\}$ specify a level .01 test?
   b. Show that even though H_a is two-sided, no two-tailed test is a level .01 test.
   c. Sketch a graph of power as a function of p' for this test. Is this desirable?

100. A service station has six gas pumps. When no vehicles are at the station, let p_i denote the probability that the next vehicle will select pump i ($i = 1, 2, \ldots, 6$). Based on a sample of size n, we wish to test H_0: $p_1 = \cdots = p_6$ versus the alternative H_a: $p_1 = p_3 = p_5$, $p_2 = p_4 = p_6$ (note that H_a is not a simple hypothesis). Let X be the number of customers in the sample that select an even-numbered pump.

   a. Show that the likelihood ratio test rejects H_0 if either $X \geq c$ or $X \leq n - c$. [*Hint*: When H_a is true, let θ denote the common value of p_2, p_4, and p_6.]
   b. Let $n = 10$ and $c = 9$. Determine the power of the test both when H_0 is true

and also when $p_2 = p_4 = p_6 = 1/10$, $p_1 = p_3 = p_5 = 7/30$.

101. Consider testing a pair of simple hypotheses H_0: $\theta = \theta_0$ versus H_a: $\theta = \theta_a$. Rather than prescribing the significance level and minimizing P(type II error), imagine trying to minimize the linear combination $a \cdot \alpha + b \cdot \beta$ for some specified constants $a > 0$ and $b > 0$. Show that $a \cdot \alpha + b \cdot \beta$ is minimized by using the rejection region

$$R^* = \left\{ (x_1, \ldots, x_n) : \frac{f(x_1, \ldots, x_n; \theta_a)}{f(x_1, \ldots, x_n; \theta_0)} \geq \frac{a}{b} \right\}$$

   [*Hint*: Imitate the first half of the proof of the Neyman-Pearson Lemma, but use $a \cdot \alpha + b \cdot \beta$ in place of $k\alpha + \beta$].

102. Refer back to the scenario introduced in Example 9.23, where H_0: $\mu = 1$ versus H_a: $\mu = 2$ was tested based on a sample from a Poisson(μ) distribution. Suppose committing a type II error is considered 3 times as problematic as a type I error, and so the manufacturers wish to minimize $\alpha + 3\beta$.

   a. Determine the test procedure that minimizes $\alpha + 3\beta$ when $n = 5$. [*Hint*: Refer back to the previous exercise.]
   b. For the test procedure in part (a), what are α, β, and the (minimized) value of $\alpha + 3\beta$?
   c. Repeat parts (a)–(b) for $n = 10$.

Inferences Based on Two Samples

10

Introduction

Chapters 8 and 9 presented confidence intervals (CIs) and hypothesis-testing procedures for single parameters, such as a population mean μ and a population proportion p. In this chapter, we extend these methods to situations involving the means, proportions, and variances of two different population distributions. For example, let μ_1 and μ_2 denote the true average decrease in cholesterol for two drugs. Then an investigator might wish to use results from patients assigned at random to two different groups as a basis for testing the null hypothesis $\mu_1 = \mu_2$ versus the alternative hypothesis $\mu_1 \neq \mu_2$. As another example, let p_1 denote the true proportion of all metal-on-metal hip replacements that fail, and let p_2 represent the true proportion of all ceramic-on-ceramic replacements that fail. Based on surveys of 500 people with each type of hip replacement, we might like an interval estimate for the difference $p_1 - p_2$.

10.1 The Two-Sample z Confidence Interval and Test

The inferences discussed in this section concern a difference $\mu_1 - \mu_2$ between the means of two different population distributions. An investigator might, for example, wish to test hypotheses about the difference between the true mean stopping distances of two different braking systems under identical conditions. One such hypothesis would state that $\mu_1 - \mu_2 = 0$, i.e., that $\mu_1 = \mu_2$. Alternatively, it may be appropriate to estimate $\mu_1 - \mu_2$ by computing a 95% CI. Such inferences would be based on a sample of stopping distances for each braking system.

ASSUMPTIONS
1. $X_1, X_2, \ldots, X_m$ is a random sample from a population with mean μ_1 and standard deviation σ_1.
2. $Y_1, Y_2, \ldots, Y_n$ is a random sample from a population with mean μ_2 and standard deviation σ_2.
3. The X and Y samples are independent of each other.

The natural estimator of $\mu_1 - \mu_2$ is $\overline{X} - \overline{Y}$, the difference between the corresponding sample means. The test statistic results from standardizing this estimator, so we need expressions for the expected value and standard deviation of $\overline{X} - \overline{Y}$.

PROPOSITION   The expected value of $\overline{X} - \overline{Y}$ is $\mu_1 - \mu_2$, so $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$. The standard deviation of $\overline{X} - \overline{Y}$ is

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

Proof Both these results depend on the rules of expected value and variance presented in Chapter 5. By linearity of expectation,

$$E(\overline{X} - \overline{Y}) = E(\overline{X}) - E(\overline{Y}) = \mu_1 - \mu_2$$

Because the X and Y samples are independent, $\overline{X}$ and $\overline{Y}$ are independent quantities, so the variance of the difference is the *sum* of $V(\overline{X})$ and $V(\overline{Y})$:

$$V(\overline{X} - \overline{Y}) = V(\overline{X}) + V(\overline{Y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

The standard deviation of $\overline{X} - \overline{Y}$ is the square root of this expression. ∎

Regarding $\mu_1 - \mu_2$ as a parameter θ, its estimator is $\hat{\theta} = \overline{X} - \overline{Y}$ with standard deviation $\sigma_{\hat{\theta}}$ given by the proposition. When σ_1 and σ_2 both have known values, the test statistic will have the form $(\hat{\theta} - \text{null value})/\sigma_{\hat{\theta}}$; this form of a test statistic was used in several one-sample problems in the previous chapter. If σ_1 and σ_2 are unknown, the sample standard deviations must be used to estimate $\sigma_{\hat{\theta}}$ (the topic of Section 10.2).

Confidence Interval for $\mu_1 - \mu_2$ With Known σ's

In Chapters 8 and 9, the first CI and test procedure for a population mean μ were based on the assumption that the population distribution was normal with the value of the population standard deviation σ known to the investigator. Similarly, we first assume here that *both* population distributions are normal and that the values of *both* σ_1 and σ_2 are known.

Because the population distributions are normal, both $\overline{X}$ and $\overline{Y}$ have normal distributions. This implies that $\overline{X} - \overline{Y}$ is normally distributed, with expected value $\mu_1 - \mu_2$ and standard deviation $\sigma_{\overline{X}-\overline{Y}}$ given in the foregoing proposition. Standardizing $\overline{X} - \overline{Y}$ gives the standard normal variable

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}} \tag{10.1}$$

Since the area under the z curve between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $1 - \alpha$, it follows that

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}} < z_{\alpha/2} \right) = 1 - \alpha$$

Manipulation of the inequalities inside the parentheses to isolate $\mu_1 - \mu_2$ yields the equivalent probability statement

$$P\left(\overline{X} - \overline{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}\right) = 1 - \alpha$$

This implies that a $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$ has lower limit $(\bar{x} - \bar{y}) - z_{\alpha/2} \cdot \sigma_{\overline{X}-\overline{Y}}$ and upper limit $(\bar{x} - \bar{y}) + z_{\alpha/2} \cdot \sigma_{\overline{X}-\overline{Y}}$. This interval is a special case of the general formula $\hat{\theta} \pm z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$.

If both m and n are large, the CLT implies that both $\overline{X}$ and $\overline{Y}$ are approximately normal. In that case, this interval remains valid with a confidence level of *approximately* $100(1 - \alpha)\%$ irrespective of the population distributions.

TWO-SAMPLE
z INTERVAL

Assuming independent random samples from normal population distributions, a CI for $\mu_1 - \mu_2$ with a confidence level of $100(1 - \alpha)\%$ has endpoints

$$\bar{x} - \bar{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

An upper or lower confidence bound can also be calculated by retaining the appropriate sign (+ or −) and replacing $z_{\alpha/2}$ by z_α.

These confidence limits may also be applied to samples from nonnormal populations, provided that both sample sizes are large (say, $m > 40$ and $n > 40$); the confidence level is then approximate.

In practice, the assumption of known σ's is generally unrealistic. If information was available concerning the population standard deviations, typically both μ_1 and μ_2 would also be known. In Section 10.2, we will examine the more realistic scenario when the values of all four parameters are unknown.

Example 10.1 The article "Reflective Tape Applied to Bicycle Frame and Conspicuity Enhancement at Night" (*Hum. Factors* 2017: 485–500) describes a series of studies to determine the distance at which drivers can see a bicyclist ahead in the road ("detection distance") at night. One study compared detection distance when the bicycle had a typical red reflector mounted on the rear of the bike versus having reflective tape wrapped around the posterior forks, seat post, and rear reflector panel. The sample mean detection distances under these two conditions were $\bar{x} = 67.66$ m and $\bar{y} = 168.28$ m, respectively.

Suppose these observations were based on independent random samples of $m = n = 64$ drivers, and that the population standard deviations under these two conditions are $\sigma_1 = 30$ m and $\sigma_2 = 40$ m (values consistent with information in the article). Then a 95% CI for $\mu_1 - \mu_2$, the true difference in mean detection distance under these two settings, is

$$(67.66 - 168.28) \pm 1.96\sqrt{\frac{30^2}{64} + \frac{40^2}{64}} = -100.62 \pm 1.96(6.25) = (-112.87, -88.37)$$

Note that the confidence level is approximate, because with large sample sizes but no assumption of normality we are relying on the CLT. The interval indicates that average nighttime detection distance is between 88.37 and 112.87 m greater (that is, better) for bikes using reflective tape versus those just relying on the standard red rear reflector. ∎

If the standard deviations σ_1 and σ_2 are known and the investigator uses equal sample sizes, then the sample size $m = n$ for each sample that yields a $100(1 - \alpha)\%$ interval of width w is

$$m = n = \frac{z_{\alpha/2}^2(\sigma_1^2 + \sigma_2^2)}{(w/2)^2}$$

which will generally have to be rounded up to an integer. (Recall that $w/2$ represents the desired bound on the interval's *margin of error*. The sample size formula results from setting $w/2 = z_{\alpha/2}\sqrt{\sigma_1^2/n + \sigma_2^2/n}$ and solving for n.)

Test Procedures for $\mu_1 - \mu_2$ with Known σ's

In a hypothesis-testing problem, the null hypothesis will state that $\mu_1 - \mu_2$ has a specified value. Denoting this null value by Δ_0, the null hypothesis becomes H_0: $\mu_1 - \mu_2 = \Delta_0$. For example, the null hypothesis might state that the difference in true average fuel efficiencies (mpg) between cars having a turbocharged engine and a nonturbo engine is -3 (that is, on average turbocharging decreases fuel efficiency by 3 mpg). The null value would be $\Delta_0 = +3$ if the subscripts 1 and 2 instead referred to nonturbo and turbo engines, in that order. Often $\Delta_0 = 0$, in which case H_0 is equivalent to asserting that $\mu_1 = \mu_2$. A test statistic results from replacing $\mu_1 - \mu_2$ in Expression (10.1) by the null value Δ_0. Because the test statistic Z is obtained by standardizing $\overline{X} - \overline{Y}$ under the assumption that H_0 is true, it has a standard normal distribution in this case.

Consider the alternative hypothesis H_a: $\mu_1 - \mu_2 > \Delta_0$. A value $\bar{x} - \bar{y}$ that considerably exceeds Δ_0 (the expected value of $\overline{X} - \overline{Y}$ when H_0 is true) provides evidence against H_0 and for H_a. Such a value of $\bar{x} - \bar{y}$ corresponds to a positive and large value of z. Thus H_0 should be rejected in favor of H_a if z is greater than or equal to an appropriately chosen critical value. Because the test statistic Z has a standard normal distribution when H_0 is true, the upper-tailed rejection region $z \geq z_\alpha$ gives a test with significance level (type I error probability) α. Rejection regions for the other two alternatives H_a: $\mu_1 - \mu_2 < \Delta_0$ and H_a: $\mu_1 - \mu_2 \neq \Delta_0$ that yield tests with desired significance level α are lower-tailed and two-tailed, respectively.

As in the confidence interval discussion, z-based inference is still approximately correct here for samples from nonnormal populations provided both m and n are large. (Note, though, that here we still assume known population standard deviations.)

TWO-SAMPLE z TEST

Null hypothesis: H_0: $\mu_1 - \mu_2 = \Delta_0$

Test statistic value: $z = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}}$

Alternative Hypothesis	Rejection Region for Level α Test
H_a: $\mu_1 - \mu_2 > \Delta_0$	$z \geq z_\alpha$ (upper-tailed test)
H_a: $\mu_1 - \mu_2 < \Delta_0$	$z \leq -z_\alpha$ (lower-tailed test)
H_a: $\mu_1 - \mu_2 \neq \Delta_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)

Because these are z tests, a P-value is computed as it was for the z tests in Chapter 9: P-value = $1 - \Phi(z)$ for an upper-tailed test, $= \Phi(z)$ for a lower-tailed test, and $= 2[1 - \Phi(|z|)]$ for a two-tailed test.

These test procedures may also be applied to samples from nonnormal populations, provided that both sample sizes are large (say, $m > 40$ and $n > 40$).

Example 10.2 Each student in a class of 21 responded to a questionnaire that requested their grade point average (GPA) and the number of hours they studied each week. For those who studied less than 10 h/week the GPAs were

| 2.80 | 3.40 | 4.00 | 3.60 | 2.00 | 3.00 | 3.47 | 2.80 | 2.60 | 2.00 |

and for those who studied at least 10 h/week the GPAs were

| 3.00 | 3.00 | 2.20 | 2.40 | 4.00 | 2.96 | 3.41 | 3.27 | 3.80 | 3.10 | 2.50 |

Normal probability plots for both sets are reasonably linear, so the normality assumption is tenable. Because the standard deviation of GPAs for the whole campus is $\sigma = .6$, it is reasonable to apply that value here to both (conceptual) populations. The sample means are 2.97 for the <10 study hours group and 3.06 for the ≥ 10 study hours group. Treating the two samples as random, is there evidence that true average GPA is higher for students who study more? Let's carry out a test of significance at level .05 using the seven-step procedure outline in Section 9.2.

1. Parameter: $\mu_1 - \mu_2$, the difference between true mean GPA for the (conceptual) <10 population and true mean GPA for the ≥ 10 population
2. Hypotheses:

$$H_0: \mu_1 - \mu_2 = 0 \text{ (i.e., } \mu_1 = \mu_2)$$

$$H_a: \mu_1 - \mu_2 < 0 \text{ (i.e., } \mu_1 < \mu_2)$$

3. Assumptions/requirements: We have assumed underlying normal distributions for the GPAs of both populations, each with a known population standard deviation.
4. Test statistic value: With $\Delta_0 = 0$, the test statistic value is

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}}$$

5. Rejection region: The inequality in H_a implies that the test is lower-tailed. For $\alpha = .05$, $z_\alpha = z_{.05} = 1.645$. H_0 will be rejected if $z \leq -1.645$.
6. Substituting $m = 10, \bar{x} = 2.97, \sigma_1 = .6, n = 11, \bar{y} = 3.06$, and $\sigma_2 = .6$ into the formula for z yields

$$z = \frac{2.97 - 3.06}{\sqrt{\dfrac{.6^2}{10} + \dfrac{.6^2}{11}}} = \frac{-.09}{.262} = -.34$$

That is, the value of $\bar{x} - \bar{y}$ is only one-third of a standard deviation below what would be expected when H_0 is true.
7. Because the value of z is not even close to the rejection region, there is no reason to reject the null hypothesis. This test does not provide convincing statistical evidence that students who study ≥ 10 h per week have a higher mean GPA than those studying <10 h per week. ■

Power, β, and Sample Size Determination
Both power and β (the probability of a type II error) are easily calculated when the population distributions are normal with known values of σ_1 and σ_2. Consider the case in which the alternative

hypothesis is $H_a: \mu_1 - \mu_2 > \Delta_0$. Let Δ' denote a value of $\mu_1 - \mu_2$ that exceeds Δ_0, a value for which H_0 is false and H_a is true. The upper-tailed rejection region $z \geq z_\alpha$ can be re-expressed in the form $\bar{x} - \bar{y} \geq \Delta_0 + z_\alpha \sigma_{\overline{X}-\overline{Y}}$. Thus the probability of a type II error when $\mu_1 - \mu_2 = \Delta'$ is

$$\begin{aligned}\beta(\Delta') &= P(\text{not rejecting } H_0 \text{ when } \mu_1 - \mu_2 = \Delta') \\ &= P(\overline{X} - \overline{Y} < \Delta_0 + z_\alpha \sigma_{\overline{X}-\overline{Y}} \text{ when } \mu_1 - \mu_2 = \Delta')\end{aligned}$$

When $\mu_1 - \mu_2 = \Delta'$, $\overline{X} - \overline{Y}$ is normally distributed with mean value Δ' and standard deviation $\sigma_{\overline{X}-\overline{Y}}$ (the same standard deviation as when H_0 is true); using these values to standardize the inequality in parentheses gives β.

Alternative Hypothesis $\boldsymbol{\beta(\Delta') = P(\text{type II error when } \mu_1 - \mu_2 = \Delta')}$

$H_a: \mu_1 - \mu_2 > \Delta_0$ $\Phi\left(z_\alpha - \dfrac{\Delta' - \Delta_0}{\sigma_{\overline{X}-\overline{Y}}}\right)$

$H_a: \mu_1 - \mu_2 < \Delta_0$ $1 - \Phi\left(-z_\alpha - \dfrac{\Delta' - \Delta_0}{\sigma_{\overline{X}-\overline{Y}}}\right)$

$H_a: \mu_1 - \mu_2 \neq \Delta_0$ $\Phi\left(z_{\alpha/2} - \dfrac{\Delta' - \Delta_0}{\sigma_{\overline{X}-\overline{Y}}}\right) - \Phi\left(-z_{\alpha/2} - \dfrac{\Delta' - \Delta_0}{\sigma_{\overline{X}-\overline{Y}}}\right)$

where $\sigma_{\overline{X}-\overline{Y}} = \sqrt{(\sigma_1^2/m) + (\sigma_2^2/n)}$. For each case, power $= 1 - \beta(\Delta')$.

Example 10.3 (Example 10.2 continued) If $\mu_1 = \mu_2 - .5$ (true average GPA is .5 lower for the less-studious group), what is the probability of detecting such a departure from H_0 based on a level .05 test with sample sizes $m = 10$ and $n = 11$? The value of $\sigma_{\overline{X}-\overline{Y}}$ for these sample sizes (the denominator of z) was previously calculated as .262. The probability of a type II error for the lower-tailed level .05 test when $\mu_1 - \mu_2 = \Delta' = -.5$ is

$$\beta(-.5) = 1 - \Phi\left(-1.645 - \frac{-.5 - 0}{.262}\right) = 1 - \Phi(0.263) = .396$$

Thus the probability of detecting such a departure is power $= 1 - \beta(-.5) = .604$. Clearly, we have a mediocre chance of detecting a difference of $-.5$ with these sample sizes. Perhaps we should not conclude from Example 10.2 that there is no relationship between study time and GPA, because the sample sizes were insufficient.   ∎

As in Chapter 9, sample sizes m and n can be determined that will satisfy both $P(\text{type I error}) = $ a specified α and $P(\text{type II error when } \mu_1 - \mu_2 = \Delta') = $ a specified β. For an upper-tailed test, equating the previous expression for $\beta(\Delta')$ to the specified value of β gives

$$\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} = \frac{(\Delta' - \Delta_0)^2}{(z_\alpha + z_\beta)^2}$$

When the two sample sizes are equal, this equation yields

$$m = n = \frac{(\sigma_1^2 + \sigma_2^2)(z_\alpha + z_\beta)^2}{(\Delta' - \Delta_0)^2}$$

This expression is also correct for a lower-tailed test, whereas α is replaced by $\alpha/2$ for a two-tailed test.

Using a Comparison to Identify Causality

Investigators are often interested in comparing either the effects of two different treatments on some outcome, or the response after treatment with the response after no treatment (treatment vs. control). If the individuals or objects to be used in the comparison are not assigned by the investigators to the two different conditions, the study is said to be **observational**. The difficulty with drawing conclusions based on an observational study is that although statistical analysis may indicate a significant difference in response between the two groups, the difference may be due to some underlying factors that had not been controlled, rather than to any difference in the effects of the treatments.

Example 10.4 Many investigations in the last several years have explored the potential benefits of consuming a moderate amount of alcohol. As reported by CNN (July 27, 2017), a Danish study found that light-to-moderate drinkers (those consuming a few glasses of wine 3–4 days per week) had a lower risk of developing diabetes than those who rarely consumed alcohol. The study was based on tracking more than 70,000 Danes over a five-year period, and a test of the difference between the new diabetes development rates between these two groups resulted in an extremely low P-value. The difference was also considered clinically meaningful; that is, the low P-value was not simply the result of the very large sample size.

Should we conclude that moderate alcohol consumption *causes* a decreased likelihood of diabetes? Should health professionals recommend a few glasses of wine per day to help prevent diabetes onset? Not necessarily: since individuals in the study decided for themselves how much to drink, there could be some other underlying factor that the moderate drinkers have in common that would explain their lower risk of diabetes. For instance, most drinkers in the study specifically consumed wine, and wine-drinkers tend to be wealthier. Perhaps other lifestyle features of those wealthier individuals can help explained the observed relationship. (Using advanced statistical methods, the researchers "adjusted" for several factors including age, diet, and education, but they can't account for every possible alternative explanation.)   ∎

Once upon a time, it was argued that the studies linking smoking and lung cancer were all observational, and therefore that nothing had been proved. This was the view of the great statistician R. A. Fisher, who maintained till his death in 1962 that the observational studies did not show causation. He said that people who choose to smoke might be more susceptible to lung cancer. This explanation for the relationship had plenty of opposition then, and few would support it now. At that time few women got lung cancer because few women smoked, but when smoking increased among women, so did lung cancer. Furthermore, the incidence of lung cancer was higher for those who smoked more, and quitters had reduced incidence. Eventually, the physiological effects on the body were better understood, and nonobservational animal studies made it clear that smoking does, in fact, cause lung cancer.

To establish causation through a statistical study, we must try to eliminate the possibility that the groups being compared (e.g., drinkers and nondrinkers) have some other distinguishing feature (e.g., wealth) that could explain the study results. A **randomized controlled experiment** results when investigators *assign* subjects to the two treatments in a random fashion. When statistical significance is observed in such an experiment, the investigator and other interested parties will have more confidence in the conclusion that the difference in response has been caused by a difference in treatments.

Example 10.5 Many advertisers have touted "green labeling"—discussing the environmental impact of a product in advertisements (e.g., "now with less phosphates")—as a way to increase sales. But is this really effective? The article "How Green Should You Be: Can Environmental Associations Enhance Brand Performance?" (*J. Mark. Res.* 2008: 547–563) discussed a randomized controlled experiment in which shoppers were shown one of three brochures describing a (made-up) brand of detergent: one brochure that provided generic information about the brand, one that included information on the environmental performance of the brand, and one that additionally included an "environmental certification" label. Brochure types were *randomly assigned* to the study participants. The authors of the article then assessed participants' attitude toward the brand, including the likelihood of purchasing that detergent.

What the researchers found contradicted conventional wisdom: shoppers who saw the "green" advertisements were no more positively disposed to the product than those who had seen the generic advertisement. In other words, the presence of environmental information did *not* cause people to be more apt to purchase that brand of detergent. (To be more precise, the experiment uncovered no statistically significant differences in customer attitudes between the three brochures.) ∎

Observational studies, experiments, and the issue of establishing causality are discussed at greater length in the (nonmathematical) books by Utts, Moore, and Freedman et al., listed in the bibliography.

Exercises: Section 10.1 (1–12)

1. An article in *Consumer Reports* compared various types of batteries. The average lifetimes of Duracell AA batteries and Energizer AA batteries were given as 4.1 h and 4.5 h, respectively. Suppose these are the population average lifetimes.

   a. Let $\overline{X}$ be the sample average lifetime of 100 Duracell batteries and $\overline{Y}$ be the sample average lifetime of 100 Energizer batteries. What is the mean value of $\overline{X} - \overline{Y}$ (i.e., where is the distribution of $\overline{X} - \overline{Y}$ centered)? How does your answer depend on the specified sample sizes?

   b. Suppose the population standard deviations of lifetime are 1.8 h for Duracell batteries and 2.0 h for Eveready batteries. With the sample sizes given in part (a), what is the variance of the statistic $\overline{X} - \overline{Y}$, and what is its standard deviation?

   c. For the sample sizes given in part (a), draw a picture of the approximate distribution curve of $\overline{X} - \overline{Y}$ (include a measurement scale on the horizontal axis). Would the shape of the curve necessarily be the same for sample sizes of 10 batteries of each type? Explain.

2. According to a 2018 report by the CDC, the mean body mass index (BMI) for American adult men is 29.1 kg/m^2, while the mean for women is 29.6 kg/m^2. Suppose these are population averages.

   a. Let $\overline{X}$ be the sample average BMI of 50 randomly selected American adult men $\overline{Y}$ be the sample average BMI of 75 randomly selected American adult women. What is the expected value of $\overline{X} - \overline{Y}$? How does your answer depend on the specified sample sizes?

   b. Suppose the population standard deviations of BMI are 4.7 for men and 6.2 for women (these values are consistent with the study). With the sample sizes given in part (a), what is the variance of the statistic $\overline{X} - \overline{Y}$, and what is its standard deviation?

   c. For the sample sizes given in part (a), what is the approximate distribution of $\overline{X} - \overline{Y}$, and why?

d. Would the shape of the distribution in part (c) necessarily be the same for sample sizes of 5 men and 7 women? Explain.

3. Let μ_1 and μ_2 denote true average tread lives (miles) for two competing brands of size P205/65R15 tires.

   a. Test $H_0: \mu_1 - \mu_2 = 0$ versus $H_a: \mu_1 - \mu_2 \neq 0$ at level .05 using the following information: $m = 45$, $\bar{x} = 42{,}500$, $\sigma_1 = 2200$, $n = 45$, $\bar{y} = 40{,}400$, and $\sigma_2 = 1900$.

   b. Use the information in part (a) to compute a 95% CI for $\mu_1 - \mu_2$. Does the resulting interval suggest that $\mu_1 - \mu_2$ has been precisely estimated?

4. Let μ_1 denote true average tread life for a premium brand of P205/65R15 tire and let μ_2 denote the true average tread life for an economy brand of the same size.

   a. Test $H_0: \mu_1 - \mu_2 = 5000$ versus the alternative $H_a: \mu_1 - \mu_2 > 5000$ at level .01 using the following information: $m = 45$, $\bar{x} = 42{,}500$, $\sigma_1 = 2200$, $n = 45$, $\bar{y} = 36{,}800$, and $\sigma_2 = 1500$.

   b. Use the information in part (a) to compute a 99% lower confidence bound for $\mu_1 - \mu_2$. Is your answer consistent with the test in part (a)?

5. Persons having Raynaud's syndrome are apt to suffer a sudden impairment of blood circulation in fingers and toes. In an experiment to study the extent of this impairment, each subject immersed a forefinger in water and the resulting heat output (cal/cm^2/min) was measured. For $m = 10$ subjects with the syndrome, the average heat output was $\bar{x} = .64$, and for $n = 10$ nonsufferers, the average output was 2.05. Let μ_1 and μ_2 denote the true average heat outputs for the two types of subjects. Assume that the two distributions of heat output are normal with $\sigma_1 = .2$ and $\sigma_2 = .4$.

   a. Consider testing $H_0: \mu_1 - \mu_2 = -1.0$ versus $H_a: \mu_1 - \mu_2 < -1.0$ at level .01.

Describe in words what H_a says, and then carry out the test.

   b. Compute the P-value for the value of Z obtained in part (a).

   c. What is the probability of a type II error when the actual difference between μ_1 and μ_2 is $\mu_1 - \mu_2 = -1.2$? What is the power?

   d. Assuming that $m = n$, what sample sizes are required to ensure that $\beta = .1$ when $\mu_1 - \mu_2 = -1.2$?

6. An experiment to compare the tension bond strength of polymer latex modified mortar to that of unmodified mortar resulted in $\bar{x} = 18.12$ kgf/cm^2 for the modified mortar ($m = 40$) and $\bar{y} = 16.87$ kgf/cm^2 for the unmodified mortar ($n = 32$). Let μ_1 and μ_2 be the true average tension bond strengths for the modified and unmodified mortars, respectively. Assume that the bond strength distributions are both normal.

   a. Assuming that $\sigma_1 = 1.6$ and $\sigma_2 = 1.4$, test $H_0: \mu_1 - \mu_2 = 0$ versus $H_a: \mu_1 - \mu_2 > 0$ at level .01.

   b. Compute the probability of a type II error for the test of part (a) when $\mu_1 - \mu_2 = 1$.

   c. Suppose the investigator decided to use a level .05 test and wished $\beta = .10$ when $\mu_1 - \mu_2 = 1$. If $m = 40$, what value of n is necessary?

7. What affects the time a consumer spends looking at a product on the shelf prior to selection? The following data summarized elapsed time (in seconds) for purchasers of fabric softener and washing-up liquid; the former is much more expensive than the latter. These products were chosen because they're similar with respect to shelf space and number of brands available.

Product	Sample size	Sample mean
Fabric softener	15	30.42
Washing-up liquid	19	26.53

a. What assumptions, if any, are necessary for the inferential procedures of this section to be valid in this situation? Why?

b. Assuming that $\sigma_1 = \sigma_2 = 8.5$ s, test to see if there is a significant difference in the true average time purchasers spend looking at these two products, at the $\alpha = .01$ significance level.

8. An experiment was performed to compare the fracture toughness of high-purity nickel maraging steel with commercial-purity steel of the same type. The sample average toughness was $\bar{x} = 65.6$ for $m = 32$ specimens of the high-purity steel, whereas for $n = 38$ specimens of commercial steel $\bar{y} = 59.8$. Because the high-purity steel is more expensive, its use for a certain application can be justified only if its fracture toughness exceeds that of commercial-purity steel by more than 5. Suppose that both toughness distributions are normal.

a. Assuming that $\sigma_1 = 1.2$ and $\sigma_2 = 1.1$, test the relevant hypotheses using $\alpha = .001$.

b. Compute β and power for the test conducted in part (a) when $\mu_1 - \mu_2 = 6$.

9. A study seeks to compare hospitals based on the performance of their intensive care units. The response variable is the mortality ratio, the ratio of the number of deaths over the predicted number of deaths based on the condition of the patients. The comparison will be between hospitals with nurse staffing problems and hospitals without such problems. Assume, based on past experience, that the standard deviation of the mortality ratio will be around .2 in both types of hospital. How many of each type of hospital should be included in the study in order to have both the type I and type II error probabilities be .05, if the true difference of mean mortality ratio for the two types of hospital is .2? If we conclude that hospitals with nurse staffing problems have a higher mortality ratio, does this imply a causal relationship? Explain.

10. To decide whether chemistry or physics majors have higher starting salaries in industry, n B.S. graduates of each type are surveyed, yielding $\bar{x} = \$61{,}500$ for chemistry and $\bar{y} = \$61{,}000$ for physics. Assume $\sigma = \$2500$ for both populations.
Calculate the P-value for the appropriate two-sample z test, assuming that the data was based on $n = 100$. Then repeat the calculation for $n = 400$. Is the small P-value for $n = 400$ indicative of a difference that has practical significance? Would you have been satisfied with just a report of the P-value? Comment briefly.

11. a. Show for the upper-tailed test with σ_1 and σ_2 known that as either m or n increases, β decreases when $\mu_1 - \mu_2 > \Delta_0$.

b. For the case of equal sample sizes ($m = n$) and fixed α, what happens to the necessary sample size n as β is decreased, where β is the desired type II error probability at a fixed alternative?

12. The level of monoamine oxidase (MAO) activity in blood platelets (nm/mg protein/h) was determined for each individual in a sample of 43 chronic schizophrenics, resulting in $\bar{x} = 2.69$, as well as for 45 normal subjects, resulting in $\bar{y} = 6.35$. Assume that $\sigma_1 = 2.3$ and $\sigma_2 = 4.0$. Does this data strongly suggest that true average MAO activity for normal subjects is *more than twice* the activity level for schizophrenics? Derive a test procedure and carry out the test using $\alpha = .01$. [*Hint*: Let μ_1 and μ_2 refer to true average MAO activity for schizophrenics and normal subjects, respectively, and consider the parameter $\theta = 2\mu_1 - \mu_2$. Write H_0 and H_a in terms of θ, estimate θ, and derive $\sigma_{\hat{\theta}}$.]

10.2   The Two-Sample *t* Confidence Interval and Test

In the previous section, we illustrated the use of a CI and test procedure for the difference of two means under the assumptions of normally distributed populations with known standard deviations. For large samples, the CLT allows us to use these methods even when the two populations of interest are not normal.

In practice, though, it is virtually always the case that the values of the population standard deviations are unknown. We now proceed by extending the one-sample *t* procedures from Chapters 8 and 9 to the analysis of a difference of means. Such inferential methods still assume normal population distributions, though (as discussed below) that assumption can be relaxed for large sample sizes.

We continue under the assumptions 1–3 stated at the beginning of Section 10.1. Since it is no longer assumed that the population standard deviations σ_1 and σ_2 are known, they will be replaced in Expression (10.1) by the sample standard deviations S_1 and S_2, respectively. The following theorem stems from a result first presented by B. L. Welch in 1938.

WELCH'S THEOREM   When the population distributions are both normal, the standardized variable

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{m} + \dfrac{S_2^2}{n}}} \tag{10.2}$$

has approximately a *t* distribution with df *v* estimated from the data by

$$v = \frac{\left(\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}\right)^2}{\dfrac{(s_1^2/m)^2}{m-1} + \dfrac{(s_2^2/n)^2}{n-1}} = \frac{\left[(se_1)^2 + (se_2)^2\right]^2}{\dfrac{(se_1)^4}{m-1} + \dfrac{(se_2)^4}{n-1}} \tag{10.3}$$

where $se_1 = s_1/\sqrt{m}$ and $se_2 = s_2/\sqrt{n}$ (round *v* down to the nearest integer).

The cumbersome Expression (10.3) is called *Welch's degrees of freedom* (or Welch–Satterthwaite, after another statistician researching this problem around the same time). Of course, statistical software packages have (10.3) built in. Manipulating *T* from (10.2) in a probability statement to isolate $\mu_1 - \mu_2$ gives a CI, whereas a test statistic results from replacing $\mu_1 - \mu_2$ by the null value Δ_0.

**TWO-SAMPLE
t PROCEDURES**   The **two-sample *t* confidence interval for $\mu_1 - \mu_2$** with approximate confidence level $100(1 - \alpha)\%$ is

$$\overline{x} - \overline{y} \pm t_{\alpha/2, v}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where *v* = Welch's df formula (10.3). One-sided confidence bounds can be calculated by retaining the appropriate sign (+ or −) and replacing $t_{\alpha/2, v}$ by $t_{\alpha, v}$.

The **two-sample *t* test** for testing $H_0: \mu_1 - \mu_2 = \Delta_0$ is as follows:

Alternative Hypothesis	**Rejection Region for Approximate Level α Test**
$H_a: \mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha,v}$ (upper-tailed test)
$H_a: \mu_1 - \mu_2 < \Delta_0$	$t \leq -t_{\alpha,v}$ (lower-tailed test)
$H_a: \mu_1 - \mu_2 \neq \Delta_0$	either $t \geq t_{\alpha/2,v}$ or $t \leq -t_{\alpha/2,v}$ (two-tailed test)

Test statistic value: $t = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}}}$

A *P*-value can be computed as described in Section 9.4 for the one-sample *t* test.

Example 10.6 Which way of dispensing champagne, the traditional vertical method or a tilted "beer-like" pour, preserves more of the tiny gas bubbles that improve flavor and aroma? The following data was reported in the article "On the Losses of Dissolved CO_2 during Champagne Serving" (*J. Agr. Food Chem.* 2010: 8768–6775).

Temperature (°C)	Type of pour	*n*	Mean (g/L)	SD
18	Traditional	4	4.0	.5
18	Slanted	4	3.7	.3
12	Traditional	4	3.3	.2
12	Slanted	4	2.0	.3

Assuming that the sampled distributions are normal, let's calculate confidence intervals for the difference between true average dissolved CO_2 loss for the traditional pour and that for the slanted pour at each of the two temperatures.

For the 18°C temperature, Welch's df is

$$v = \frac{\left(\dfrac{.5^2}{4} + \dfrac{.3^2}{4}\right)^2}{\dfrac{(.5^2/4)^2}{3} + \dfrac{(.3^2/4)^2}{3}} = \frac{.007225}{.00147083} = 4.91$$

Rounding down, the CI will be based on 4 df. For a confidence level of (approximately) 99%, we need $t_{.005,4} = 4.604$. The desired interval is

$$4.0 - 3.7 \pm (4.604)\sqrt{\frac{.5^2}{4} + \frac{.3^2}{4}} = .3 \pm (4.604)(.2915) = .3 \pm 1.3 = (-1.0, 1.6)$$

Thus we can be highly confident that $-1.0 < \mu_1 - \mu_2 < 1.6$, where μ_1 and μ_2 are true average losses for the traditional and slant methods, respectively. Notice that this CI contains 0, so at the 99% confidence level, it is plausible that $\mu_1 - \mu_2 = 0$, that is, that $\mu_1 = \mu_2$.

The df formula for the 12°C comparison yields df = .00105625/.00020208 = 5.23, necessitating the use of $t_{.005,5} = 4.032$ for a 99% CI. The resulting interval is (.6, 2.0). Thus 0 is not a plausible value for this difference. It appears from the CI that the true average loss when the slant method is used is smaller than that when the traditional method is used, so that the slant method is better at this temperature. This in fact was the conclusion reported in the popular media. ∎

Example 10.7 What color should you use for your Web site's background? The authors of "Waiting for the Web: How Screen Color Affects Time Perception" (*J. Mark. Res.* 2004) compared subjects' time perception based on the background color of a Web site being downloaded. Subjects were randomly assigned to see a blue background or a yellow background; the Web sites were otherwise identical, *including the actual download time*. Data consistent with the information in the article appears in Figure 10.1. Values of summary statistics appear in Table 10.1.
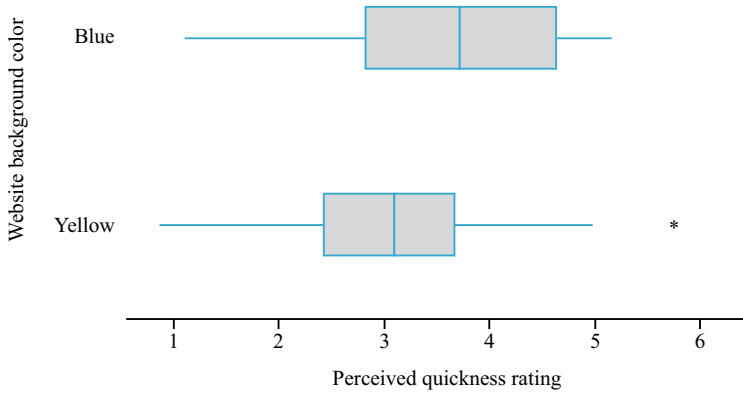
Figure 10.1 Comparative boxplot for Example 10.7

Table 10.1 Values of summary statistics for Example 10.7

		Perceived quickness rating	
	n	Mean	SD
Blue	25	3.67	1.07
Yellow	24	3.04	1.07

Let's test to see whether background color affects users' average perception of download time, at the 5% level. (A higher "perceived quickness" rating indicates the subject *thought* the page downloaded faster.)

1. The parameters of interest are
   μ_1 = true mean perceived quickness rating with a blue background
   μ_2 = true mean perceived quickness rating with a yellow background
2. $H_0: \mu_1 - \mu_2 = 0$
   $H_a: \mu_1 - \mu_2 \neq 0$
3. Subjects were randomly assigned to blue or yellow background color, so it is reasonable to treat the groups' responses as independent. Normal probability plots of data consistent with the article appear in Figure 10.2; the patterns in both plots are reasonably linear, so neither one suggests a marked deviation from normality. (We can be a little forgiving about some curvature here, since $m = 25$ and $n = 24$ are not too small.) It is therefore valid to use the two-sample *t* test for this analysis.
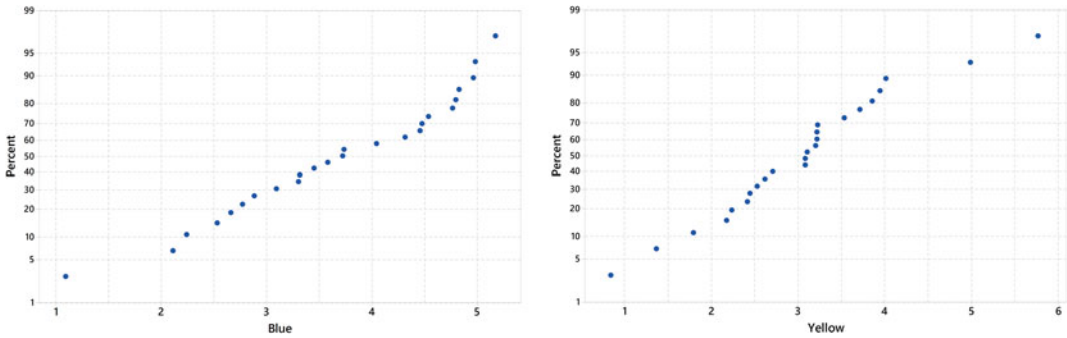
Figure 10.2 Normal probability plots for Example 10.7

4. The null value is $\Delta_0 = 0$, so the test statistic value is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}}}$$

5. Welch's df formula (10.3) gives $\nu = 46.9$, which we round down to 46 df. From software, $t_{.025,46} = 2.013$, so we will reject H_0 if either $t \geq 2.013$ or $t \leq -2.013$.
6. Using the values in Table 10.1,

$$t = \frac{(3.67 - 3.04) - 0}{\sqrt{\dfrac{(1.07)^2}{25} + \dfrac{(1.07)^2}{24}}} = 2.06$$

7. Since $2.06 \geq 2.013$, using a significance level of .05 we can (barely) reject the null hypothesis in favor of the alternative hypothesis, confirming the conclusion stated in the article: users' perceptions of the speed at which a Web site downloads differ depending on whether the background color is blue or yellow. However, someone demanding more compelling evidence might select $\alpha = .01$, a level for which H_0 cannot be rejected.

Using the P-value approach, for this two-tailed test and with the aid of software,

$$P\text{-value} \approx 2 \cdot P(T \geq 2.06 \text{ when } T \sim t_{46}) = 2(.023) = .046$$

Because $.046 \leq .05$, H_0 would again barely be rejected at the $\alpha = .05$ significance level (but not rejected at the .01 level, since $.046 > .01$).

This isn't the whole story: the same study also measured the sense of relaxation users felt when viewing the Web sites. They found that an increased sense of relaxation associated with the color blue accounted for subjects' higher average perceived quickness. Blue backgrounds don't make downloads seem quicker *per se*; they relax the user more and make download time less noticeable.   ∎

Motivation for Welch's Theorem

Dividing numerator and denominator of (10.2) by the standard deviation of the numerator gives

$$\frac{[\overline{X} - \overline{Y} - (\mu_1 - \mu_2)] \left/ \sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}\right.}{\sqrt{\dfrac{S_1^2}{m} + \dfrac{S_2^2}{n}} \left/ \sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}\right.}$$

The numerator of this ratio is a standard normal rv because it results from standardizing the normally distributed difference $\overline{X} - \overline{Y}$. The denominator is independent of the numerator because the sample variances are independent of the sample means for normal samples. However, in order for (10.2) to be a t random variable, the denominator needs to be the square root of a chi-squared rv divided by its df, and unfortunately this is not the case. So let us try to express the denominator at least approximately as $\sqrt{W/v}$ with $W \sim \chi_v^2$, yielding

$$\frac{S_1^2}{m} + \frac{S_2^2}{n} = \left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)\frac{W}{v}$$

To determine v we equate the means and variances of both sides, with the help of $E(W) = v$, $V(W) = 2v$, $(m-1)S_1^2/\sigma_1^2 \sim \chi_{m-1}^2$, and $(n-1)S_2^2/\sigma_2^2 \sim \chi_{n-1}^2$ from Sections 6.3 and 6.4. It follows that $E(S_1^2) = \sigma_1^2$, $V(S_1^2) = 2\sigma_1^4/(m-1)$, and similarly for S_2^2. The mean of the left-hand side is

$$E\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

which is also the mean of the right-hand side, so the means are equal. The variance of the left-hand side is

$$V\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right) = \frac{2\sigma_1^4}{(m-1)m^2} + \frac{2\sigma_2^4}{(n-1)n^2}$$

and the variance of the right-hand side is

$$V\left[\left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)\frac{W}{v}\right] = \left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)^2 \cdot \frac{2v}{v^2} = \left(\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)^2 \cdot \frac{2}{v}$$

Now equate these two variances, substituting sample variances for the unknown population variances, and solve for v. This gives Expression (10.3) in Welch's Theorem.

Large-Sample t Procedures

We have seen in previous chapters that t-based CIs and hypothesis tests can be applied to data from nonnormal populations provided that the sample sizes are sufficiently large. The same is true for the two-sample t procedures: Welch's Theorem is still approximately correct even if the X's and Y's are not sampled from normal distributions, so long as both sample sizes are large enough. We'll continue to use the convention that $m > 40$ and $n > 40$ qualify as "large" samples.

Also in parallel with previous chapters, for large samples there is little practical difference between using z or t critical values for inference. It can be shown (Exercise 97) that Welch's df satisfies $v \geq \min(m-1, n-1)$, so that v is large if both m and n are. In that situation, using z values for the

procedures of this section—equivalently, substituting s_1 and s_2 for σ_1 and σ_2 in the two-sample z procedures of Section 10.1—will yield similar results to the two-sample t procedures.

Example 10.8 A study was carried out in an attempt to improve student performance in a low-level university mathematics course. Experience had shown that many students had fallen by the wayside, meaning that they had dropped out or completed the course with minimal effort and low grades. The study involved assigning the students to sections based on odd or even Social Security number. It is important that the assignment to sections not be on the basis of student choice, because then the differences in performance might be attributable to differences in student attitude or ability. Half of the sections were taught traditionally, whereas the other half were taught in a way that hopefully would keep the students involved. They were given frequent assignments that were collected and graded, they had frequent quizzes, and they were allowed retakes on exams.

Prof. Lotus Hershberger conducted the experiment and he supplied the final exam scores, out of 40 points possible, for the 79 students taught traditionally (the control group) and for the 85 students taught with more involvement (the experimental group). Table 10.2 summarizes the data. Does this information suggest that true mean for the experimental condition exceeds that for the control condition? Let's use a test with $\alpha = .05$.

Table 10.2 Summary results for Example 10.8

Group	Sample size	Sample mean	Sample SD
Control	79	23.87	11.60
Experimental	85	27.34	8.85

Let μ_1 and μ_2 denote the true mean scores for the control condition and the experimental condition, respectively. The two hypotheses are H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 < 0$. Welch's degrees of freedom $\nu = 145$ here; since the t_{145} and z curves are virtually indistinguishable, we'll use the latter here. H_0 will be rejected if $z \leq -z_{.05} = -1.645$. Then

$$z = \frac{(23.87 - 27.34) - 0}{\sqrt{\dfrac{11.60^2}{79} + \dfrac{8.85^2}{85}}} = \frac{-3.47}{1.620} = -2.14$$

Since $-2.14 \leq -1.645$, H_0 is rejected at significance level .05. Alternatively, the P-value for a lower-tailed z test is

$$P\text{-value} = \Phi(z) = \Phi(-2.14) = .016$$

which implies rejection at significance level .05.

We have shown fairly conclusively that the experimental method of instruction is an improvement. Nevertheless, there is more to be said. It is important to view the data graphically to see if there is anything strange. Figure 10.3 combines a boxplot and dotplot.

The plot shows that both groups have outlying observations at the low end; some students showed up for the final but performed very poorly. What happens if we compare the groups while ignoring the low performers whose scores are below 10? The resulting summary information is in Table 10.3.
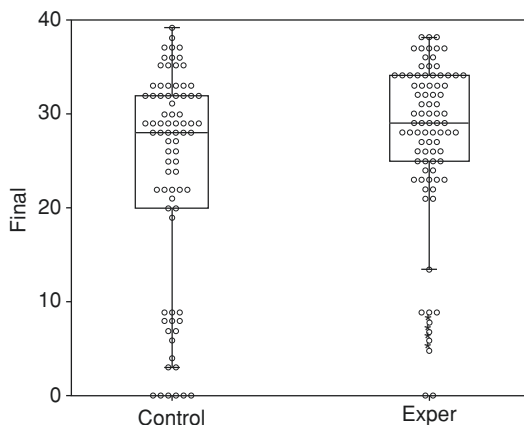
Figure 10.3   Boxplot/dotplot for the teaching experiment

Table 10.3   Summary results without poor performers

Group	Sample size	Sample mean	Sample SD
Control	61	29.59	5.005
Experimental	76	29.88	4.950

Notice that the means and standard deviations for the two groups are now very similar. Indeed, based on Table 10.3 the test statistic value is $-.34$, giving no reason to reject the null hypothesis. For the majority of the students, there appears to be not much effect from the experimental treatment. It is the low performers who make a big difference in the results. There were 18 low performers in the control group but only 9 in the experimental group. The effect of the experimental instruction is to decrease the number of students who perform at the bottom of the scale. This is in accord with the goals of the experimental treatment, which was designed to keep students on track. ∎

Pooled *t* Procedures

Alternatives to the two-sample *t* procedures described in this section result from assuming not only that the two population distributions are normal but also that they have equal, albeit unknown, standard deviations $(\sigma_1 = \sigma_2)$. That is, the two population distribution curves are assumed normal with equal spreads, the only possible difference between them being where they are centered (i.e., at μ_1 and μ_2).

Let σ denote the common population standard deviation. Then standardizing $\overline{X} - \overline{Y}$ gives

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma^2}{m} + \dfrac{\sigma^2}{n}}} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\dfrac{1}{m} + \dfrac{1}{n}}}$$

which has a standard normal distribution. Before this variable can be used as a basis for making inferences about $\mu_1 - \mu_2$, the common variance must be estimated from sample data. One estimator of σ^2 is S_1^2, the variance of the *m* observations in the first sample, and another is S_2^2, the variance of the second sample. Intuitively, a better estimator than either individual sample variance results from combining the two sample variances. A first thought might be to use $(S_1^2 + S_2^2)/2$, the ordinary average of the two sample variances. However, if $m > n$ then the first sample contains more

information about σ^2 than does the second sample, and an analogous comment applies if $m < n$. The following *weighted* average of the two sample variances, called the **pooled** (i.e., combined) **estimator** of σ^2, adjusts for any difference between the two sample sizes:

$$S_p^2 = \frac{m-1}{m+n-2}S_1^2 + \frac{n-1}{m+n-2}S_2^2$$

It can be shown (Exercise 39) that S_p^2 is proportional to a chi-squared rv with $m + n - 2$ df. In turn, the rv that results if S_p^2 replaces σ^2 in the above Z statistic follows a t distribution with $m + n - 2$ df (Exercise 40). In the same way that earlier standardized variables were used as a basis for deriving confidence intervals and test procedures, this t variable immediately leads to the *pooled t confidence interval* for estimating $\mu_1 - \mu_2$ and the *pooled t test* for testing hypotheses about a difference between means. In particular, the pooled t test statistic for testing $H_0: \mu_1 - \mu_2 = \Delta_0$ is

$$T_p = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{\sqrt{\dfrac{S_p^2}{m} + \dfrac{S_p^2}{n}}} = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{S_p\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}},$$

and $T_p \sim t_{m+n-2}$ when H_0 is true.

   In the past, many statisticians recommended these pooled t procedures over the two-sample t procedures. The pooled t test, for example, can be derived from the likelihood ratio principle, whereas the two-sample t test is not a likelihood ratio test. Furthermore, the significance level for the pooled t test is exact, whereas it is only approximate for the two-sample t test. Finally, power and sample sizes calculations using the pooled t procedure can easily be performed by software using a noncentral t distribution (Exercise 123).

   However, statistical research has shown that while the pooled t test does outperform the two-sample t test by a bit (more power for the same α) when $\sigma_1 = \sigma_2$, the former test can easily lead to erroneous conclusions if applied when the population standard deviations are different. Analogous comments apply to the behavior of the two confidence intervals. That is, the pooled t procedures are *not* robust to violations of their equal variance assumption.

   It has been suggested that one could carry out a preliminary test of $H_0: \sigma_1 = \sigma_2$ and use a pooled t procedure if this null hypothesis is not rejected. Unfortunately, the usual "F test" of equal variances (Section 10.5) is quite sensitive to the assumption of normal population distributions, much more so than t procedures. We therefore recommend the conservative approach of using two-sample t procedures unless there is really compelling evidence for doing otherwise, particularly when the two sample sizes are different.

Power and Type II Error Probabilities

Determining power and β for the two-sample t test is complicated. The most recent versions of R, SAS, Minitab, and JMP will calculate power for the pooled t test—that is, assuming a common value for σ_1 and σ_2—but not for the two-sample t test. However, Prof. Russell Lenth (Univ. of Iowa) has developed a Java software package that performs such power calculations for the two-sample t test; the package can be downloaded for free from his Web site. The software will also calculate sample sizes necessary to obtain a specified power for a particular value of $\mu_1 - \mu_2$.

   In general, power will increase (β will decrease) as the sample sizes increase, as α increases, and as $\mu_1 - \mu_2$ moves farther from Δ_0. When m and n are both large, the quantity T in (10.2) also has an approximately normal distribution, and so the power, β, and sample size formulas from Section 10.1 provide approximately correct values. Population sd's in those formulas can be replaced by sample sd's.

Exercises: Section 10.2 (13–42)

13. Determine the number of degrees of freedom for the two-sample *t* test or CI in each of the following situations:

    a. $m = 10$, $n = 10$, $s_1 = 5.0$, $s_2 = 6.0$
    b. $m = 10$, $n = 15$, $s_1 = 5.0$, $s_2 = 6.0$
    c. $m = 10$, $n = 15$, $s_1 = 2.0$, $s_2 = 6.0$
    d. $m = 12$, $n = 24$, $s_1 = 5.0$, $s_2 = 6.0$

14. A 2008 study in the *J. Family Econ. Issues* compared the work and home habits of female and male lawyers in Canada. All participants in the survey had at least one child; single parents and couples who are both lawyers were excluded. Two of the variables measured were the weekly number of hours spent in the office and the number of hours spent with their children on weekdays.

		Weekly work hours		Weekday hours w/kids	
	n	Mean	SD	Mean	SD
Mothers	230	41.38	11.90	3.27	1.68
Fathers	604	48.09	10.30	1.82	1.29

    a. Estimate, with 95% confidence, the difference in the average weekly number of work hours between mothers and fathers who practice law.
    b. Estimate, with 95% confidence, the difference in the average number of hours female and male lawyers spend with their kids on weekdays. Then, convert this interval into an estimate for the difference in average *weekly* hours spent with kids (*Hint:* The first interval is a daily average, and a work week has 5 days).

    [*Note:* Interestingly, the study also found that "contrary to assumptions in the literature and the workplace, mothers practicing law are significantly more committed to their careers than fathers."]

15. The article "Return Migration, Investment in Children, and Intergenerational Mobility: Comparing Sons of Foreign- and Native-Born Fathers" (*J. Hum. Res.* 2008: 299–324)

presented the following summary data on years of education both for sample of sons of native-born fathers in Germany and another sample of sons of foreign-born fathers.

	n	$\bar{x}$	*s*
Foreign-born	251	9.2	1.9
Native-born	640	11.7	2.6

Does the true average years of education for sons of native-born fathers appear to exceed that for those with foreign-born fathers? State and test the appropriate hypotheses using a significance level of .01.

16. The accompanying time-to-repair (min) data for both high rail and low rail breaks on curved track appeared in the article "Uncertainty Estimation in Railway Track Life-Cycle Cost" (*J. Rail Rapid Transit* 2009: 285–293). (On a curved track, the high rail is the outer rail with the larger radius, while the low rail is the inner rail with the smaller radius.)

High:	159	120	480	149	270	547	340
	43	228	202	240	218		

Low:	258	154	216	240	169	75	340
	202	202	216				

    Normal probability plots of both samples show reasonably linear patterns.

    a. Construct a comparative boxplot and comment on interesting features.
    b. Carry out a test of hypotheses at significance level .10 to decide if there is evidence for concluding that true average repair time for high rails exceeds that for low rails by more than 30 min.
    c. Obtain and interpret a confidence interval at the 90% confidence level for the difference between true average repair times for high and low rails.

17. Due to recent concerns about player concussions, football helmets have recently increased in both size and mass. Have these changes made a difference? The article

"The Effects of Helmet Weight on Hybrid III Head and Neck Responses by Comparing Unhelmeted and Helmeted Impacts" (*J. Biomech. Engr.* 2016) reports on an experiment in which repeated impact trials were performed on an artificial human head both wearing and not wearing a football helmet.

a. The following summary information for the variable head acceleration (g) is consistent with information in the article.

	Sample size	Mean	SD
Helmet	24	43.1	4.5
No helmet	24	75.4	7.2

Test whether the average head acceleration is reduced by helmet wear at the .05 significance level.

b. The researchers were concerned that the mass of the helmet might increase the force experienced by the upper neck. The following summary information for resultant neck force (Newtons) is consistent with information in the article.

	Sample size	Mean	SD
Helmet	24	1331	93
No helmet	24	945	77

Test whether the average resultant neck force is increased by helmet wear at the .05 level.

[*Note:* The authors conclude that "the increased neck forces provide a possible explanation as to why there has not been a … reduction in concussion rates despite improvements in helmets' ability to reduce head accelerations."]

c. If the null hypotheses in (a) and (b) are in fact both true, what can be said about the chance that at least one type I error is committed by the two tests?

18. The article "Evaluation of a Ventilation Strategy to Prevent Barotrauma in Patients at High Risk for Acute Respiratory Distress Syndrome" (*New Engl. J. Med.* 1998:

355–358) reported on an experiment in which 120 patients with similar clinical features were randomly divided into a control group and a treatment group, each consisting of 60 patients. The sample mean ICU stay (days) and sample standard deviation for the treatment group were 19.9 and 39.1, respectively, whereas these values for the control group were 13.7 and 15.8.

a. Calculate a point estimate for the difference between true average ICU stay for the treatment and control groups. Does this estimate suggest that there is a significant difference between true average stays under the two conditions?

b. Answer the question posed in part (a) by carrying out a formal test of hypotheses. Is the result different from what you conjectured in part (a)?

c. Does it appear that ICU stay for patients given the ventilation treatment is normally distributed? Explain your reasoning.

d. Estimate true average length of stay for patients given the ventilation treatment in a way that conveys information about precision and reliability.

19. What impact does fast-food consumption have on various dietary and health characteristics? The article "Effects of Fast-Food Consumption on Energy Intake and Diet Quality among Children in a National Household Study" (*Pediatrics* 2004: 112–118) reported the accompanying summary data on daily calorie intake both for a sample of teens who said they did not typically eat fast food and another sample of teens who said they did usually eat fast food.

Eat fast food	Sample size	Sample mean	Sample SD
No	663	2258	1519
Yes	413	2637	1138

a. Estimate the difference between true average calorie intake for teens who typically don't eat fast foods and true average intake for those who do eat fast foods, and

do so in a way that conveys information about reliability and precision.

b. Does this data provide strong evidence for concluding that true average calorie intake for teens who typically eat fast food exceeds true average intake for those who don't typically eat fast food by more than 200 cal/day? Carry out a test at significance level .05 based on determining the *P*-value.

20. Much research has focused on comparing business environment cultures across several countries. The article "Perception of Internal Factors for Corporate Entrepreneurship: A Comparison of Canadian and U.S. Managers" (*Entrep. Theory Pract.* 1999: 9–24) presented the following summary data on hours per week managers spent thinking about new ideas.

Country	Sample size	Sample mean	Sample SD
U.S.	174	5.8	6.0
Canada	353	5.1	4.6

Does it appear that true average time per week that U.S. managers spend thinking about new ideas differs from that for Canadian managers? State and test the relevant hypotheses.

21. Credit card spending and resulting debt pose very real threats to consumers in general, and the potential for abuse is especially serious among college students. It has been estimated that about two-thirds of all college students possess credit cards, and 80% of these students received cards during their first year of college. The article "College Students' Credit Card Debt and the Role of Parental Involvement: Implications for Public Policy" (*J. Public Policy Mark.* 2001: 105–113) reported that for 209 students whose parents had no involvement whatsoever in credit card acquisition or payments, the sample mean total account balance was $421 with a sample standard deviation of $686, whereas for 75 students whose parents assisted with payments even

though they were under no legal obligation to do so, the sample mean and sample standard deviation were $666 and $1048, respectively. All sampled students were at most 21 years of age.

a. Do you think it is plausible that the distributions of total debt for these two types of students are normal? Why or why not? Is it necessary to assume normality in order to compare the two groups using an inferential procedure described in this chapter? Explain.

b. Estimate the true average difference between total balance for noninvolvement students and postacquisition-involvement students using a method that incorporates precision into the estimate. Then interpret the estimate. [*Note*: Data was also reported in the article for preacquisition involvement only and for both pre- and postacquisition involvement.]

22. Returning to the previous exercise, the mean and standard deviation of the number of credit cards for the no-involvement group were 2.22 and 1.58, respectively, whereas the mean and standard deviation for the payment-help group were 2.09 and 1.65, respectively. Does it appear that the true average number of cards for no-involvement students exceeds the average for payment-help students? Carry out an appropriate test of significance.

23. Expert and amateur pianists were compared in a study "Maintaining Excellence: Deliberate Practice and Elite Performance in Young and Older Pianists" (*J. Exp. Psychol. Gen.* 1996: 331–340). The researchers used a keyboard that allowed measurement of the force applied by a pianist in striking a key. All 48 pianists played Prelude Number 1 from Bach's Well-Tempered Clavier. For 24 amateur pianists the mean force applied was 74.5 with standard deviation 6.29, and for 24 expert pianists the mean force was 81.8 with standard deviation 8.64. Do

expert pianists hit the keys harder? Assuming normally distributed data, state and test the relevant hypotheses, and interpret the results.

24. The article "Supervised Exercise Versus Non-Supervised Exercise for Reducing Weight in Obese Adults" (*J. Sport. Med. Phys. Fit.* 2009: 85–90) reported on an investigation in which participants were randomly assigned either to a supervised exercise program or a control group. Those in the control group were told only that they should take measures to lose weight. After 4 months, the sample mean decrease in body fat for the 17 individuals in the experimental group was 6.2 kg with a sample standard deviation of 4.5 kg, whereas the sample mean and sample standard deviation for the 17 people in the control group were 1.7 kg and 3.1 kg, respectively. Assume normality of the two body fat loss distributions (as did the investigators).

   a. Calculate a 99% lower prediction bound for the body fat loss of a single randomly selected individual subjected to the supervised exercise program. Can you be highly confident that such an individual will actually lose body fat?
   b. Does it appear that true average decrease in body fat is more than 2 kg larger for the experimental condition than for the control condition? Carry out a test of appropriate hypotheses using a significance level of .01.

25. Fusible interlinings are being used with increasing frequency to support outer fabrics and improve the shape and drape of various pieces of clothing. The article "Compatibility of Outer and Fusible Interlining Fabrics in Tailored Garments" (*Textile Res. J.* 1997: 137–142) gave the accompanying data on extensibility (%) at 100 g/cm for both high-quality fabric (H) and poor-quality fabric (P) specimens.

H	1.2	.9	.7	1.0	1.7	1.7	1.1	.9	1.7
	1.9	1.3	2.1	1.6	1.8	1.4	1.3	1.9	1.6
	.8	2.0	1.7	1.6	2.3	2.0			
P	1.6	1.5	1.1	2.1	1.5	1.3	1.0	2.6	

   a. Construct normal probability plots to verify the plausibility of both samples having been selected from normal population distributions.
   b. Construct a comparative boxplot. Does it suggest that there is a difference between true average extensibility for high-quality fabric specimens and that for poor-quality specimens?
   c. The sample mean and standard deviation for the high-quality sample are 1.508 and .444, respectively, and those for the poor-quality sample are 1.588 and .530. Use the two-sample t test to decide whether true average extensibility differs for the two types of fabric.

26. Imaging of the colon with a contrast dye to evaluate for injury requires that the colon first be distended by pumping in carbon dioxide. The article "Determination of Normal Distribution of Distended Colon Volumes to Guide Performance of Colonic Imaging With Fluid Distention" (*Curr. Probl. Diagn. Radiol.* 2016: 185–188) reported that for a sample of 85 female patients undergoing this procedure, the mean colon length after distention was 201.8 cm and the standard deviation was 32.2 cm, whereas for a sample of 31 males the mean and standard deviation were 180.2 cm and 38.6 cm, respectively.

   a. Carry out a test at significance level .1 to decide whether true average length differs by sex (the article reported a *P*-value for this test).
   b. Construct and interpret 90% CI for the difference in true average colon length between the two sexes under these settings. Is your interval consistent with the result of the test in part (a)? Explain.

27. Research has shown that good hip range of motion and strength in throwing athletes results in improved performance and decreased body stress. The article "Functional Hip Characteristics of Baseball Pitchers and Position Players" (*Am. J. Sport. Med.* 2010: 383–388) reported on a study involving samples of 40 professional pitchers and 40 professional position players. For the pitchers, the sample mean trail leg total arc of motion (degrees) was 75.6 with a sample standard deviation of 5.9, whereas the sample mean and sample standard deviation for position players were 79.6 and 7.6, respectively. Assuming normality, test appropriate hypotheses to decide whether true average range of motion for the pitchers is less than that for the position players (as hypothesized by the investigators). In reaching your conclusion, what type of error might you have committed?

28. Tennis elbow is thought to be aggravated by the impact experienced when hitting the ball. The article "Forces on the Hand in the Tennis One-Handed Backhand" (*Int. J. Sport Biomech.* 1991: 282–292) reported the force (Newtons) on the hand just after impact on a one-handed backhand drive for six advanced players and for eight intermediate players.

Type of player	Sample size	Sample mean	Sample SD
1. Advanced	6	40.3	11.3
2. Intermediate	8	21.4	8.3

In their analysis of the data, the authors assumed that both force distributions were normal. Calculate a 95% CI for the difference between true average force for advanced players (μ_1) and true average force for intermediate players (μ_2). Does your interval provide compelling evidence for concluding that the two μ's are different? Would you have reached the same conclusion by calculating a CI for $\mu_2 - \mu_1$

(i.e., by reversing the 1 and 2 labels on the two types of players)? Explain.

29. As the population ages, there is increasing concern about accident-related injuries to the elderly. The article "Age and Gender Differences in Single-Step Recovery from a Forward Fall" (*J. Gerontol A Biol. Sci. Med. Sci.* 1999 54(1):M44–50) reported on an experiment in which the maximum lean angle—the farthest a subject is able to lean and still recover in one step—was determined for both a sample of younger females (21–29 years) and a sample of older females (67–81 years). The following observations are consistent with summary data given in the article:

Younger: 29, 34, 33, 27, 28, 32, 31, 34, 32, 27
Older: 18, 15, 23, 13, 12

Does the data suggest that true average maximum lean angle for older females is more than 10 degrees smaller than it is for younger females? State and test the relevant hypotheses at significance level. 10 by obtaining a *P*-value.

30. The article "Effect of Internal Gas Pressure on the Compression Strength of Beverage Cans and Plastic Bottles" (*J. Test. Eval.* 1993: 129–131) includes the accompanying data on compression strength (lb) for a sample of 12-oz aluminum cans filled with strawberry drink and another sample filled with cola. Does the data suggest that the extra carbonation of cola results in a higher average compression strength? Base your answer on a *P*-value. What assumptions are necessary for your analysis?

Beverage	Sample size	Sample mean	Sample SD
Strawberry drink	15	540	21
Cola	15	554	15

31. Which foams more when you pour it, Coke or Pepsi? Here are measurements by Diane Warfield on the foam volume (mL) after

pouring a 12-oz can of Coke, based on a sample of 12 cans:

312.2 292.6 331.7 355.1 362.9 331.7
292.6 245.8 280.9 320.0 273.1 288.7

and here are measurements for Pepsi, based on a sample of 12 cans:

148.3 210.7 152.2 117.1   89.7 140.5
128.8 167.8 156.1 136.6 124.9 136.6

a. Verify graphically that normality is an appropriate assumption.
b. Calculate a 99% confidence interval for the population difference in mean volumes.
c. Does the upper limit of your interval in (b) give a 99% lower confidence bound for the difference between the two μ's? If not, calculate such a bound and interpret it in terms of the relationship between the foam volumes of Coke and Pepsi.
d. Summarize in a sentence what you have learned about the foam volumes of Coke and Pepsi.

32. In a comparative study conducted at Virginia Tech, two Principles of Economics classes were run in an identical fashion except for one respect: one class used an interactive electronic teaching system (called WITS) for seven "research exercises," while the other class discussed the research exercises but did not use the interactive devices. Final exam score results are summarized below ("Technology Improves Learning in Large Principles of Economics Classes: Using Our WITS," *Am. Econ. Rev.* 2004: 442–446).

	Sample size	Sample mean	Sample SD
WITS	62	77.45	11.1
Traditional	64	74.25	8.7

a. Test the see whether the true mean final exam scores using WITS and using traditional instruction are different, at the $\alpha = .10$ significance level.

b. Construct a 90% CI for the difference in true mean final exam score for WITS instruction and traditional instruction. Is your interval consistent with the test in part (a)?
c. What does the interval in part (b) say about the practical significance of the test?

33. The article "Characterization of Bearing Strength Factors in Pegged Timber Connections" (*J. Struct. Engr.* 1997: 326–332) gave the following summary data on proportional stress limits for specimens constructed using two different types of wood:

Type of wood	Sample size	Sample mean	Sample SD
Red oak	14	8.48	.79
Douglas fir	10	6.65	1.28

Assuming that both samples were selected from normal distributions, carry out a test of hypotheses to decide whether the true average proportional stress limit for red oak joints exceeds that for Douglas fir joints by more than 1 MPa.

34. According to the article "Fatigue Testing of Condoms" (*Polym. Test.* 2009: 567–571), "tests currently used for condoms are surrogates for the challenges they face in use," including a test for holes, an inflation test, a package seal test, and tests of dimensions and lubricant quality (all fertile territory for the use of statistical methodology!). The investigators developed a new test that adds cyclic strain to a level well below breakage and determines the number of cycles to break. The cited article reported that for a sample of 20 natural latex condoms of a certain type, the sample mean and sample standard deviation of the number of cycles to break were 4358 and 2218, respectively, whereas a sample of 20 polyisoprene condoms gave a sample mean and sample standard deviation of 5805 and 3990, respectively. Is there strong evidence for concluding that the true average number of

cycles to break for the polyisoprene condom exceeds that for the natural latex condom by more than 1000 cycles? [*Note*: The article presented the results of hypothesis tests based on the *t* distribution; the validity of these depends on assuming normal population distributions.]

35. Exercise 22 from Chapter 9 gave the following data on amount (oz) of alcohol poured into a short, wide tumbler glass by a sample of experienced bartenders: 2.00, 1.78, 2.16, 1.91, 1.70, 1.67, 1.83, 1.48. The cited article also gave summary data on the amount poured by a different sample of experienced bartenders into a tall, slender (highball) glass; the following observations are consistent with the reported summary data: 1.67, 1.57, 1.64, 1.69, 1.74, 1.75, 1.70, 1.60.

    a. What does a comparative boxplot suggest about similarities and differences in the data?
    b. Carry out a test of hypotheses to decide whether the true average amount poured is different for the two types of glasses; be sure to check the validity of any assumptions necessary to your analysis, and report a *P*-value.

36. Is the incidence of head or neck pain among video display terminal users related to the monitor angle (degrees from horizontal)? The paper, "An Analysis of VDT Monitor Placement and Daily Hours of Use for Female Bifocal Users" (*Work* 2003: 77–80), reported the accompanying data. Carry out an appropriate test of hypotheses (be sure to include a *P*-value in your analysis).

Pain	Sample size	Sample mean	Sample SD
Yes	32	2.20	3.42
No	40	3.20	2.52

37. The article "Gender Differences in Individuals with Comorbid Alcohol Dependence and Post-Traumatic Stress Disorder" (*Am. J. Addict.* 2003: 412–423) reported the accompanying data on total score on the Obsessive-Compulsive Drinking Scale (OCSD).

Gender	Sample size	Sample mean	Sample SD
Male	44	19.93	7.74
Female	40	16.26	7.58

Formulate hypotheses and carry out an appropriate analysis. Does your conclusion depend on whether a significance level of .05 or .01 was employed? (The cited paper reported *P*-value < .05; presumably .05 would have been replaced by .01 if the *P*-value were really that small).

38. Which factors are relevant to the time a consumer spends looking at a product on the shelf prior to selection? The article "Effects of Base Price upon Search Behavior of Consumers in a Supermarket" (*J. Econ. Psychol.* 2003: 637–652) reported the following data on elapsed time (sec) for fabric softener purchasers and washing-up liquid purchasers; the former product is significantly more expensive than the latter. These products were chosen because they are similar with respect to allocated shelf space and number of alternative brands.

Product	Sample size	Sample mean	Sample SD
Fabric softener	15	30.47	19.15
Washing-up liquid	19	26.53	15.37

    a. What if any assumptions are needed before an inferential procedure can be used to compare true average elapsed times?
    b. If just the two sample means had been reported, would they provide persuasive evidence for a significant difference between true average elapsed times for the two products?

c. Carry out an appropriate test of significance and state your conclusion.

39. Let $X_1, \ldots, X_m \sim N(\mu_1, \sigma)$ and $Y_1, \ldots, Y_n \sim N(\mu_2, \sigma)$ be independent random samples from the specified normal population distributions (note that the population sd's are equal). Let S_1^2 and S_2^2 denote the sample variance of the two samples, and define a pooled variance estimator of σ^2 by

$$S_p^2 = \frac{m-1}{m+n-2}S_1^2 + \frac{n-1}{m+n-2}S_2^2$$

Show that $(m+n-2)S_p^2/\sigma^2$ has a chi-squared distribution with $m + n - 2$ df. [*Hint:* Recall from Chapter 6 that $(m-1)S_1^2/\sigma^2 \sim \chi^2_{m-1}$ and similarly for the second sample variance. What is the distribution of the sum of two independent chi-squared rvs?]

40. Refer back to the scenario of the previous exercise.

a. Verify that the standardized variable $[(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)]/\sqrt{\sigma^2(1/m + 1/n)}$ has a standard normal distribution.

b. Show that the pooled t variable

$$T_p = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}}$$

has a t distribution with $m + n - 2$ df. [*Hint*: create a t-distributed variable using the standard normal rv from part

(a) and the chi-squared rv from the previous exercise.]

41. Consider the pooled t variable T_p from part (b) of the previous exercise.

a. Use this t variable to obtain a pooled t confidence interval formula for $\mu_1 - \mu_2$.

b. The article "Effect of Welding on a High-Density Polyethylene Liner" (*J. Mater. Civil Engr.* 1996: 94–100) reported the following data on tensile strength (psi) of liner specimens both when a certain fusion process was used and when this process was not used.

No fusion	2748	2700	2655	2822	2511	3213	3220	2753
	3149	3257						
Fused	3027	3356	3359	3297	3125	2910	2889	2902

Use the pooled t formula from part (a) to estimate the difference between true average tensile strength for the two processes with a 95% confidence interval.

c. Estimate the difference between the two μ's using the two-sample t interval discussed in this section, and compare it to the interval of part (b).

42. Refer to the previous two exercises. Describe the pooled t test for testing H_0: $\mu_1 - \mu_2 = 0$ when both population distributions are normal with $\sigma_1 = \sigma_2$. Then use this test procedure to test the hypotheses in Example 10.7.

10.3 Analysis of Paired Data

In Sections 10.1 and 10.2, we considered estimating or testing for a difference between two means μ_1 and μ_2. This was done by utilizing the results of a random sample $X_1, ..., X_m$ from the distribution with mean μ_1 and a completely independent (of the X's) sample $Y_1, ..., Y_n$ from the distribution with mean μ_2. That is, either m individuals were selected from population 1 and n different individuals from population 2, or m individuals were given one treatment and another n individuals were given the other treatment. In contrast, there are a number of experimental situations in which there is only one set of n individuals or experimental objects, and two observations are made on each individual or object, resulting in a natural pairing of values.

Example 10.9 Homes are typically appraised before sale. Appraisers hired by lenders such as banks have an incentive to assign a higher value to a house (so the home loan will be larger), while borrowers' appraisers might be inclined to value the same house at a lower price. The article "Distressed Properties: Valuation Bias and Accuracy" (*J. Real Estate Fin. Econ.* 2010) describes a study in which a random sample of 20 residential properties being purchased in New Orleans after foreclosure was selected. Each property was appraised both by the borrower and by the lender, resulting in the following data (thousands of dollars).

House	1	2	3	4	5	6	7	8
Lender's appraisal	24.3	31.1	108.5	20.0	58.2	23.6	38.7	54.2
Borrower's appraisal	18.6	21.8	98.1	10.2	50.2	15.7	29.8	45.5

House	9	10	11	12	13	14	15	16
Lender's appraisal	21.3	145.3	123.4	171.0	41.2	123.1	47.4	26.1
Borrower's appraisal	14.6	135.8	111.4	156.5	31.2	109.7	39.7	18.6

House	17	18	19	20
Lender's appraisal	76.9	52.5	101.2	33.6
Borrower's appraisal	67.5	42.2	90.0	26.4

Figure 10.4 displays a plot of this data. At first glance, it appears that lenders' appraisals are perhaps a little higher on average than borrowers', but there is a great deal of variability in both samples. So, perhaps any differences between the samples can be attributed to this variability.
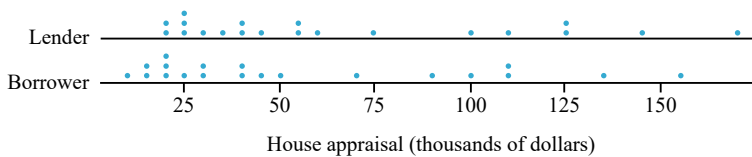
Figure 10.4 Plot of original data from Example 10.9

However, looking back at the original data, a clearer picture emerges: *for every single house*, the lender's appraisal exceeds the borrower's appraisal. Figure 10.5 displays the *difference* in appraised value (lender's appraisal minus borrower's appraisal) for these 20 homes. As we will see, a correct analysis of this data focuses on these differences.

Difference in appraisals (Lender minus Borrower, thousands of dollars)
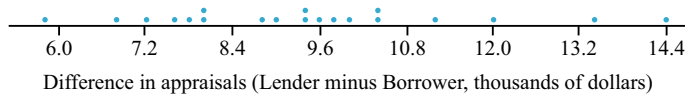
Figure 10.5 Plot of differences from Example 10.9   ∎

ASSUMPTIONS	The data consists of n independently selected pairs (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n), with $E(X_i) = \mu_1$ and $E(Y_i) = \mu_2$. Let $D_1 = X_1 - Y_1$, …, $D_n = X_n - Y_n$, so the D_i's are the differences within pairs. Then the D_i's are assumed to be normally distributed with mean value μ_D and standard deviation σ_D.

We are again interested in hypothesis testing or estimation for the difference $\mu_1 - \mu_2$. The denominator of the two-sample t statistic was obtained by first applying the rule $V(\overline{X} - \overline{Y}) = V(\overline{X}) + V(\overline{Y})$. However, with paired data, the X and Y observations within each pair are often not independent, so $\overline{X}$ and $\overline{Y}$ are not independent of each other, and the rule is not valid. We must therefore abandon the two-sample t procedures and look for an alternative method of analysis.

A Confidence Interval for μ_D

Because different pairs are independent, the D_i's are independent of each other. If we let $D = X - Y$, where X and Y are the first and second observations, respectively, within a randomly selected pair, then the expected difference is

$$\mu_D = E(X - Y) = E(X) - E(Y) = \mu_1 - \mu_2$$

(recall that linearity of expectation is valid even when X and Y are dependent). Thus a confidence interval for μ_D is equivalent to one for $\mu_1 - \mu_2$. An analogous comment applies to a test of hypotheses. But since the D_i's constitute a normal random sample (of differences) with mean μ_D, inferences about μ_D can be performed using one-sample t procedures from Chapters 8 and 9. That is, *to draw conclusions about $\mu_1 - \mu_2$ when data is paired, form the differences D_1, D_2,…, D_n and carry out a one-sample t procedure, based on $n - 1$ df, on the D_i's.*

Let $\overline{D}$ and S_D denote the sample mean and standard deviation, respectively, of the n paired differences D_1, …, D_n. In the same way that the t CI for a single population mean μ is based on the t variable $T = (\overline{X} - \mu)/(S/\sqrt{n})$, a t confidence interval for μ_D $(= \mu_1 - \mu_2)$ is based on the fact that

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \tag{10.4}$$

has a t distribution with $n - 1$ df. Manipulation of this t variable, as in previous derivations of CIs, yields the following interval.

PAIRED
t INTERVAL

The **paired t CI for μ_D** with confidence level $100(1 - \alpha)$% has endpoints

$$\bar{d} \pm t_{\alpha/2,n-1} \cdot \frac{s_D}{\sqrt{n}}$$

where $\bar{d}$ and s_D are the observed values of the sample mean and standard deviation of the paired differences. A one-sided confidence bound results from retaining the relevant sign (+ or −) and replacing $t_{\alpha/2}$ by t_{α}.

When n is small, the validity of this interval requires that the distribution of differences be at least approximately normal. For large n, the CLT ensures that the interval is at least approximately valid without any restrictions on the distribution of differences.

Example 10.10 Adding computerized medical images to a database promises to provide great resources for physicians. However, there are other methods of obtaining such information, so the issue of efficiency of access needs to be investigated. The article "The Comparative Effectiveness of Conventional and Digital Image Libraries" (*J. Audio Media Med.* 2001: 8–15) reported on an experiment in which 13 computer-proficient medical professionals were timed both while retrieving an image from a library of slides and while retrieving the same image from a computer database with a Web front end.

Subject	1	2	3	4	5	6	7	8	9	10	11	12	13
Slide	30	35	40	25	20	30	35	62	40	51	25	42	33
Digital	25	16	15	15	10	20	7	16	15	13	11	19	19
Difference	5	19	25	10	10	10	28	46	25	38	14	23	14

Let μ_D denote the true mean difference between slide retrieval time (sec) and digital retrieval time. Using the paired t confidence interval to estimate μ_D requires that the difference distribution be at least approximately normal. The slight curvature in the normal probability plot from JMP (Figure 10.6) isn't enough to invalidate the normality assumption.
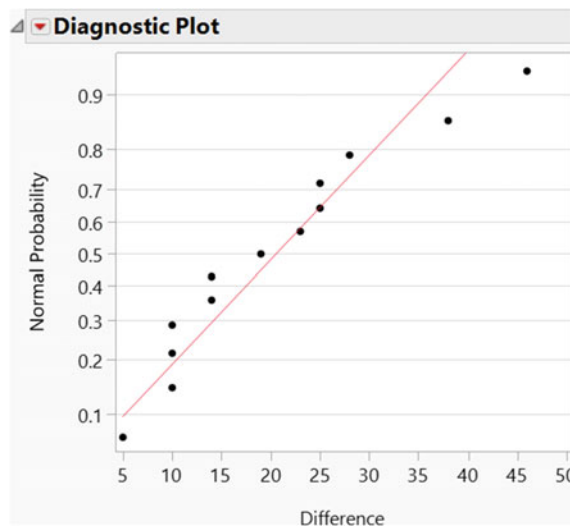
Figure 10.6 Normal probability plot of the differences in Example 10.10

For the 13 differences, $\bar{d} = 20.5$ and $s_D = 11.96$. The t critical value required for a 95% confidence level is $t_{.025,12} = 2.179$, and the 95% CI is

$$\bar{d} \pm t_{\alpha/2, n-1} \cdot \frac{s_D}{\sqrt{n}} = 20.5 \pm 2.179 \cdot \frac{11.96}{\sqrt{13}} = 20.5 \pm 7.2 = (13.3, 27.7)$$

Thus we can be highly confident (at the 95% confidence level) that $13.3 < \mu_D < 27.7$. This interval of plausible values is rather wide, a consequence of the sample standard deviation being large relative to the sample mean. A sample size much larger than 13 would be required to estimate with substantially more precision. Notice, however, that 0 lies well outside the interval, suggesting that $\mu_D > 0$; this is confirmed by a formal hypothesis test. We can conclude from the experiment that computer retrieval appears to be faster on average. ∎

The Paired t Test

Hypothesis testing for paired data also involves calculating the n paired differences and working with that single sample of values. The T variable in (10.4) forms the basis for such tests. And since $\mu_D = \mu_1 - \mu_2$, any hypothesis about the mean difference is equivalent to a hypothesis about the difference between means.

PAIRED t TEST	Null hypothesis: H_0: $\mu_D = \Delta_0$ Test statistic value: $t = \dfrac{\bar{d} - \Delta_0}{s_D/\sqrt{n}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
H_a: $\mu_D > \Delta_0$	$t \geq t_{\alpha, n-1}$
H_a: $\mu_D < \Delta_0$	$t \leq -t_{\alpha, n-1}$
H_a: $\mu_D \neq \Delta_0$	either $t \geq t_{\alpha/2, n-1}$ or $t \leq -t_{\alpha/2, n-1}$

A P-value can be calculated as was done for earlier t tests.

Example 10.11 Musculoskeletal neck-and-shoulder disorders are all too common among office staff who perform repetitive tasks using visual display units. The article "Upper-Arm Elevation During Office Work" (*Ergonomics* 1996: 1221–1230) reported on a study to determine whether more varied work conditions would have any impact on arm movement. The accompanying data was obtained from a sample of $n = 16$ subjects. Each observation is the amount of time, expressed as a proportion of total time observed, during which arm elevation was below 30°. The two measurements from each subject were obtained 18 months apart. During this period, work conditions were changed, and subjects were allowed to engage in a wider variety of work tasks. Does the data suggest that true average time during which elevation is below 30° differs after the change from what it was before the change? This particular angle is important because in Sweden, where the research was conducted, workers' compensation regulations assert that arm elevation less than 30° is not harmful.

Subject	1	2	3	4	5	6	7	8
Before	81	87	86	82	90	86	96	73
After	78	91	78	78	84	67	92	70
Difference	3	−4	8	4	6	19	4	3
Subject	9	10	11	12	13	14	15	16
Before	74	75	72	80	66	72	56	82
After	58	62	70	58	66	60	65	73
Difference	16	13	2	22	0	12	−9	9

Figure 10.7 shows a normal probability plot of the 16 differences; the pattern in the plot is quite straight, supporting the normality assumption. A boxplot of these differences appears in Figure 10.8; the box is located considerably to the right of zero, suggesting that perhaps $\mu_D > 0$ (note also that 13 of the 16 differences are positive and only two are negative).

Average: 6.75
Std Dev: 8.23408
N of data: 16

W-test for Normality
R   0.9916
p value (approx): > 0.1000
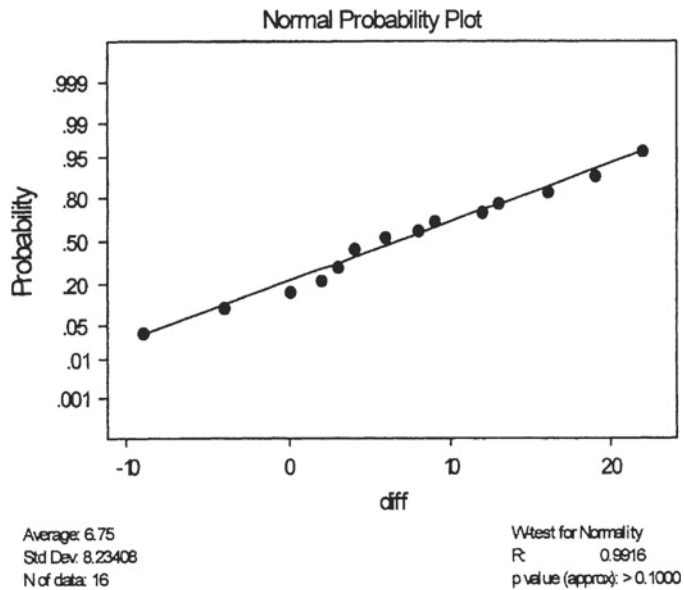
Figure 10.7   A normal probability plot from Minitab of the differences in Example 10.11
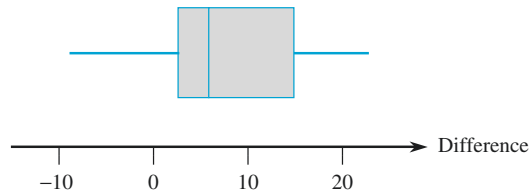
Figure 10.8   A boxplot of the differences in Example 10.11

Let's now use the recommended sequence of steps (refer back to Example 9.20) to test the appropriate hypotheses using the *P*-value method.

1. Let μ_D denote the true average difference between elevation time before the change in work conditions and time after the change.
2. H_0: $\mu_D = 0$ (there is no difference between true average time before the change and true average time after the change)

   H_a: $\mu_D \neq 0$ (there is a difference)
3. The paired *t* test requires data from a normal population. Figure 10.7 validates the plausibility of this assumption.
4. $t = \dfrac{\bar{d} - 0}{s_D/\sqrt{n}} = \dfrac{\bar{d}}{s_D/\sqrt{n}}$
5. From the $n = 16$ differences, $\bar{d} = 6.75$ and $s_D = 8.234$, so

$$t = \frac{6.75}{8.234/\sqrt{16}} = 3.28 \approx 3.3$$

6. Appendix Table A.7 shows that the area to the right of 3.3 under the *t* curve with 15 df is .002. The inequality in H_a implies that a two-tailed test is appropriate, so the *P*-value is approximately $2(.002) = .004$ (software gives .0051).
7. Since $.004 \leq .01$, the null hypothesis can be rejected at either significance level .05 or .01. It does appear that the true average difference between times is something other than zero; that is, true average time after the change is different from that before the change. Recalling that arm elevation should be kept under 30°, we can conclude that the situation became worse because the amount of time below 30° decreased.   ∎

Paired *t* Versus Two-Sample *t* Procedures

Consider using the two-sample *t* test on paired data. The numerators of the paired *t* and two-sample *t* test statistics are identical, since

$$\bar{d} = \frac{1}{n}\sum d_i = \frac{1}{n}\sum (x_i - y_i) = \frac{1}{n}\sum x_i - \frac{1}{n}\sum y_i = \bar{x} - \bar{y}$$

The difference between the two test statistics is due entirely to the denominators. Each test statistic is obtained by standardizing $\overline{X} - \overline{Y}\ (=\overline{D})$, but in the presence of dependence the two-sample *t* standardization is incorrect. To see this, recall from Section 5.3 that

$$V(X \pm Y) = V(X) + V(Y) \pm 2\mathrm{Cov}(X, Y)$$

Since the correlation between X and Y is $\rho = \mathrm{Corr}(X, Y) = \mathrm{Cov}(X, Y)/[\sqrt{V(X)} \cdot \sqrt{V(Y)}]$, it follows that

$$V(X - Y) = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$$

Applying this to $\overline{X} - \overline{Y}$ yields

$$V(\overline{X} - \overline{Y}) = V(\overline{D}) = V\left(\frac{1}{n}\sum D_i\right) = \frac{V(D_i)}{n} = \frac{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}{n}$$

The two-sample t test is based on the assumption of independence, in which case $\rho = 0$. But in many paired experiments, there will be a strong *positive* dependence between X and Y (large X associated with large Y), so that ρ will be positive and the variance of $\overline{X} - \overline{Y}$ will be smaller than $\sigma_1^2/n + \sigma_2^2/n$. Thus whenever there is positive dependence within pairs, the denominator for the paired t statistic should be *smaller* than for t of the independent-samples test, resulting in a larger test statistic and a smaller P-value.

Similarly, when data is paired, the paired t CI will usually be narrower than the (incorrect) two-sample t CI. This is because there is typically much less variability in the differences than in the x and y values.

The paired t and two-sample t procedures described in this section and the previous section apply when we want to compare two populations, treatments, or conditions based upon a quantitative measurement (profit, sales, time, etc.). Many situations exist in which researchers can design their study using their choice of either "matched pairs" or two independent samples. However, once that design decision is made, *only one analysis procedure is correct.* In Examples 10.9–10.11, it would be wrong to use the two-sample t procedures from Section 10.2, since they are predicated on having two independent random samples of data. Similarly, even when investigators gather two independent samples of the same size ($m = n$), the paired t procedures would not be appropriate because no natural pairing would exist between the individuals in sample #1 and the unrelated individuals in sample #2.

Sometimes, as in our examples, paired data results from two observations being taken on the same individual or object. Even when this cannot be done, paired data with dependence within pairs can be obtained by *matching* on one or more characteristics thought to influence responses. For example, in a pharmaceutical study to compare the efficacy of two drugs for lowering blood pressure, the experimenter's budget might allow for the treatment of 100 patients. If 50 patients are randomly selected for treatment with the first drug and another 50 independently selected for treatment with the second drug, an independent-samples experiment results.

However, the experimenter, knowing that blood pressure is influenced by age and weight, might decide to create pairs of patients so that within each of the resulting 50 pairs, age and weight were approximately equal (though there might be sizable differences *between* pairs). Then the two drugs would be randomly assigned to the subjects *within each pair*, for a total of 50 observations on each drug. The benefit of this matching (or "blocking") is that we can account for unwanted sources of variation (e.g., age and weight) that might otherwise have masked differences in the two treatments.

Exercises: Section 10.3 (43–55)

43. The Weaver–Dunn procedure with a fiber mesh tape augmentation is commonly used to treat AC joint (a joint in the shoulder) separations requiring surgery. The article "TightRope Versus Fiber Mesh Tape Augmentation of Acromioclavicular Joint Reconstruction" (*Am. J. Sport Med.* 2010: 1204–1208) described the investigation of a new method which was hypothesized to provide superior stability (less movement) compared to the W–D procedure. The authors of the cited article kindly provided the accompanying data on anteposterior (forward–backward) movement (mm) for six matched pairs of shoulders:

Subject	1	2	3	4	5	6
Fiber mesh	20	30	20	32	35	33
TightRope	15	18	16	19	10	12

Carry out a test of hypotheses at significance level .01 to see if true average movement for the TightRope treatment is

indeed less than that for the Fiber Mesh treatment. Be sure to check any assumptions underlying your analysis.

44. Hexavalent chromium has been identified as an inhalation carcinogen and an air toxin of concern in a number of different locales. The article "Airborne Hexavalent Chromium in Southwestern Ontario" (*J. Air Waste Manage.* 1997: 905–910) gave the accompanying data on both indoor and outdoor concentration (nanograms/m^3) for a sample of houses selected from a certain region.

House	1	2	3	4	5	6	7	8	9
Indoor	.07	.08	.09	.12	.12	.12	.13	.14	.15
Outdoor	.29	.68	.47	.54	.97	.35	.49	.84	.86

House	10	11	12	13	14	15	16	17
Indoor	.15	.17	.17	.18	.18	.18	.18	.19
Outdoor	.28	.32	.32	1.55	.66	.29	.21	1.02

House	18	19	20	21	22	23	24	25
Indoor	.20	.22	.22	.23	.23	.25	.26	.28
Outdoor	1.59	.90	.52	.12	.54	.88	.49	1.24

House	26	27	28	29	30	31	32	33
Indoor	.28	.29	.34	.39	.40	.45	.54	.62
Outdoor	.48	.27	.37	1.26	.70	.76	.99	.36

a. Calculate a confidence interval for the population mean difference between indoor and outdoor concentrations using a confidence level of 95%, and interpret the resulting interval.

b. If a 34th house was to be randomly selected from the population, between what values would you predict the difference in concentrations to lie?

45. Shoveling is not exactly a high-tech activity, but will continue to be a required task even in our information age. The article "A Shovel with a Perforated Blade Reduces Energy Expenditure Required for Digging Wet Clay" (*Hum. Factors* 2010: 492–502) reported on an experiment in which each of 13 workers was provided with both a conventional shovel and a shovel whose blade was perforated with small holes. The authors of the cited article provided the following data on stable energy expenditure [kcal/kg(subject)/lb(clay)]:

Worker	1	2	3	4	5	6	7
Conventional	.0011	.0014	.0018	.0022	.0010	.0016	.0028
Perforated	.0011	.0010	.0019	.0013	.0011	.0017	.0024

Worker	8	9	10	11	12	13
Conventional	.0020	.0015	.0014	.0023	.0017	.0020
Perforated	.0020	.0013	.0013	.0017	.0015	.0013

a. Calculate a confidence interval at the 95% confidence level for the true average difference between energy expenditure for the conventional shovel and the perforated shovel (a normal probability plot of the sample differences shows a reasonably linear pattern). Based on this interval, does it appear that the shovels differ with respect to true average energy expenditure? Explain.

b. Carry out a test of hypotheses at significance level .05 to see if true average energy expenditure using the conventional shovel exceeds that using the perforated shovel; include a *P*-value in your analysis.

46. The article "Effect of Wearable Technology Combined With a Lifestyle Intervention on Long-term Weight Loss" (*JAMA* 2017: 1161–1171) describes a study in which adults in a large cohort were provided the same weight-loss regimen for six months. Then, participants were randomly assigned to either (1) self-monitor diet and physical activity using a Web site or (2) track diet and physical activity with a wearable device and accompanying Web interface. The weights of all subjects (in kg) were recorded at the beginning of the study and 24 months later. The following summary is consistent with information in the article.

Treatment			Baseline	24 Months	Difference
(1)	$n = 233$	Mean:	95.2	89.3	–5.9
		SD:	16.4	15.1	6.9
(2)	$n = 237$	Mean:	96.3	92.8	–3.5
		SD:	16.5	15.7	7.3

a. Construct and interpret a 95% confidence interval for the population mean weight loss under the first treatment (self-monitoring with a Web site).
b. Construct and interpret a 95% confidence interval for the population mean weight loss under the second treatment (tracking with a wearable device).
c. How confident can you be that *both* the intervals in (a) and (b) contain the values of population mean weight loss?
d. Does the data show at the .05 level that the two population means in parts (a) and (b) are different? Perform the appropriate hypothesis test. [*Note*: This does *not* require a paired t procedure!]

47. Refer back to Example 10.9. Here are the differences in appraisals displayed in Figure 10.5:

5.7	9.3	10.4	9.8	8.0	7.9	8.9	8.7	6.7	9.5
12.0	14.5	10.0	13.4	7.7	7.5	9.4	10.3	11.2	7.2

a. Construct a normal probability plot of these 20 differences. Is it plausible that the population distribution of differences is normal?
b. Construct and interpret a 95% upper confidence bound for the true mean difference in appraised home value.
c. Test the hypothesis that the true mean difference in appraised home value is less than $10,000 at the .05 level. Is your answer consistent with part (b)?

48. Management at a large retail appliance chain required all full-time sales staff to attend a one-day training session on improving sales technique. To evaluate the effectiveness of this rather expensive training, the number of sales in the week prior to the training and the number of sales in the week following the training was recorded for each salesperson. Data for the 10 full-times salespersons at one branch of the store appears in the accompanying table.

Salesperson	1	2	3	4	5	6	7	8	9	10
After sales training	44	53	30	41	53	63	55	68	35	41
Before sales training	50	45	25	40	45	55	40	54	33	49
Difference	–6	8	5	1	8	8	15	14	2	–8

Does the data provide convincing statistical evidence that, on average, employees make a greater number of weekly sales after the training? Test the appropriate hypotheses at the .01 level. Validate any necessary conditions.

49. The article "Less Is Better: When Low-value Options Are Valued More Highly than High-value Options" (*J. Behav. Decis. Making* 1998: 107–121) describes several experiments pertaining to consumer behavior. In one experiment, 46 students were split into two groups: 23 who were shown 7 oz of ice cream in a 5-oz cup (the cup was overflowing) and 23 who were shown 8 oz of ice cream in a 10-oz cup (there was a lot of empty space left over). Each student was then asked, "What is the most you are willing to pay for a serving?" The researchers theorized that students would pay more, on average, for the overflowing cup even though it contained less ice cream.

a. Which is the correct method of analysis for this situation: the paired t test, or the two-sample t procedure from the previous section? Why?
b. The sample averages for the 7 oz and 8 oz groups were $2.26 and $1.66, respectively; information in the article suggests the corresponding standard deviations are $0.84 and $0.81, respectively. Test the researchers' hypothesis at the $\alpha = .05$ level. Indicate any assumptions required for your method to be valid.

c. In a second experiment, a different group of 23 students were shown both of the aforementioned ice cream cups, side by side. Each student then indicated how much s/he was willing to pay for each ice cream serving. Which is the correct method of analysis for this second experiment: the paired t test, or the two-sample t procedure from the previous section? Why?

d. It is hypothesized that under this condition, students will be willing to pay more for the 8 oz serving. Test this hypothesis at the .05 level using the following information: sample average for 7 oz serving = $1.56; sample average for 8 oz serving = $1.85; standard deviation of sample differences = $0.32. Indicate any assumptions required for your method to be valid.

e. Can you explain why the two experiments gave "opposite" results? [*Hint:* This is not a statistics question.]

50. The article discussed in the previous exercise also describes an experiment in which 35 students were asked to price two boxes of silverware: a 24-piece box with all 24 pieces intact, and a 40-piece box with only 31 pieces of silverware intact. Each student indicated the amount s/he would be willing to pay for each box. The sample average amount students were willing to pay for the 24-piece and 40-piece boxes were $29.70 and $32.03, respectively. The standard deviation of the differences was $6.41. Test the hypothesis that, on average, students are willing to pay more for the box with more silverware even though it was not completely intact. Use a .05 level of significance.

51. Chapter 1 Exercise 81 describes a study of children's private speech (talking to themselves). The 33 children were each observed in about 100 ten-second intervals in the first grade, and again in the second and third grades. Because private speech occurs more in challenging circumstances, the children were observed while doing their mathematics. The speech was classified as on task (about the math lesson), off task, or mumbling (the observer could not tell what was said). Here are the 33 first-grade mumble scores, followed by the third-grade scores:

20.8	24.4	19.4	33.3	26.0	56.6	39.5	24.7	21.6
19.2	43.0	26.3	22.7	49.4	35.4	56.8	45.4	28.7
34.0	26.9	48.4	27.6	52.6	5.9	38.5	22.1	22.2
32.1	48.1	19.5	42.2	20.3	20.0			

28.8	57.0	23.9	46.9	50.0	64.6	54.2	55.3	21.4
44.3	11.7	58.6	76.1	76.4	48.6	37.2	69.8	29.1
46.5	50.0	69.6	69.8	59.4	22.7	84.9	42.0	67.2
38.3	78.5	38.1	60.4	57.8	38.7			

The numbers are in the same order for each grade; for example, the third student mumbled in 19.4% of the intervals in the first grade and 23.9% of the intervals in the third grade.

a. Verify graphically that normality is plausible for the population distribution of differences.

b. Find a 95% confidence interval for the difference of population means, and interpret the result.

52. Can people operate touch screen devices more quickly with their index finger or their thumb? Holding the device in landscape or portrait position? The article "Evaluation of a Psychomotor Vigilance Task (PVT) for Touch Screen Devices" (*Hum. Factors* 2017: 661–670) describes a study in which 13 participants performed a series of tasks on an iPod holding it two different ways: (1) in portrait position using their index finger and (2) in landscape position using their thumb. The median response time, in

milliseconds, was recorded for each participant under both settings. Those observations are summarized in the accompanying table.

iPod position	Mean	SD
Portrait/index	224.09	39.30
Landscape/thumb	211.50	30.74
Difference	12.59	15.28

Test whether the results under the two positions are significantly different at the .05 significance level.

53. It has been estimated that between 1945 and 1971, as many as 2 million children were born to mothers treated with diethylstilbestrol (DES), a nonsteroidal estrogen recommended for pregnancy maintenance. The FDA banned this drug in 1971 because research indicated a link with the incidence of cervical cancer. The article "Effects of Prenatal Exposure to Diethylstilbestrol (DES) on Hemispheric Laterality and Spatial Ability in Human Males" (*Hormones Behav.* 1992: 62–75) discussed a study in which 10 males exposed to DES and their unexposed brothers underwent various tests. This is the summary data on the results of a spatial ability test: $\bar{x} = 12.6$ (exposed), $\bar{y} = 13.7$, and standard error of mean difference = .5. Test at level .05 to see whether exposure is associated with reduced spatial ability by obtaining the *P*-value.

54. As integrated circuits operate at ever-smaller resolutions, the clean handling of wafers in the manufacturing process has become even more important. The article "Particle Free Handling of Substrates"

(*IEEE Trans. Semicond. Manuf.* 2016: 314–319) provides the following data on the number of particles detected pre- and post-handling for a sample of 16 wafers:

Pre	5	5	32	14	2	2	17	13
Post	236	684	1256	3605	40	92	173	44
Diff.	231	679	1224	3591	38	90	156	31

Pre	18	27	5	18	52	20	17	35
Post	51	88	189	610	124	1218	2023	3057
Diff.	33	61	184	592	72	1198	2006	3022

a. The researchers desired a confidence interval for μ_D, the average increase in number of particles per wafer due to handling. Why should the paired *t* interval *not* be applied here? [*Hint:* Construct a normal probability plot of the differences.]

b. A normal probability plot of the logarithms of the difference values shows that the population of $\ln(D)$ values is plausibly normal (i.e., *D* may be lognormal). Take the logarithm of the differences, and use those values to construct a 95% CI for $E[\ln(D)]$.

c. It can be shown that exponentiating the endpoints of the interval from part (b) produces a confidence interval not for μ_D, but rather the population median $\tilde{\mu}_D$. Exponentiate the limits of the interval from part (b), and interpret this interval.

55. Construct a paired data set for which $t = \infty$, so that the data is highly significant when the correct analysis is used, yet *t* for the two-sample *t* test is quite near zero, so the incorrect analysis yields an insignificant result.

10.4   Inferences About Two Population Proportions

Having presented methods for comparing the means of two different populations, we now turn to the comparison of two population proportions. The notation for this scenario is an extension of the notation used in the corresponding one-population problem. Let p_1 and p_2 denote the proportions of individuals in populations 1 and 2, respectively, who possess a particular characteristic. Equivalently, if we use the label S (success) for an individual who possesses the characteristic of interest—favors a particular proposition, has read at least one book within the last month, etc.—then p_1 and p_2 represent the probabilities of seeing the label S on a randomly chosen individual from populations 1 and 2, respectively.

We will assume the availability of a sample of m individuals from the first population and n from the second. The variables X and Y will represent the number of individuals in each sample possessing the characteristic that defines p_1 and p_2. Provided the population sizes are much larger than the sample sizes, the distribution of X can be taken to be binomial with parameters m and p_1, and similarly, $Y \sim \text{Bin}(n, p_2)$. Furthermore, the samples are assumed to be independent of each other, so that X and Y are independent rvs.

The obvious estimator for $p_1 - p_2$, the difference in population proportions, is the corresponding difference in sample proportions. With $\hat{P}_1 = X/m$ and $\hat{P}_2 = Y/n$, the estimator of $p_1 - p_2$ can be expressed as $\hat{P}_1 - \hat{P}_2 = X/m - Y/n$.

PROPOSITION   Let $X \sim \text{Bin}(m, p_1)$ and $Y \sim \text{Bin}(n, p_2)$ with X and Y independent variables. Define $\hat{P}_1 = X/m$ and $\hat{P}_2 = Y/n$. Then

$$E(\hat{P}_1 - \hat{P}_2) = p_1 - p_2,$$

so $\hat{P}_1 - \hat{P}_2$ is an unbiased estimator of $p_1 - p_2$, and

$$V(\hat{P}_1 - \hat{P}_2) = \frac{p_1 q_1}{m} + \frac{p_2 q_2}{n} \quad \left(\text{where } q_i = 1 - p_i\right)$$

Proof   Since $E(X) = mp_1$ and $E(Y) = np_2$,

$$E\left(\frac{X}{m} - \frac{Y}{n}\right) = \frac{1}{m}E(X) - \frac{1}{n}E(Y) = \frac{1}{m}mp_1 - \frac{1}{n}np_2 = p_1 - p_2$$

Since $V(X) = mp_1 q_1$, $V(Y) = np_2 q_2$, and X and Y are independent,

$$V\left(\frac{X}{m} - \frac{Y}{n}\right) = V\left(\frac{X}{m}\right) + (-1)^2 V\left(\frac{Y}{n}\right) = \frac{1}{m^2}V(X) + \frac{1}{n^2}V(Y) = \frac{p_1 q_1}{m} + \frac{p_2 q_2}{n} \qquad \blacksquare$$

We will focus first on situations in which both m and n are large. Then because $\hat{P}_1$ and $\hat{P}_2$ individually have approximately normal distributions, the estimator $\hat{P}_1 - \hat{P}_2$ also has approximately a normal distribution. Standardizing $\hat{P}_1 - \hat{P}_2$ yields a variable Z whose distribution is approximately standard normal:

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1 q_1}{m} + \dfrac{p_2 q_2}{n}}}$$

A Large-Sample Test Procedure

Analogous to the hypotheses for $\mu_1 - \mu_2$, the most general null hypothesis an investigator might consider would be of the form $H_0: p_1 - p_2 = \Delta_0$, where Δ_0 is again a specified number. Although for population means the case $\Delta_0 \neq 0$ presented no difficulties, for population proportions the cases $\Delta_0 = 0$ and $\Delta_0 \neq 0$ must be considered separately. Since the vast majority of actual problems of this sort involve $\Delta_0 = 0$ (i.e., the null hypothesis $p_1 = p_2$), we will concentrate on this case. When $H_0: p_1 - p_2 = 0$ is true, let p denote the common value of p_1 and p_2 (and similarly for q). Then the standardized variable

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - 0}{\sqrt{pq\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}}$$

has approximately a standard normal distribution when H_0 is true. However, this Z cannot serve as a test statistic because the value of p is unknown—H_0 asserts only that there is a common value of p, but H_0 does not say what that value is. To obtain a usable test statistic having approximately a standard normal distribution when H_0 is true, p must be estimated from the sample data.

Assuming then that $p_1 = p_2 = p$, instead of separate samples of size m and n from two different populations (two different binomial distributions), we really have a single sample of size $m + n$ from one population with proportion p. Since the total number of individuals in this combined sample having the characteristic of interest is $X + Y$, the estimator of p is

$$\hat{P} = \frac{X + Y}{m + n} = \frac{m}{m + n}\hat{P}_1 + \frac{n}{m + n}\hat{P}_2 \tag{10.5}$$

The second expression for $\hat{P}$ shows that it is actually a weighted average of estimators $\hat{P}_1$ and $\hat{P}_2$ obtained from the two samples. If we take (10.5) and substitute back into Z with $\hat{Q} = 1 - \hat{P}$, the resulting statistic has approximately a $N(0, 1)$ distribution when H_0 is true.

**TWO-PROPORTION
z TEST**

Null hypothesis: $H_0: p_1 - p_2 = 0$

Test statistic value (large samples): $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}}$

Alternative Hypothesis	Rejection Region for Approximate Level α Test
$H_a : p_1 - p_2 > 0$	$z \geq z_\alpha$
$H_a : p_1 - p_2 < 0$	$z \leq -z_\alpha$
$H_a : p_1 - p_2 \neq 0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

A P-value is calculated in the same way as for previous z tests.
These procedures are valid provided that $n_1\hat{p}_1 \geq 10$, $n_1\hat{q}_1 \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2\hat{q}_2 \geq 10$.

Example 10.12   Are customers more willing to buy a product or service just because they're offered multiple purchase plans? In a study published in *Land Econ.* (2006), home-owning residents of

Madison, WI were offered the chance to buy wind-generated electricity for their homes as part of a pilot program in the city. A random sample of 237 residents was offered seven different purchase plans (ranging from 50 kWh to 600 kWh per month; the extra monthly cost of wind-generated power was about $1 for 25 kWh). An independent random sample of 649 residents was offered just a single choice for the amount they could purchase, with that amount randomly selected for each resident from among the same seven options. Thirty-seven percent of those offered multiple options purchased wind-generated electricity, compared to 24% of the single-option group.

Does this data suggest that customers are more willing to buy wind-generated electricity when they're offered multiple purchase plans, or could the disparity be attributed to chance? Test at the $\alpha = .01$ level.

1. The population being studied consists of all homeowners in Madison, WI. Within this population, the parameter of interest is $p_1 - p_2$, the difference in the proportions who would buy wind-generated electricity under the multiple-option scheme and the single-option scheme.
2. Researchers believed a priori that customers were more likely to buy in under the first scheme, so the competing hypotheses are

$$H_0: p_1 - p_2 = 0 \text{ (i.e., } p_1 = p_2)$$
$$H_a: p_1 - p_2 > 0 \text{ (i.e., } p_1 > p_2)$$

3. The data consists of two independent random samples. From the numbers provided, $n_1 \hat{p}_1 = (237)(.37) \approx 88$, $n_1 \hat{q}_1 \approx 149$, $n_2 \hat{p}_2 = (649)(.24) \approx 156$, and $n_2 \hat{q}_2 \approx 493$; all of these values are at least 10. The requirements for this large-sample z hypothesis test are satisfied.
4. The two-proportion z test statistic value is $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}}$.
5. For this upper-tailed test, we reject H_0 if $z \geq z_{.01} = 2.33$.
6. The combined (pooled) estimate of the common proportion p under H_0 is

$$\hat{p} = \frac{237}{237 + 649}(.37) + \frac{649}{237 + 649}(.24) = .275,$$

which results in a test statistic value of

$$z = \frac{(.37) - (.24)}{\sqrt{(.275)(.725)\left[\dfrac{1}{237} + \dfrac{1}{649}\right]}} = 3.84$$

That is, the observed value of $\hat{P}_1 - \hat{P}_2$ is almost four standard deviations larger than what we'd expect if H_0 were true.

7. Since $3.84 \geq 2.33$, H_0 is rejected at the .01 significance level. The data very strongly suggests that Madison homeowners are *more* likely to purchase wind-generated electricity if they are offered several options for the amount of electricity they can buy.

Using a P-value approach, based on the direction of H_a, the P-value equals $1 - \Phi(3.84) = .0003$. If residents were equally likely to buy wind-generated electricity under both schemes, the chance of observing a disparity at least as large as the one in this study would be extremely small. Hence, again, we would reject H_0 in favor of the alternative hypothesis. ∎

Power, β, and Sample Sizes

Here the determination of power and β is a bit more cumbersome than it was for other large-sample tests. The reason is that the denominator of Z is an estimate of the standard deviation of $\hat{P}_1 - \hat{P}_2$ assuming that $p_1 = p_2 = p$. When H_0 is false, $\hat{P}_1 - \hat{P}_2$ must be re-standardized using

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{p_1 q_1}{m} + \frac{p_2 q_2}{n}} \tag{10.6}$$

The form of σ in (10.6) implies that power and β are functions of both p_1 and p_2, not just the difference $p_1 - p_2$. So we denote the chance of a type II error by $\beta(p_1, p_2)$.

Alternative Hypothesis	$\boldsymbol{\beta(p_1, p_2)}$

H_a: $p_1 - p_2 > 0$

$$\Phi\left[\frac{z_\alpha \sqrt{\bar{p}\bar{q}\left(\frac{1}{m} + \frac{1}{n}\right)} - (p_1 - p_2)}{\sigma}\right]$$

H_a: $p_1 - p_2 < 0$

$$1 - \Phi\left[\frac{-z_\alpha \sqrt{\bar{p}\bar{q}\left(\frac{1}{m} + \frac{1}{n}\right)} - (p_1 - p_2)}{\sigma}\right]$$

H_a: $p_1 - p_2 \neq 0$

$$\Phi\left[\frac{z_{\alpha/2} \sqrt{\bar{p}\bar{q}\left(\frac{1}{m} + \frac{1}{n}\right)} - (p_1 - p_2)}{\sigma}\right] - \Phi\left[\frac{-z_{\alpha/2} \sqrt{\bar{p}\bar{q}\left(\frac{1}{m} + \frac{1}{n}\right)} - (p_1 - p_2)}{\sigma}\right]$$

where $\bar{p} = (mp_1 + np_2)/(m+n)$, $\bar{q} = (mq_1 + nq_2)/(m+n)$, and σ is given by (10.6). For each case, power $= 1 - \beta$.

Alternatively, for specified p_1 and p_2, the sample sizes necessary to achieve $\beta(p_1, p_2) = \beta$ can be determined. For example, for the upper-tailed test, we equate $-z_\beta$ to the argument of $\Phi(\cdot)$ (i.e., what's inside the parentheses) in the foregoing box. If $m = n$, there is a simple expression for the common value:

$$n = \frac{\left[z_\alpha \sqrt{(p_1 + p_2)(q_1 + q_2)/2} + z_\beta \sqrt{p_1 q_1 + p_2 q_2}\right]^2}{(p_1 - p_2)^2} \tag{10.7}$$

for an upper- or lower-tailed test, with $\alpha/2$ replacing α for a two-tailed test.

Example 10.13 One of the truly impressive applications of statistics occurred in connection with the design of the 1954 Salk polio vaccine experiment and analysis of the resulting data. Part of the experiment focused on the efficacy of the vaccine in combating paralytic polio. Because it was thought that without a control group of children, there would be no sound basis for assessment of the vaccine, it was decided to administer the vaccine to one group and a placebo injection (visually indistinguishable from the vaccine but known to have no effect) to a control group. For ethical reasons and also because it was thought that the knowledge of vaccine administration might have an effect on treatment and diagnosis, the experiment was conducted in a **double-blind** manner. That is, neither the

individuals receiving injections nor those administering them actually knew who was receiving vaccine and who was receiving the placebo (samples were numerically coded)—remember, at that point it was not at all clear whether the vaccine was beneficial.

Let p_1 and p_2 be the probabilities of a child getting paralytic polio for the control and treatment conditions, respectively. The objective was to test H_0: $p_1 - p_2 = 0$ versus H_a: $p_1 - p_2 > 0$ (the alternative hypothesis states that a vaccinated child is less likely to contract polio than an unvaccinated child). Supposing the true value of p_1 is .0003 (an incidence rate of 30 per 100,000), the vaccine would be a significant improvement if the incidence rate was halved—that is, $p_2 = .00015$. Using a level $\alpha = .05$ test, it would then be reasonable to ask for sample sizes for which power = 90% (i.e., $\beta = .1$) when $p_1 = .0003$ and $p_2 = .00015$. Assuming equal sample sizes, the required n is obtained from (10.7) as

$$n = \frac{\left[1.645\sqrt{(.5)(.00045)(.199955)} + 1.28\sqrt{(.00015)(.99985) + (.0003)(.9997)}\right]^2}{(.0003 - .00015)^2}$$

$$= [(.0349 + .0271)/.00015]^2 \approx 171{,}000$$

The actual data for this experiment follows. Sample sizes of approximately 200,000 were used. The reader can easily verify that $z = 6.43$, a highly significant value. The vaccine was judged a resounding success!

Placebo: $m = 201{,}229$ x = number of cases of paralytic polio = 110
Vaccine: $n = 200{,}745$ $y = 33$ ∎

A Large-Sample Confidence Interval for $p_1 - p_2$

As with means, many two-sample problems involve the objective of comparison through hypothesis testing, but sometimes an interval estimate for $p_1 - p_2$ is appropriate. Both $\hat{P}_1 = X/m$ and $\hat{P}_2 = Y/n$ have approximate normal distributions when m and n are both large. If we identify θ with $p_1 - p_2$, then $\hat{\theta} = \hat{P}_1 - \hat{P}_2$ satisfies the conditions necessary for obtaining a large-sample CI (see Section 9.6). In particular, the estimated standard deviation of $\hat{\theta}$ is $\sqrt{(\hat{p}_1\hat{q}_1/m) + (\hat{p}_2\hat{q}_2/n)}$. The $100(1 - \alpha)\%$ interval $\hat{\theta} \pm z_{\alpha/2} \cdot \hat{\sigma}_{\hat{\theta}}$ then becomes the **two-proportion z interval**

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m} + \frac{\hat{p}_2\hat{q}_2}{n}}$$

Like the two-proportion z test, this formula is suitable for large samples. Notice that the estimated standard deviation of $\hat{p}_1 - \hat{p}_2$ (the square root expression) is different here from what it was for hypothesis testing when $\Delta_0 = 0$.

Statistical research has shown that the actual confidence level for the two-proportion z CI can sometimes deviate substantially from the nominal level (the level you think you are getting when you use a particular z critical value—e.g., 95% when $z_{\alpha/2} = 1.96$). A suggested improvement is to add one success and one failure to each of the two samples and then replace the $\hat{p}$'s and $\hat{q}$'s in the foregoing formula by $\tilde{p}$'s and $\tilde{q}$'s where $\tilde{p}_1 = (x+1)/(m+2)$, etc. This adjusted interval can also be used reliably when sample sizes are quite small.

Example 10.14 *Agritourism* (visiting farms and participating in farm activities) is an increasingly large part of the American tourism industry. The authors of "Examination of the Use of E-Marketing by Small Farms in the Northeast," (*J. Food Distrib. Res.* 2006 37(1)) investigated the relationship between agritourism and presence on the Internet. In a survey of 640 farms in the northeastern United States, 261 farms had a Web site for their farm business and 379 did not. Among farms with a Web site, 167 had some type of agritourism activities, compared to 152 of the farms that did not have a business Web site.

Let p_1 = the proportion of all northeastern farms with Web sites that provide agritourism activities and p_2 = the proportion of all northeastern farms without Web sites that provide agritourism activities. With $\hat{p}_1 = 167/261 = .640$ and $\hat{p}_2 = 152/379 = .401$, a 95% confidence interval for $p_1 - p_2$ is

$$(.640 - .401) \pm 1.96\sqrt{\frac{.640(.360)}{261} + \frac{.401(.599)}{379}} = .239 \pm .076 = (.163, .315)$$

We are 95% confident that the *difference* in the proportion of northeastern farms with Web sites that offer agritourism activities and the proportion of northeastern farms without Web sites that offer agritourism activities is between .163 and .315. (Using $\tilde{p}_1 = 168/263$ and $\tilde{p}_2 = 153/381$ based on sample sizes of 263 and 381, respectively, the adjusted interval here is essentially identical to the original interval.)

In particular, the agritourism rate is much higher among those farms which use a business Web site to advertise. We observe here a positive association between having a Web site and providing agritourism activities. One caveat: since this is only an observational study, we cannot conclude that presence on the Internet *causes* farms to make money off agritourism. ∎

Small-Sample Inferences

On occasion an inference concerning $p_1 - p_2$ may have to be based on samples for which at least one sample size is small. Appropriate methods for such situations are not as straightforward as those for large samples, and there is less agreement among statisticians as to recommended procedures.

The main issue here is that $\hat{P}_1 - \hat{P}_2$ is no longer approximately normal when *m* or *n* is small, and no expression exists for the exact distribution of the difference of two (scaled) binomial rvs. Some software packages will "adjust" the data by adding one success and one failure to each sample, as was mentioned briefly in the context of confidence intervals. Alternatively, statistical software can be used to simulate the sampling distribution of $\hat{P}_1 - \hat{P}_2$ using the underlying binomial models, and *P*-values can be estimated therefrom.

One frequently used test in this situation, called *Fisher's exact test*, is based on the hypergeometric distribution. This method has its own deficiencies, as it assumes that both the sample sizes *and* the total number of successes across the two samples are fixed. But Fisher's exact test is the most-commonly used alternative procedure for comparing two proportions from small samples. Please consult an appropriate reference for more information.

Exercises: Section 10.4 (56–70)

56. Independent random samples of 237 African-Americans and 396 Caucasian-Americans were asked to identify their "favorite" category of television programming ("Television Types and TV Attitudes of African-Americans, Latinos, and Caucasians," *J. Advert. Res.* 2008: 235–246). In the survey, 13.1% of African-Americans and 18.7% of Caucasians indicated that they prefer to watch the news more than any other category.

    a. Test the hypothesis that the proportion of all people whose favorite TV viewing category is news differs between the populations of all African-Americans and Caucasians, at the $\alpha = .01$ significance level.
    b. Would your answer to part (a) be different if you used a .10 significance level? Explain.
    c. The same survey found that 33.3% of 213 randomly selected Latinos chose the news as their favorite television program. Repeat part (a), but compare the populations of Latinos and African-Americans.

57. A sample of 300 urban adult residents of a particular state revealed 63 who favored increasing the highway speed limit from 55 to 65 mph, whereas a sample of 180 rural residents yielded 75 who favored the increase. Does this data indicate that the sentiment for increasing the speed limit is different for the two groups of residents?

    a. Test H_0: $p_1 = p_2$ versus H_a: $p_1 \neq p_2$ using $\alpha = .05$, where p_1 refers to the urban population.
    b. If the true proportions favoring the increase are actually $p_1 = .20$ (urban) and $p_2 = .40$ (rural), what is the probability that H_0 will be rejected using a level .05 test with $m = 300$, $n = 180$?

58. It is thought that the front cover and the nature of the first question on mail surveys influence the response rate. The article "The Impact of Cover Design and First Questions on Response Rates for a Mail Survey of Skydivers" (*Leisure Sci.* 1991: 67–76) tested this theory by experimenting with different cover designs. One cover was plain; the other used a picture of a skydiver. The researchers speculated that the return rate would be lower for the plain cover.

Cover	Number sent	Number returned
Plain	207	104
Skydriver	213	109

    Does this data support the researchers' hypothesis? Test the relevant hypotheses using $\alpha = .10$ by first calculating a *P*-value.

59. Do teachers find their work rewarding and satisfying? The article "Work-Related Attitudes" (*Psych. Rep.* 1991: 443–450) reports the results of a survey of 395 elementary school teachers and 266 high school teachers. Of the elementary school teachers, 224 said they were very satisfied with their jobs, whereas 126 of the high school teachers were very satisfied with their work. Estimate the difference between the proportion of all elementary school teachers who are satisfied and all high school teachers who are satisfied by calculating a CI.

60. Several states have an annual "sales tax holiday" to encourage spending. A survey of 695 randomly selected shoppers at a large retail center in Texas asked how important people felt the tax holiday was (*Am. J. Bus.* 2007). 565 shoppers indicated that the tax holiday was important in their decision to shop.

a. Estimate the proportion of all Texas shoppers for whom the tax holiday is important in their decision to shop.

b. In the same study, 195 of 250 men said the tax holiday was important in their decision to shop, compared to 370 of 445 women. Test the hypothesis that women are more likely than men to consider the tax holiday important, at the 5% significance level.

61. The author of the article "Food and Eating on Television: Impacts and Influences" (*Nutrition Food Sci.* 2000: 24–29) examined hundreds of hours of BBC television footage and categorized food images for both TV programs and commercials. Out of 1785 food images in TV programs, 322 showed sugary and/or fatty foods, while 511 out of 1186 commercial food images were sugary and/or fatty.

    a. Construct a 99% CI for the difference in the proportion of food images that include sugary/fatty foods in TV programs and in commercials. Assume these two samples are representative of all food images on the BBC.

    b. What does the CI in part (a) say about the disparity between food images in TV programs and those in commercials?

62. The authors of the article "Adjuvant Radiotherapy and Chemotherapy in Node-Positive Premenopausal Women with Breast Cancer" (*New Engl. J. Med.* 1997: 956–962) reported on the results of an experiment designed to compare treating cancer patients with only chemotherapy to treatment with a combination of chemotherapy and radiation. Of the 154 individuals who received the chemotherapy-only treatment, 76 survived

at least 15 years, whereas 98 of the 164 patients who received the hybrid treatment survived at least that long.

    a. With p_1 denoting the proportion of all such women who, when treated with just chemotherapy, survive at least 15 years and p_2 denoting the analogous proportion for the hybrid treatment, calculate a 99% CI for $p_1 - p_2$.

    b. Based on the interval from part (a), can either treatment be judged superior to the other? Why or why not?

63. The article "Luck of the Draw: Creating Chinese Brand Names" (*J. Advertising Res.* 2008: 523–530) explores the use of "lucky" brand names in China, and whether that use varies by the uncertainty in a brand's business environment (its market sector). In a sample of 654 brands from sectors with low uncertainty, 372 names were considered lucky; among 548 "high-uncertainty" brands, 343 had lucky names. (In Chinese culture, the number of strokes required to write a name determines its luck.) The authors of the article theorized that companies would use "lucky" brand names more often in high-uncertainty business environments. Test the authors' hypothesis at the $\alpha = .05$ significance level. [*Hint:* The two populations are all Chinese brands in low-uncertainty business environments and all Chinese brands in high-uncertainty business environments.]

64. Air travelers often complain that recirculated air in the cabin leads to the spread of colds, while the airline industry generally disputes this claim. A 2002 study in the *J. Am. Med. Assoc.* reported the following information for passengers on flights with and without recirculated air:

	Post-flight respiratory symptoms?	
Recirculated air on the flight?	Yes	No
Yes	111	472
No	108	409

Assume these represent independent random samples from the two relevant populations. Does the data suggest that the likelihood of catching a cold is higher on flights with recirculated air? Test at the $\alpha = .05$ significance level.

65. In a 2017 class project, two team members each approached 50 students on campus (randomly selected using systematic sampling). Each student was asked to participate in a survey, but the survey itself was a ruse: the real goal was to see who would agree to be surveyed by Melissa (who has a British accent) or Kristine (an American accent). In the end, 41 of 50 students agreed to be surveyed by Melissa, while 27 of 50 took Kristine's (fake) survey.

    a. Test the hypothesis of equal population proportions at the .01 significance level. Find the *P*-value for the test, and interpret your results. Be sure to clearly define your parameters!
    b. Can it be concluded that there is a causal relationship between interviewer's accent and willingness to be surveyed? Explain.

66. Statin drugs are used to decrease cholesterol levels and therefore hopefully to decrease the chances of a heart attack. In a British study ("MRC/BHF Heart Protection Study of Cholesterol Lowering with Simvastin in 20,536 High-Risk Individuals: A Randomized Placebo-Controlled Trial," *Lancet* 2002: 7–22) 20,536 at-risk adults were assigned randomly to take either a 40-mg statin pill or placebo. The subjects had coronary disease, artery blockage, or diabetes. After 5 years there were 1328 deaths (587 from heart attack) among the 10,269 in the statin group and 1507 deaths (707 from heart attack) among the 10,267 in the placebo group.

    a. Give a 95% confidence interval for the difference in population death proportions.
    b. Give a 95% confidence interval for the difference in population heart attack death proportions.
    c. Is it reasonable to say that most of the difference in death proportions is due to heart attacks, as would be expected?

67. Using the traditional formula, a 95% CI for $p_1 - p_2$ is to be constructed based on equal sample sizes from the two populations. For what value of $n \ (= m)$ will the resulting interval have width at most .1 irrespective of the results of the sampling?

68. In medical investigations, the ratio $\theta = p_1/p_2$ is often of more interest than the difference $p_1 - p_2$ (e.g., individuals given treatment 1 are how many times as likely to recover as those given treatment 2?). Let $\hat{\theta} = \hat{P}_1/\hat{P}_2$. When m and n are both large, the statistic $\ln(\hat{\theta})$ has approximately a normal distribution with approximate mean value $\ln(\theta)$ and approximate standard deviation $[(m - x)/(mx) + (n - y)/(ny)]^{1/2}$.

    a. Use these facts to obtain a large-sample 95% CI formula for estimating $\ln(\theta)$, and then a CI for θ itself.

b. The article "Low-Dose Aspirin for Preventing Recurrent Venous Thromboembolism (VT)" (*New Engl. J. Med.* 2012: 1979–1987) reports a study in which VT recurred in 73 of 411 patients randomly assigned a placebo and in 57 of the 411 assigned to an aspirin regimen. Calculate an interval of plausible values for θ at the 95% confidence level. What does this interval suggest about the efficacy of the aspirin treatment?

69. All the examples of this section featured success/failure data from two independent samples. **McNemar's Test** handles *paired* binary responses. For example, suppose that before a major policy speech by a political candidate, n individuals are selected and asked whether (S) or not (F) they favor the candidate. Then after the speech the same n people are asked the same question. The responses can be entered in a table as follows:

		After	
		S	F
Before	S	X_1	X_2
	F	X_3	X_4

where $X_1 + X_2 + X_3 + X_4 = n$. Let p_1, p_2, p_3, and p_4 denote the four cell probabilities, so that $p_1 = P(S$ before and S after), and so on. We wish to test the hypothesis that the true proportion of supporters (S) after the speech has not increased against the alternative that it has increased.

a. State the two hypotheses of interest in terms of p_1, p_2, p_3, and p_4.
b. Construct an estimator for the after/before difference in success probabilities.
c. When n is large, it can be shown that the random variable $(X_i - X_j)/n$ has approximately a normal distribution with variance $[p_i + p_j - (p_i - p_j)^2]/n$. Construct a test statistic with approximately a standard normal distribution when H_0 is true.
d. If $x_1 = 350$, $x_2 = 150$, $x_3 = 200$, and $x_4 = 300$, what do you conclude?

70. McNemar's test, developed in the previous exercise, can also be used when individuals are "matched" to yield n pairs and then one member of each pair is given treatment 1 and the other is given treatment 2. Then X_1 is the number of pairs in which both treatments were successful, and similarly for X_2, X_3, and X_4. Suppose the following data is obtained from such a matched pairs design to assess the effectiveness of a certain migraine headache medicine. Use McNemar's test to determine whether the medicine is effective in the treatment of migraines.

		Medicine	
		S	F
Placebo	S	44	34
	F	46	30

10.5   Inferences About Two Population Variances

Methods for comparing two population variances (or standard deviations) are occasionally needed, though such problems arise much less frequently than those involving means or proportions. For the case in which the populations under investigation are normal, the procedures are based on the F distribution from Sections 6.3 and 6.4.

Testing Hypotheses

A test procedure for hypotheses concerning the ratio σ_1/σ_2, as well as a CI for this ratio, is based on the following result from Section 6.4.

THEOREM   Let $X_1, \ldots, X_m$ be a random sample from a normal distribution with standard deviation σ_1, let $Y_1, \ldots, Y_n$ be another random sample (independent of the X_i's) from a normal distribution with standard deviation σ_2, and let S_1 and S_2 denote the two sample standard deviations. Then the rv

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \qquad (10.8)$$

has an F distribution with $v_1 = m - 1$ and $v_2 = n - 1$.

Under the null hypothesis of equal population standard deviations, (10.8) reduces to the ratio of sample variances. For a test statistic we use this ratio of sample variances, and the claim that $\sigma_1 = \sigma_2$ is rejected if the ratio differs by too much from 1.

THE _F_ TEST FOR EQUALITY OF VARIANCES

Null hypothesis: H_0: $\sigma_1 = \sigma_2$ (equivalently, $\sigma_1^2 = \sigma_2^2$)

Test statistic value: $f = s_1^2/s_2^2$

Alternative Hypothesis	Rejection Region for a Level α Test
H_a: $\sigma_1 > \sigma_2$	$f \geq F_{\alpha, m-1, n-1}$
H_a: $\sigma_1 < \sigma_2$	$f \leq F_{1-\alpha, m-1, n-1}$
H_a: $\sigma_1 \neq \sigma_2$	either $f \geq F_{\alpha/2, m-1, n-1}$ or $f \leq F_{1-\alpha/2, m-1, n-1}$

Since critical values are tabled only for $\alpha = .10, .05, .01,$ and $.001$ in Appendix Table A.8, the two-tailed test can be performed only at levels .20, .10, .02, and .002 without statistical software.

Example 10.15 Is there less variation in weights of some baked goods than others? Here are the weights (in grams) for a sample of Bruegger's bagels and another sample of Wolferman's English muffins:

B:	99.8	105.4	94.7	107.8	114.3	106.3		
W:	99.0	98.2	98.1	102.1	102.9	104.1	98.8	99.5

The normality assumption is very important for the use of the F test. Normal probability plots from Minitab are shown in Figure 10.9. There is no apparent reason to doubt normality of either population distribution here.
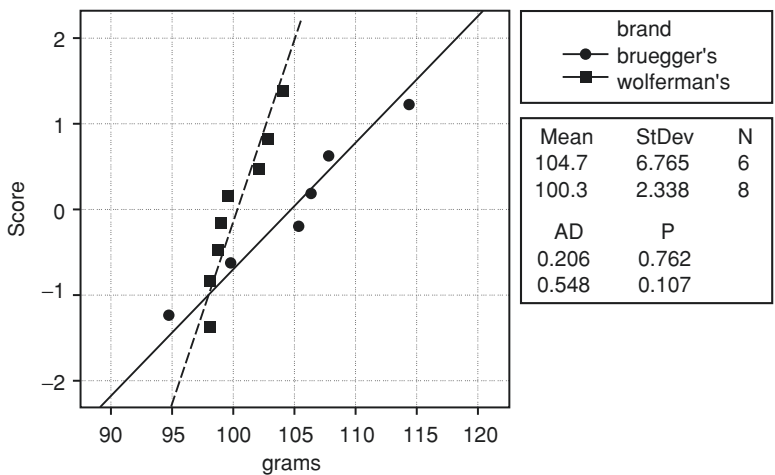
Figure 10.9   Normal plot for baked goods

Notice the difference in slopes for the two sources. This suggests different variabilities because the z-score vertical axis is related to the horizontal axis (grams) by z = (grams − mean)/(std dev). Thus, when score is plotted against grams the slope is the reciprocal of the standard deviation. Now let's test H_0: $\sigma_1 = \sigma_2$ against a two-sided alternative with α = .02. We need the critical values $F_{.01,5,7}$ = 7.46 and $F_{.99,5,7}$ = $1/F_{.01,7,5}$ = 1/10.46 = .0956; here we have used the reciprocal property

$$F_{p,v_1 v_2} = 1/F_{1-p,v_2,v_1} \tag{10.9}$$

from Section 6.3. From the sample data

$$f = \frac{s_1^2}{s_2^2} = \frac{6.765^2}{2.338^2} = 8.37$$

which exceeds 7.46, so the hypothesis of equal standard deviations is rejected. We conclude that there is a difference in weight variation, and the English muffins are less variable.

Notice that it is not really necessary to use the lower-tailed critical value here if the groups are chosen so the first group has the larger variance, and therefore the value of $f = s_1^2/s_2^2$ exceeds 1. Because $f > 1$, the only comparison is between the computed f and the upper critical value 7.46. It does not change the result of the test to fix things so $f > 1$, so it is not cheating to simplify the test in this way. ∎

P-Values for F Tests

Recall that the P-value for an upper-tailed t test is the area under the relevant t curve (the one with appropriate df) to the right of the calculated t. In the same way, the P-value for an upper-tailed F test is the area under the F curve with appropriate numerator and denominator df to the right of the calculated f. Figure 10.10 illustrates this for a test based on v_1 = 4 and v_2 = 6.
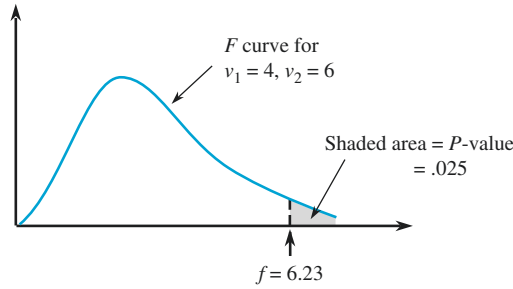
Figure 10.10   A P-value for an upper-tailed F test

Unfortunately, tabulation of F curve upper-tail areas is much more cumbersome than for t curves because two df's are involved. For each combination of v_1 and v_2, our F table gives only the four critical values that capture areas .10, .05, .01, and .001. Figure 10.11 shows what can be said about the P-value depending on where f falls relative to the four critical values.
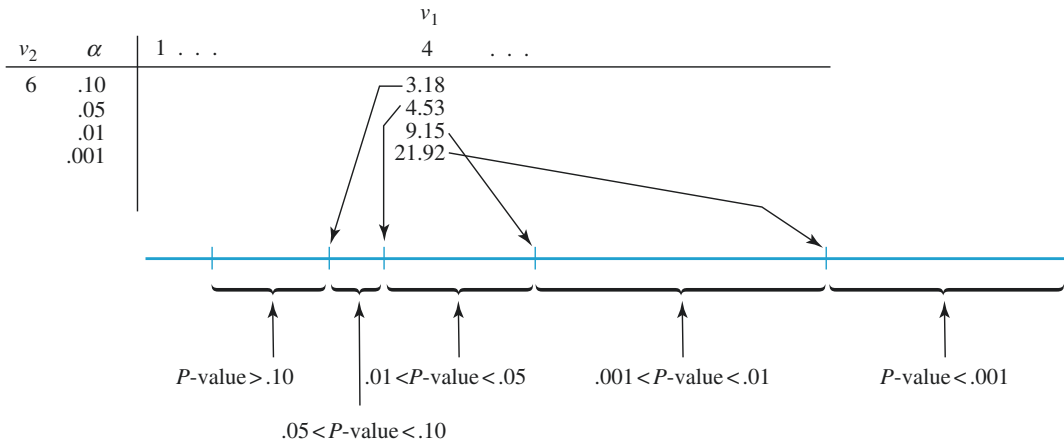
Figure 10.11   Obtaining P-value information from the F table for an upper-tailed F test

For example, for a test with $v_1 = 4$ and $v_2 = 6$,

$$f = 2.16 \qquad \Rightarrow P\text{-value} > .10$$
$$f = 5.70 \qquad \Rightarrow .01 < P\text{-value} < .05$$
$$f = 25.03 \qquad \Rightarrow P\text{-value} < .001$$

Once we know that $.01 < P$-value $< .05$, H_0 would be rejected at a significance level of .05 but not at a level of .01. When P-value $< .001$, H_0 should be rejected at any reasonable significance level.

The F tests discussed in succeeding chapters will all be upper-tailed. If, however, a lower-tailed F test is appropriate, then (10.9) should be used to obtain lower-tailed critical values so that bounds on the P-value can be established. In the case of a two-tailed test, the bounds from a one-tailed test should be multiplied by 2. For example, if $f = 5.82$ when $v_1 = 4$ and $v_2 = 6$, then since 5.82 falls between the .05 and .01 critical values, $2(.01) < P$-value $< 2(.05)$, giving $.02 < P$-value $< .10$. H_0 would then be

rejected if $\alpha = .10$ but not if $\alpha = .01$. In this case, we cannot say from our table what conclusion is appropriate when $\alpha = .05$ (since we don't know whether the P-value is smaller or larger than this). However, statistical software shows that the area to the right of 5.82 under the $F_{4,6}$ curve is .029, so the P-value is $2(.029) = .058$ and the null hypothesis should therefore not be rejected at level .05. More generally, good statistical software will provide an exact P-value for any test based on an F distribution.

A Confidence Interval for σ_1/σ_2

The CI for σ_1/σ_2 is based on the probability statement implied by (10.8):

$$P\left(F_{1-\alpha/2,v_1,v_2} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{\alpha/2,v_1,v_2}\right) = 1 - \alpha$$

Manipulating the inequalities to isolate σ_1^2/σ_2^2 yields

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2,v_1,v_2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{1-\alpha/2,v_1,v_2}} = \frac{s_1^2}{s_2^2} \cdot F_{\alpha/2,v_2,v_1}$$

Equation (10.9) has been used here to simplify the upper bound and enable use of Table A.8. Thus the confidence interval for σ_1^2/σ_2^2 is

$$\left(\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2,m-1,n-1}}, \frac{s_1^2}{s_2^2} \cdot F_{\alpha/2,n-1,m-1}\right)$$

An interval for σ_1/σ_2 results from taking the square root of each limit:

$$\left(\frac{s_1}{s_2} \cdot \frac{1}{\sqrt{F_{\alpha/2,m-1,n-1}}}, \frac{s_1}{s_2} \cdot \sqrt{F_{\alpha/2,n-1,m-1}}\right)$$

In the interval for the ratio of population standard deviations, notice that the limits of the interval are proportional to the ratio of sample standard deviations. Of course, the lower limit is less than the ratio of sample standard deviations, and the upper limit exceeds it.

Example 10.16 Let's calculate a confidence interval using the data of Example 10.15. The sample standard deviations are $s_1 = 6.765$ for 6 Bruegger's bagels and $s_2 = 2.338$ for 8 Wolferman English muffins. Then a 98% confidence interval for the ratio σ_1/σ_2 is

$$\left(\frac{6.765}{2.338} \cdot \frac{1}{\sqrt{F_{.01,5,7}}}, \frac{6.765}{2.338} \cdot \sqrt{F_{.01,7,5}}\right) = \left(2.89 \cdot \frac{1}{\sqrt{7.46}}, 2.89 \cdot \sqrt{10.46}\right) = (1.06, \ 9.35)$$

Because 1 is not included in the interval, it suggests that the two standard deviations differ. By comparing the CI calculation with the hypothesis test calculation, it should be clear that a two-tailed test would reject equality at the 2% level, and this is consistent with the results of Example 10.15. ∎

It is important to emphasize that the methods of this section are strongly dependent on the normality assumption. Expression (10.8) is valid only in the case of normal data or nearly normal data. Otherwise, the F distribution in (10.8) does not apply. The t procedures of this chapter are robust to the normality assumption, meaning that the procedures still work in the case of moderate departures from normality, but this is not true for comparison of standard deviations based on (10.8).

For nonnormal data, alternative tests for equal variance are available, including Levene's test and an extension of the method by Bonett mentioned in Section 8.4. Consult your local statistician for more information.

Exercises: Section 10.5 (71–78)

71. Obtain or compute the following quantities:

    a. $F_{.05,5,8}$

    b. $F_{.05,8,5}$

    c. $F_{.95,5,8}$

    d. $F_{.95,8,5}$

    e. The 99th percentile of the F distribution with $v_1 = 10$, $v_2 = 12$

    f. The 1st percentile of the F distribution with $v_1 = 10$, $v_2 = 12$

    g. $P(F \leq 6.16)$ for $v_1 = 6$, $v_2 = 4$

    h. $P(.177 \leq F \leq 4.74)$ for $v_1 = 10$, $v_2 = 5$

72. Give as much information as you can about the P-value of the F test in each of the following situations:

    a. $v_1 = 5$, $v_2 = 10$, upper-tailed test, $f = 4.75$

    b. $v_1 = 5$, $v_2 = 10$, upper-tailed test, $f = 2.00$

    c. $v_1 = 5$, $v_2 = 10$, two-tailed test, $f = 5.64$

    d. $v_1 = 5$, $v_2 = 10$, lower-tailed test, $f = .200$

    e. $v_1 = 35$, $v_2 = 20$, upper-tailed test, $f = 3.24$

73. Refer to Exercise 41. Does the data suggest that the standard deviation of the strength distribution for fused specimens is smaller than that for not-fused specimens? Carry out a test at significance level .01 by obtaining as much information as you can about the P-value.

74. Return to the data on maximum lean angle given in Exercise 29. Carry out a test at significance level .10 to see whether the population standard deviations for the two age groups are different (normal probability plots support the necessary normality assumption).

75. Refer to the railway repair time data in Exercise 16. Carry out a test at significance level .01 to see whether the population standard deviation for time-to-repair is larger for high rail breaks than for low rail breaks.

76. Exercise 35 presented data on the pour size of two groups of experienced bartenders, one group pouring into tumblers and the other into highball glasses. Test the hypothesis that the variability in the two (conceptual) populations of pour sizes is different, at the $\alpha = .02$ level.

77. Return to the fat loss experiment described in Exercise 24. Calculate a 95% CI for the ratio of the population standard deviations for the experimental and control groups.

78. For the data of Exercise 29 find a 90% confidence interval for the ratio of population standard deviations, and relate your CI to the test of Exercise 74.

10.6 Inferences Using the Bootstrap and Permutation Methods

In this chapter we have discussed how to make comparisons based on normal data. We have also considered comparisons of means when the sample sizes are large enough for t procedures to apply even in the absence of normality. What about other cases (e.g., smaller, skewed data sets), for which such methods are inappropriate? We now consider computer-intensive "resampling" techniques for confidence intervals and hypothesis tests that can be applied to many comparison situations.

The Two-Sample Bootstrap CI
The bootstrap for two samples is similar to the one-sample bootstrap of Section 8.5, except that samples with replacement are taken from the two original samples separately. That is, a resample is taken from the first group, a separate resample is taken from the second group, and then the difference of means (or some other comparison statistic) is computed. This process is repeated a large number of times, resulting in the *bootstrap distribution* of the comparison statistic.

   If the bootstrap distribution appears normal, then a *bootstrap t confidence interval* can be computed in a similar manner to the one presented in Section 8.5, with s_{boot} replacing the standard error expression from Section 10.2. If the statistic being bootstrapped is $\overline{X} - \overline{Y}$, it is common to use a conservative t critical value with df = min$(m - 1, n - 1)$; Welch's df formula (10.3) is also sometimes used in practice, partly to agree with the classic two-sample t interval for $\mu_1 - \mu_2$. (On the other hand, if we are bootstrapping a difference of medians or trimmed means there is no concern about agreement with a t interval.) Another reasonable alternative is to use a z critical value (df = ∞; the software package Stata does this).

   If the bootstrap distribution does not look normal, then the percentile interval should be calculated, just as was done in Section 8.5. A CI with confidence level approximately 95% requires determining the 2.5 and 97.5 percentiles of the bootstrap distribution. The bias corrected and adjusted (BCa) interval is a further refinement available in some software packages, including R and Stata. Once a valid $100(1 - \alpha)\%$ CI has been calculated, the hypothesis $\mu_1 - \mu_2 = \Delta_0$ is rejected at significance level α in favor of the two-sided (i.e., $\neq$) alternative if and only if the CI does not include Δ_0.

Example 10.17 As an example of the bootstrap for two samples, consider data from a study of children talking to themselves (private speech), introduced in Exercise 81 of Chapter 1. The children were each observed in many 10-second intervals (about 100) and the researchers computed the percentage of intervals in which private speech occurred. Because private speech tends to occur when there is a challenging task, the students were observed when they were doing arithmetic. The private speech is classified as on task if it is about arithmetic, off task if it is about something else, and mumbling if the subject is not clear.

   Here we consider just the off-task percentages for the 18 male and 15 female first graders:

B: 4.9, 5.5, 6.5, 0.0, 0.0, 3.0, 2.8, 6.4, 1.0, 0.9, 0.0, 28.1, 8.7, 1.6, 5.1, 17.0, 4.7, 28.1
G: 0.0, 1.3, 2.2, 0.0, 1.3, 0.0, 0.0, 0.0, 0.0, 3.9, 0.0, 10.1, 5.2, 3.2, 0.0

   The two-sample t interval of Section 10.2 should not be applied: the sample sizes are rather small, and with the large number of zeroes (a majority for the girls), the population normality assumption is clearly violated. Nevertheless, it is useful to give that CI purely for comparison purposes. The 95% interval is

$$\bar{x} - \bar{y} \pm t_{.025,v}\sqrt{\frac{s_1^2}{18} + \frac{s_2^2}{15}} = 6.906 - 1.813 \pm 2.080\sqrt{\frac{8.719^2}{18} + \frac{2.846^2}{15}}$$
$$= 5.093 \pm 2.080(2.1825) = 5.093 \pm 4.540 = (.55, 9.63)$$

Welch's formula was used to obtain $v = 21$.

Again, the *t* method is of questionable validity, because the sample sizes might not be large enough to compensate for the nonnormality. The bootstrap method involves drawing a random resample of size 18 with replacement from the 18 boys, drawing a random resample of size 15 with replacement from the 15 girls, and calculating the difference of resampled means $\bar{x}^* - \bar{y}^*$. This process is repeated a large number of times, creating a bootstrap distribution for the statistic $\overline{X} - \overline{Y}$. Here are random resamples from the boys and girls:

B: 0.0, 3.0, 2.8, 0.9, 3.0, 0.0, 0.0, 6.5, 6.4, 8.7, 6.4, 1.0, 0.9, 5.5, 17.0, 17.0, 0.0, 3.0
G: 1.3, 0.0, 0.0, 0.0, 0.0, 1.3, 1.3, 0.0, 3.2, 0.0, 1.3, 5.2, 0.0, 0.0, 0.0

For these two resamples, the difference of means is $\bar{x}^* - \bar{y}^* = 4.56 - .91 = 3.65$. Doing this 1000 times (using the R package `boot`) gives the bootstrap distribution displayed in Figure 10.12.

The bootstrap distribution looks *almost* normal, but with some positive skewness. If the original sample of boys and girls is representative of their populations, then the histogram in Figure 10.12 should resemble the true sampling distribution of $\overline{X} - \overline{Y}$ in this scenario. For example, the standard deviation of the bootstrap distribution (i.e., of the 1000 $\bar{x}^* - \bar{y}^*$ values) is $s_{\text{boot}} = 2.1874$, very close to the 2.1825 that was computed for the estimated standard error in the two-sample *t* interval above.
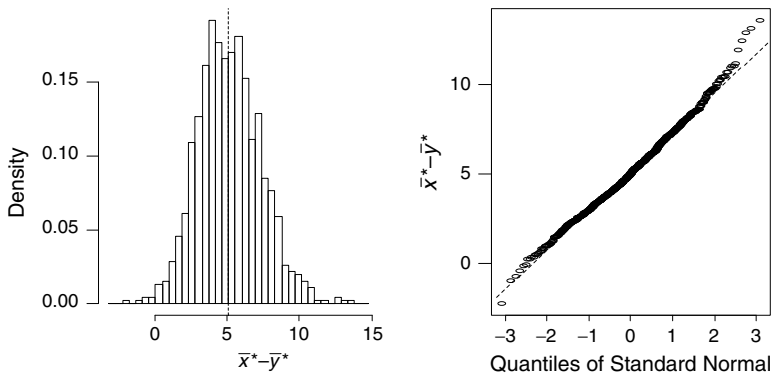
Figure 10.12 Histogram and normal plot of the bootstrapped difference in means from R

In the presence of a not-quite-normal bootstrap distribution, we use the percentile interval. The confidence limits for a 95% confidence interval are the 2.5 percentile and the 97.5 percentile of the $\bar{x}^* - \bar{y}^*$ distribution. When the 1000 bootstrap differences of means were sorted, the 25th value from the bottom was 1.029 and the 25th value from the top was 9.760. This gives a 95% CI of (1.029, 9.760). The skewness of the bootstrap distribution pushes the endpoints a little to the right of the endpoints of the two-sample *t* interval. In addition, one can use software to compute the BCa refinement, as discussed in Section 8.5. The improved interval (1.625, 10.446), obtained from R, is

moved even farther to the right compared to the previous intervals. This last interval is the most trustworthy.

Neither of these bootstrap intervals includes 0, implying that the hypothesis $\mu_1 - \mu_2 = 0$ should be rejected at the .05 significance level in favor of the conclusion that the two population means are different.   ∎

Permutation Tests

Permutation tests provide a template for comparing two (or more) populations without requiring large samples or assuming any specific distribution for the data. The relevant null hypothesis is that the two population distributions are identical (which implies both equal means and equal standard deviations). The idea behind such tests is that under the null hypothesis, every observation comes from the same distribution, and so the group labels (e.g., population 1 vs population 2 or treatment vs control) are meaningless. If that's true, then we can permute—that is, scramble or rearrange—the group labels without changing the group population means. We look at all possible label arrangements (or at least a large number), compute the difference of means for each of these, and compute a *P*-value by seeing how extreme is our original difference of means.

Example 10.18 The article "Comparison of Platelet Function and Viscoelastic Test Results between Healthy Dogs and Dogs with Naturally Occurring Chronic Kidney Disease" (*Amer. J. Veterinary Res.* 2017: 589–600), first presented in Example 1.18, provides data on the fibrinogen levels (mg/dl of blood) for two samples of dogs: 11 with chronic kidney disease (CKD) and 10 with normal kidney function. Researchers were concerned that CKD increases the production of fibrinogen, which can lead to excessive blood clotting. Boxplots of both samples show considerable skewness and the sample sizes are small, so a two-sample *t* test would be of questionable validity.

In order to demonstrate the permutation test method, consider an even smaller-scale version of this data: measurements 315, 290, 275 for the CKD dogs ($m = 3$) and 313, 250 for the healthy dogs ($n = 2$). Under the null hypothesis of equal population distributions, it should not matter if we reassign the labels "CKD" and "healthy." Therefore, we consider all ways of selecting three from among the five observations to be the CKD sample, leaving the other two for the healthy sample. Under H_0, the ten choices listed in Table 10.4 are equally likely.

Table 10.4 All possible rearrangements of $m = 3$ and $n = 2$ observations

CKD dogs			$\bar{x}$	Healthy dogs		$\bar{y}$	$\bar{x} - \bar{y}$
315	290	275	293.3	313	250	281.5	11.8
315	290	313	306.0	275	250	262.5	43.5
315	290	250	285.0	313	275	294.0	–9.0
315	275	313	301.0	290	250	270.0	31.0
315	275	250	280.0	290	313	301.5	–21.5
315	313	250	292.7	290	275	282.5	10.2
290	275	313	292.7	315	250	282.5	10.2
290	275	250	271.7	315	313	314.0	–42.3
290	313	250	284.3	275	315	295.0	–10.7
275	313	250	279.3	290	315	302.5	–23.2

How extreme is our original difference of means, 293.3 − 281.5 = 11.8 (the top row of Table 10.4), in this set of ten differences? Because it is the third-largest of the ten $\bar{x} - \bar{y}$ values, our P-value for one-sided test is 3/10 = .3, the fraction of arrangements that give a difference at least as large as our original difference.   ∎

When $m = 3$ and $n = 2$, it is simple enough to deal with all $\binom{5}{3} = 10$ arrangements. What happens when we try to use the whole set of 21 dogs?

Example 10.19 (Example 10.18 continued) Now consider a permutation test for the full dog health data:

CKD dogs: 183, 190, 250, 275, 290, 315, 320, 330, 410, 500, 821 ($m = 11$, $\bar{x} = 353.1$)
Healthy dogs: 99, 160, 165, 170, 178, 181, 190, 201, 250, 313 ($n = 10$, $\bar{y} = 190.7$)

Here we are dealing with $\binom{21}{11} = 352{,}715$ possible permutations of the 11 CKD dogs and 10 healthy dogs. Even on a reasonably fast computer it might take a while to generate this many differences and see how many are at least as large as the value $\bar{x} - \bar{y} = 353.1 - 190.7 = 162.4$ from the original data. Instead, we can take a random sample of all possible arrangements and get quite close to the exact answer. Figure 10.13 shows a histogram of 2000 values of $\bar{x} - \bar{y}$ created by randomly permuting the group labels; though this is short of all possible arrangements, this should give us a reasonable estimate of the P-value. Of the 2000 label permutations, only two resulted in an $\bar{x} - \bar{y}$ value of 162.4 or higher, for an estimated P-value of 2/2000 = .001. Thus, we have convincing statistical evidence to conclude that dogs with chronic kidney disease do have higher blood fibrinogen levels, on average, than healthy dogs.
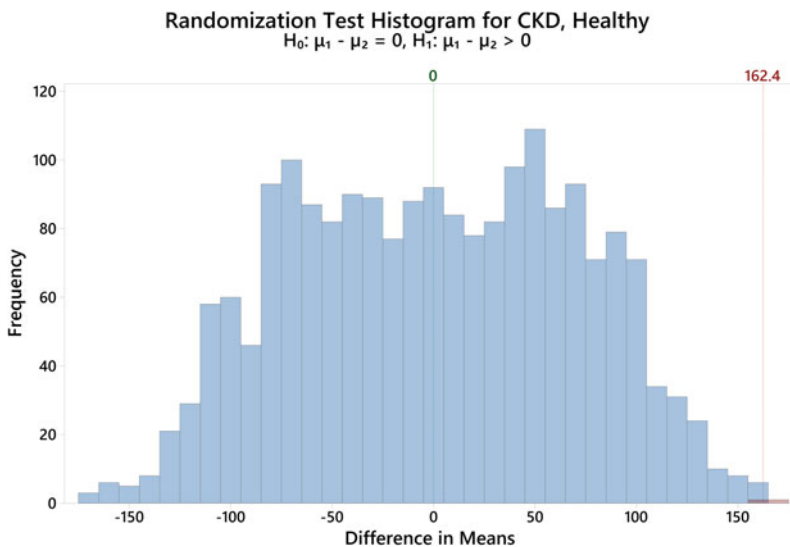
Figure 10.13 Permutation distribution for Example 10.19   ∎

The method shown in Example 10.19 could equally be applied to the difference of two sample medians or any other comparison statistic. The general permutation test method is summarized in the accompanying box.

Permutation Tests	Let θ_1 and θ_2 be the same parameters (means, medians, standard deviations, etc.) for two different populations, and consider testing H_0: $\theta_1 = \theta_2$ based on independent samples of sizes m and n, respectively. Suppose that when H_0 is true, the two population distributions are identical in all respects, so all $m + n$ observations have actually been selected from the same population distribution. In this case, the labels 1 and 2 are arbitrary, as any m of the $m + n$ observations have the same chance of ending up in the first sample (leaving the remaining n for the second sample).
	An exact permutation test computes a suitable comparison statistic for all possible rearrangements and sets the P-value equal to the fraction of these that are at least as extreme as the statistic computed on the original samples. This is the P-value for a one-tailed test, and it needs to be doubled for a two-tailed test.
	For an approximate permutation test, instead of all possible arrangements, we take a random sample with replacement from the set of all possible arrangements.

Permutation tests do not assume a specific underlying distribution, such as the normal distribution. However, this does not mean that there are no assumptions whatsoever. The null hypothesis in a permutation test is that the two *distributions* are the same, and any deviation can increase the probability of rejecting the null hypothesis. Thus, strictly speaking, we are doing a test for equal means only if the distributions are alike in all other respects, including shape and variability. See Exercise 94 for a (pathological) example in which the permutation test underestimates the true P-value.

Inferences Based on Other Statistics

The bootstrap and permutation methods are not limited to comparing means. In any of the previous examples, we could have considered the difference of two medians or two trimmed means instead. Likewise, these methods can be employed for inferences concerning the variability of two populations. Section 10.5 discussed the use of the F distribution for comparing two variances, but this inferential method is strongly dependent on normality. Bootstrapping does not require this assumption.

Example 10.20 Consider the off-task private speech data from Example 10.17. The sample standard deviations for boys and girls are 8.72 and 2.85, respectively. The method of Section 10.5 gives for the ratio of male to female variances the 95% confidence interval

$$\left(\frac{s_1^2}{s_2^2}\frac{1}{F_{.025,17,14}}, \frac{s_1^2}{s_2^2}\frac{1}{F_{.975,17,14}}\right) = \left(\frac{8.72^2}{2.85^2}\frac{1}{2.900}, \frac{8.72^2}{2.85^2}\frac{1}{.3633}\right) = (3.23, 25.77)$$

Taking the square root gives (1.80, 5.08) as the 95% confidence interval for the ratio of standard deviations. However, the legitimacy of this interval is seriously in question because of the skewed distributions.

Let's apply the bootstrap method to this problem. Take random resamples of 18 boys and 15 girls, calculate the standard deviations s_1^* and s_2^* of the two resamples, and then compute their ratio s_1^*/s_2^*. One such pair of resamples returned $s_1^* = 5.264$ for the boys and $s_2^* = 1.505$ for the girls, for a ratio of 5.264/1.505 = 3.498. This process was repeated 1000 times using the `boot` package in R.

The resulting bootstrap distribution (not shown) is strongly skewed to the right, so a percentile interval is required. The 2.5 percentile is 1.013 and the 97.5 percentile is 7.888, so the 95% confidence interval for the population ratio of standard deviations is (1.013, 7.888). The BCa refinement gives the interval (0.885, 7.382).

These two intervals differ in an important respect: the percentile interval excludes 1 but the BCa refinement includes 1. In other words, the BCa interval allows the possibility that the two population standard deviations are the same, but the percentile interval does not. We expect the BCa method to be an improvement, and this is verified in the next example, where we see that the BCa result is consistent with the results of a permutation test. ∎

Next, consider testing H_0: $\sigma_1 = \sigma_2$. Again, the traditional F test of Section 10.5 requires data from normal populations and is not robust to violations of that assumption, so its validity is questionable for nonnormal data (even when the sample sizes are large). Instead, we might use a permutation test.

It must be re-emphasized that the permutation assumes identical *distributions* under H_0, not just identical sd's. So, for instance, data from two populations with very different shapes or means but the same variability would likely result in a low P-value from the permutation test, even if the statement $\sigma_1 = \sigma_2$ is true. In fairness, the F test also assumes identical shapes (the normal curve), though not necessarily the same means. A graphical exploration of the data may illuminate the nature of the differences between two groups if a permutation test rejects its null hypothesis.

Example 10.21 (Example 10.20 continued) We know that the ratio of sample standard deviations for off-task private speech, males versus females, is 8.72/2.85 = 3.064. The idea of the permutation test is to find out how unusual this value is if we blur the distinction between males and females. That is, we remove the labels from the 18 males and 15 females and then consider all possible choices of 18 from the 33 children. For each of these possible choices we find the ratio of the standard deviation of the first 18 to the standard deviation of the last 15. The one-tailed P-value is the fraction that is at least as big as the original ratio value of 3.064.

Because there are more than a billion possible choices of 18 from 33, we instead selected 5000 random choices. Of these, 432 were at least as large as 3.064, so the one-tailed P-value is 432/5000 = .0864. For a two-tailed P-value we double this and get .1728. The permutation test does not reject H_0: $\sigma_1 = \sigma_2$ in favor of H_a: $\sigma_1 \neq \sigma_2$ at the 5% level (or even the 10% level).

How does the permutation test result compare with the previous results? Recall that the F interval and the "unadjusted" percentile interval ruled out the possibility that the two standard deviations are the same, but the BCa refinement disagreed, because 1 was included in the BCa interval. Taking it for granted that the permutation test is a valid approach and the permutation test does not reject the equality of standard deviations, the BCa interval is the only one of the three CIs consistent with this result. ∎

The Analysis of Paired Data
The bootstrap can be used for paired data if we work with the paired differences, as in the paired t methods of Section 10.3.

Example 10.22 Consider once again the private speech study from Example 10.17. The study included the percentage of intervals with on-task private speech for 33 children in the first, second, and third grades. Here we will consider just the 15 girls' scores in first and second grade. Is there a change in on-task private speech when the girls go from the first to the second grade? Here are the percentages of intervals in which on-task private speech occurred, and also the differences.

Grade 1	Grade 2	Difference
25.7	18.6	7.1
36.0	17.4	18.6
27.6	2.6	25.0
29.7	0.9	28.8
36.0	1.5	34.5
35.1	14.1	21.0
42.0	3.3	38.7

(continued)

Grade 1	Grade 2	Difference
7.6	1.6	6.0
14.1	0.0	14.1
25.0	1.5	23.5
20.2	0.0	20.2
24.4	2.1	22.3
10.4	18.4	−8.0
21.1	2.6	18.5
5.6	26.0	−20.4

Figure 10.14 shows a histogram for the differences; there is a pronounced negative skew.
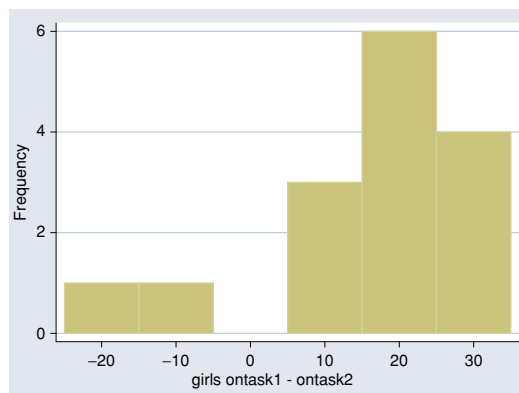
Figure 10.14   Histogram of differences for girls from Stata

The paired t method of Section 10.3 requires normality, so the skewness might invalidate this, but we present results here anyway for comparison purposes. The 95% confidence interval for the population mean difference is

$$\bar{d} \pm t_{.025,15-1}\frac{s_D}{\sqrt{15}} = 16.66 \pm 2.145\frac{15.43}{\sqrt{15}} = 16.66 \pm 8.54 = (8.12, 25.20)$$

The bootstrap focuses on the 15 differences and uses the method of Section 8.5. Using Stata, we drew 1000 resamples of size 15 with replacement from the 15 differences; the 1000 resample means $\bar{d}_1^*, \ldots, \bar{d}_{1000}^*$ constitute the bootstrap distribution. Figure 10.15 shows a histogram of these mean differences.

The histogram of $\bar{d}^*$ values is negatively skewed, which is expected because of the negative skewness shown in Figure 10.14 for the original sample. The 95% percentile interval has the 2.5th percentile of the bootstrap distribution as its lower limit and the 97.5th percentile as its upper limit: (7.91, 23.97). This interval is to the left of the paired t interval because of the negative skewness of the bootstrap distribution. The BCa refinement from Stata yields the interval (6.43, 23.12), which is even farther to the left.
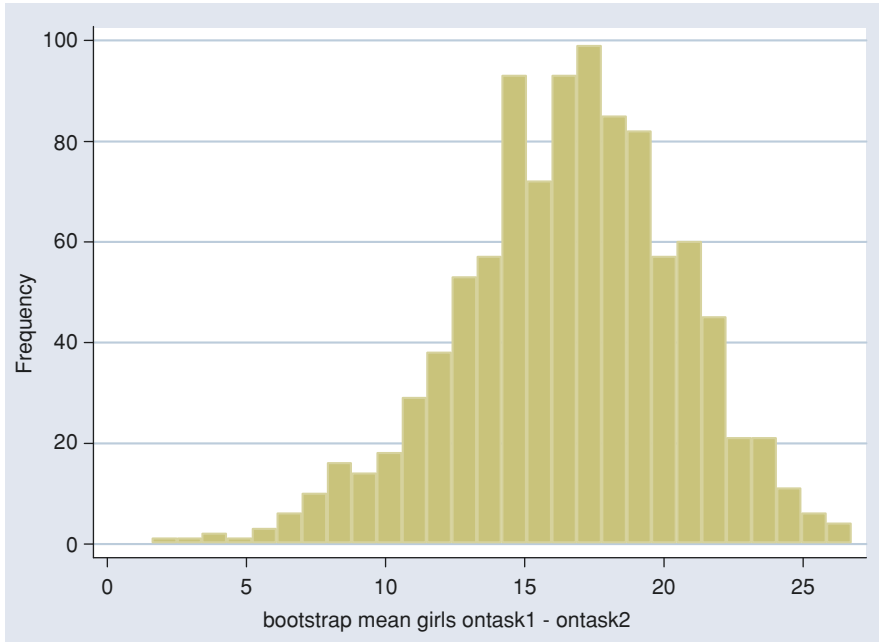
Figure 10.15 Histogram of bootstrap differences for girls from Stata

All of the intervals agree that there is a substantial population difference between first grade and second grade. There is a strong reduction in the on-task private speech of girls between first and second grades. ∎

A permutation test for paired data involves permutations *within the pairs*. Under the null hypothesis of identical population distributions, the two observations in a pair have the same population mean, so the population mean difference is zero even if the order is reversed. Therefore, we consider all possible orderings of the n pairs. Because there are two possible orderings within each pair, there are 2^n arrangements of n pairs. The one-tailed P-value is the fraction of the 2^n differences that are at least as extreme as the observed value, and the two-tailed P-value is double this.

Example 10.23 To see how the permutation test works for paired data, first consider a scaled-down version of the data from Example 10.22 with only the first three pairs: (25.7, 18.6), (36.0, 17.4), and (27.6, 2.6). They give a mean difference of $(7.1 + 18.6 + 25.0)/3 = 16.9$. Here are all $8 = 2^3$ permutations with the corresponding mean differences.

	Arrangements		Mean difference
(25.7, 18.6)	(36.0, 17.4)	(27.6, 2.6)	16.90
(25.7, 18.6)	(36.0, 17.4)	(2.6, 27.6)	.23
(25.7, 18.6)	(17.4, 36.0)	(27.6, 2.6)	4.50
(25.7, 18.6)	(17.4, 36.0)	(2.6, 27.6)	−12.17
(18.6, 25.7)	(36.0, 17.4)	(27.6, 2.6)	12.17
(18.6, 25.7)	(36.0, 17.4)	(2.6, 27.6)	−4.50
(18.6, 25.7)	(17.4, 36.0)	(27.6, 2.6)	−.23
(18.6, 25.7)	(17.4, 36.0)	(2.6, 27.6)	−16.90

Because the mean difference for the original sample is the highest value of eight, the one-tailed P-value is $1/8 = .125$, and the two-tailed P-value is $2(1/8) = .25$.

Next, let's apply the permutation test to the paired data for all 15 girls of Example 10.22. In principle it is no harder to deal with the $2^n = 2^{15} = 32,768$ arrangements when all 15 pairs are included, but this exact approach is generally approximated using a random sample. We used Stata to draw an additional 4999 samples. Of the 4999, none yielded a mean difference as large as the value $\bar{d} = 16.66$ obtained for the original sample of 15 differences. Therefore, the one-tailed P-value is $1/5000 = .0002$, and the two-tailed P-value is $2(.0002) = .0004$. Rejection of the null hypothesis at the 5% level was to be expected, given that none of the confidence intervals in Example 10.22 included 0.

It is interesting to compare the permutation test result with the paired t test of Section 10.3. For testing the null hypothesis of 0 population mean difference, the value of t is

$$\frac{\bar{d} - 0}{s_D/\sqrt{15}} = \frac{16.66}{15.425/\sqrt{15}} = 4.183$$

The two-tailed P-value for this is .0009, not very different from the result of the permutation test. ∎

Exercises: Section 10.6 (79–94)

79. A student project by Heather Kral studied students on "lifestyle floors" of a dormitory in comparison to students on other floors. On a lifestyle floor the students share a common major, and there are a faculty coordinator and resident assistant from that department. Here are the GPAs of 30 students on lifestyle floors (L) and 30 students on other floors (N):

L:	2.00	2.25	2.60	2.90	3.00	3.00	3.00	3.00
	3.00	3.20	3.20	3.25	3.30	3.30	3.32	3.50
	3.50	3.60	3.60	3.70	3.75	3.75	3.79	3.80
	3.80	3.90	4.00	4.00	4.00	4.00		
N:	1.20	2.00	2.29	2.45	2.50	2.50	2.50	2.50
	2.65	2.70	2.75	2.75	2.79	2.80	2.80	2.80
	2.86	2.90	3.00	3.07	3.10	3.25	3.50	3.54
	3.56	3.60	3.70	3.75	3.80	4.00		

Notice that the lifestyle GPAs have a large number of repeats and the distribution is skewed, so there is some question about normality.

a. Obtain a 95% confidence interval for the difference of population means using the two-sample t interval.

b. Use software to generate a bootstrap sample of differences of means. Check the bootstrap distribution for normality using a normal probability plot.

c. Use the standard deviation of the bootstrap distribution along with the mean and t critical value from (a) to get a 95% confidence interval for the difference of means.

d. Use the bootstrap sample and the percentile method to obtain a 95% confidence interval for the difference of means.

e. Compare your three confidence intervals. If they are very similar, why do you think this is the case?

f. Interpret your results. Is there a substantial difference between lifestyle and other floors? Why do you think the difference is as big as it is?

80. For the data of the previous exercise, now consider testing the hypothesis of equal population variances.

a. Carry out a two-tailed test using the method of Section 10.5. Recall that this method requires the data to be normal, and the method is sensitive to departures

from normality. Check the data for normality to see if the F test is justified.

b. Carry out a two-tailed permutation test for the hypothesis of equal population variances (or standard deviations). Why does it not matter whether you use variances or standard deviations?

c. Compare the two results and summarize your conclusions.

81. For the data of the previous two exercises, we want a 95% confidence interval for the ratio of population standard deviations.

a. Use the method of Section 10.5. Recall that this method requires the data to be normal, and the method is sensitive to departures from normality. Check the data for normality to see if the F distribution can be used for the ratio of sample variances.

b. Use software to generate a bootstrap sample of ratios of standard deviations. Then use the percentile method to obtain a 95% confidence interval for the ratio of population standard deviations.

c. Compare the two results and discuss the relationship of the results to those of the previous exercise.

82. In this application from major league baseball, the populations represent an abstraction of what the players can do, so the populations will vary from year to year. The Colorado Rockies and the Arizona Diamondbacks played nine games in Phoenix and ten games in Denver in 2001. The thinner air in Denver causes curve balls to curve less and it allows fly balls to travel farther. Does this mean that more runs are scored in Denver? The numbers of runs scored by the two teams in the nine Phoenix games (P) and ten Denver games (D) are

P:	5.09	15.88	3	8.47	11.65
	6.48	11.65	7.41	9.53	
D:	10	18	15.56	19	8.1
	14	13.76	10	20.12	10.59

The fractions occur because the numbers have been adjusted for nine innings (54 outs). For example, in the third Denver

game the Rockies won 10 to 7 on a home run with two out in the bottom of the tenth inning, so there were 59 outs instead of 54, and the number of runs is adjusted to $(54/59)(17) = 15.56$. We want to compare the average runs in Denver with the average runs in Phoenix.

a. Find a 95% confidence interval for the difference of population means using the two-sample t interval.

b. Use software to generate a bootstrap sample of differences of means. Check the bootstrap distribution for normality using a normal probability plot.

c. Use the standard deviation of the bootstrap distribution along with the mean and t critical value from (a) to get a 95% confidence interval for the difference of means.

d. Use the bootstrap sample and the percentile method to obtain a 95% confidence interval for the difference of means.

e. Compare your three confidence intervals. If you used a standard normal critical value in place of the t critical value in (c), why would that make this interval more like the one in (d)? Why should the three intervals be fairly similar for this data set?

f. Interpret your results. Is there a substantial difference between the two locations? Compare the difference with what you thought it would be. If you were a major league pitcher, would you want to be traded to the Rockies?

83. For the data of the previous exercise we want to compare population medians for the runs in Denver versus the runs in Phoenix.

a. Use software to generate a bootstrap sample of differences of medians. Check the bootstrap distribution for normality using a normal probability plot.

b. Use the standard deviation of the bootstrap distribution along with the difference of the medians in the original

sample and the *t* critical value from the previous exercise to get a 95% confidence interval for the difference of population medians.

c. Use the bootstrap sample and the percentile method to obtain a 95% confidence interval for the difference of population medians.

d. Compare the two confidence intervals.

e. How do the results for the median compare with the results for the mean? In terms of precision (measured by the width of the confidence interval) which gives the best results?

84. Can the right diet help us cope with diseases associated with aging such as Alzheimer's disease? A study ("Reversals of Age-Related Declines in Neuronal Signal Transduction, Cognitive, and Motor Behavioral Deficits with Blueberry, Spinach, or Strawberry Dietary Supplement," *J. Neurosci.* 1999: 8114–8121) investigated the effects of fruit and vegetable supplements in the diet of rats. The rats were 19 months old, which is aged by rat standards. The 40 rats were randomly assigned to four diets, of which we will consider just the blueberry diet and the control diet here. After 8 weeks on their diets, the rats were given a number of tests. We give the data for just one of the tests, which measured how many seconds they could walk on a rod. Here are the times for the ten control rats (C) and ten blueberry rats (B):

C:	15.00	7.00	2.44	5.60	3.63
	6.24	4.12	8.21	3.90	0.95
B:	5.12	9.38	18.77	15.03	6.67
	7.91	7.38	15.09	11.57	8.98

The objective is to obtain a 95% confidence interval for the difference of population means.

a. Determine a 95% confidence interval for the difference of population means using the two-sample *t* interval.

b. Use software to generate a bootstrap sample of differences of means. Check

the bootstrap distribution for normality using a normal probability plot.

c. Use the standard deviation of the bootstrap distribution along with the mean and *t* critical value from (a) to get a 95% confidence interval for the difference of means.

d. Use the bootstrap sample and the percentile method to obtain a 95% confidence interval for the difference of means.

e. Compare your three confidence intervals. If they are very similar, why do you think this is the case? If you had used a critical value from the normal table rather than the *t* table, would the result of (c) agree better with the result of (d)? Why?

f. Interpret your results. Do the blueberries make a substantial difference?

85. For the data of the previous exercise, we now want to test the hypothesis of equal population means.

a. Carry out a two-tailed test using the two-sample *t* test. Although this test requires normal data, it will still work pretty well for moderately nonnormal data. Nevertheless, you should check the data for normality to see if the test is justified.

b. Carry out a two-tailed permutation test for the hypothesis of equal population means.

c. Compare the results of (a) and (b). Would you expect them to be similar for the data of this problem? Discuss their relationship to the results of the previous exercise. Summarize your conclusions about the effectiveness of blueberries.

86. Researchers at the University of Alaska have been trying to find inexpensive feed sources for Alaska reindeer growers ("Effects of Two Barley-Based Diets on Body Mass and Intake Rates of Captive Reindeer During Winter," Poster Presentation: School of Agriculture and Land Resources

Management, University of Alaska Fairbanks, 2002). They are focusing on Alaska-grown barley because commercially available feed supplies are too expensive for farmers. Typically, reindeer lose weight in the fall and winter, and the researchers are searching for a feed to minimize this loss. Thirteen pregnant reindeer were randomly divided into two groups to be fed on two different varieties of barley, thual and finaska. Here are the weight gains between October 1 and December 15 for the seven that were fed thual barley (T) and the six that were fed finaska barley (F).

T:	−5.83	−11.5	−5.5	−1.33	−3.83	−3.33	−7.17
F:	−0.17	−0.67	−4	−3		−1.33	−0.5

The weight gains are all negative, indicating that all of the animals lost weight. The thual barley is less fibrous and more digestible, and the intake rates for the two varieties of barley were very nearly the same, so the experimenters expected less weight loss for the thual variety.

a. Determine a 95% confidence interval for the difference of population means using the two-sample t interval.
b. Use software to generate a bootstrap sample of differences of means. Check the bootstrap distribution for normality using a normal probability plot.
c. Use the standard deviation of the bootstrap distribution along with the mean and t critical value from (a) to obtain a 95% confidence interval for the difference of means.
d. Use the bootstrap sample and the percentile method to obtain a 95% confidence interval for the difference of means.
e. Compare your three confidence intervals. If they are very similar, why do you think this is the case?
f. Interpret your results. Is there a substantial difference? Is it in the direction anticipated by the experimenters?

87. Consider using the data of the previous exercise to test the hypothesis of equal population variances.

a. Carry out a two-tailed test using the method of Section 10.5. Recall that this method requires the data to be normal, and the method is sensitive to departures from normality. Check the data for normality to see if the F test is justified.
b. Carry out a two-tailed permutation test for the hypothesis of equal population variances (or standard deviations).
c. Compare the two results and summarize your conclusions.

88. Recall the scenario from Example 10.8 about the experiment in the low-level college mathematics course. Here are the 85 final exam scores for those in the experimental group (E) and the 79 final exam scores for those in the control group (C):

E:	34	27	26	33	23	37	24	34	22	23	32	5	30
	29	0	30	34	26	28	27	32	29	31	33	28	21
	28	35	30	34	9	38	9	27	25	33	9	23	32
	28	38	35	16	37	25	34	38	34	31	35	28	25
	37	28	26	29	22	33	31	23	37	34	29	33	6
	8	29	36	7	21	30	28	34	25	37	28	23	26
	34	32	34	0	24	30	36	31					
C:	37	22	29	29	33	22	32	36	29	6	4	37	0
	35	28	33	35	24	21	0	32	28	27	8	30	37
	35	25	29	3	33	33	28	32	39	20	32	22	24
	38	22	29	29	36	0	32	27	7	19	35	26	22
	28	28	32	9	33	30	36	28	3	8	31	29	9
	0	0	20	32	7	8	33	29	9	0	30	26	25
	32	29											

a. Determine a 95% confidence interval for the difference of population means using a two-sample t interval.
b. Use software to generate a bootstrap sample of differences of means. Check the bootstrap distribution for normality using a normal probability plot.
c. Use the standard deviation of the bootstrap distribution along with the mean and t critical value from (a) to get a 95% confidence interval for the difference of means.

d. Use the bootstrap sample and the percentile method to obtain a 95% confidence interval for the difference of means.

e. Compare your three confidence intervals. If they are very similar, why do you think this is the case? In the light of your results for (c) and (d), does the two-sample t interval of (a) seem to work, regardless of normality? Explain.

f. Are your results consistent with the results of Example 10.8? Explain.

89. Return to the data of Example 10.8.

a. Carry out a two-tailed permutation test for the hypothesis of equal population means.

b. Compare the results for (a) and Example 10.8. Why should you have expected (a) and Example 10.8 to give similar results?

90. For the data of Example 10.8 it might be more appropriate to compare medians.

a. Find the medians for the two groups. With the help of a stem-and-leaf display for each group, explain why the medians are much closer than the means.

b. Carry out a two-tailed permutation test to compare population medians. Given what you found in (a), explain why the result of the permutation test was to be expected.

91. Two students, Miguel Melo and Cody Watson, compared textbook prices at the campus bookstore and Amazon.com. To be fair, they included the sales tax for the local store and added shipping for Amazon. Here are the prices for a sample of 27 books.

Campus	Amazon	Campus	Amazon
100.41	106.94	59.50	69.24
99.34	113.94	87.66	73.84
51.53	61.44	26.56	33.98
20.45	31.59	44.63	40.39
28.69	29.89	96.69	117.99
70.66	83.94	18.06	27.94
98.81	107.74	103.06	115.74
111.56	115.99	14.61	24.69

(continued)

Campus	Amazon	Campus	Amazon
97.22	108.29	77.03	88.04
61.89	78.44	99.34	113.94
70.39	82.94	81.81	90.74
58.17	65.74	48.88	58.94
108.38	122.09	76.50	91.94
61.63	63.49		

a. Determine a 95% confidence interval for the difference of population means using the t method of Section 10.3. Check the data for normality. Even if the normality assumption is not valid here, explain why the t method (or the z method of Section 10.1) might still be appropriate.

b. Based on the 27 differences, use software to obtain a bootstrap sample of mean differences. Check the bootstrap distribution for normality.

c. Use the standard deviation of the bootstrap distribution along with the mean and t critical value from (a) to get a 95% confidence interval for the difference of means.

d. Use the bootstrap sample and the percentile method to obtain a 95% confidence interval for the difference of means.

e. Compare your three confidence intervals. In the light of your results for (d), does nonnormality invalidate the results of (a) and (c)? Explain.

f. Interpret your results. Is there a substantial difference between the two ways to buy books? Assuming that the populations remain unchanged and you have just these two sources, where would you buy?

92. Consider testing the hypothesis of equal population means based on the data in the previous exercise.

a. Carry out a two-tailed test using the method of Section 10.3. Is the normality assumption satisfied here? If not, why might the test be valid anyway?

b. Carry out a two-tailed permutation test for the hypothesis of equal population means.

c. Compare the results for (a) and (b). If the two results are similar, does it tend to validate (a), regardless of normality?

93. Compare bootstrapping with approximate permutation tests in which random permutations are used. Discuss the similarities and differences.

94. Assume that X is uniformly distributed on $[-1, 1]$ and the Y distribution is uniform on the two intervals $[-101, -100]$ and $[100, 101]$. Thus the means are both 0, but the variances differ substantially. We take random samples of size three from each distribution and apply a permutation test for the null hypothesis H_0: $\mu_1 = \mu_2$ against the alternative H_a: $\mu_1 < \mu_2$.

   a. Show that the probability is 1/8 that all three of the Y values come from the interval $[100, 101]$.
   b. Show that, if all three Y values come from $[100, 101]$, then the P-value for the permutation test is .05.
   c. Explain why (a) and (b) are in conflict. What is the probability that the permutation test rejects the null hypothesis at the .05 level?

Supplementary Exercises: (95–124)

95. A group of 115 University of Iowa students was randomly divided into a build-up condition group ($m = 56$) and a scale-down condition group ($n = 59$). The task for each subject was to build his or her own pizza from a menu of 12 ingredients. The build-up group was told that a basic cheese pizza costs \$5 and that each extra ingredient would cost 50 cents. The scale-down group was told that a pizza with all 12 ingredients (ugh!!!) would cost \$11 and that deleting an ingredient would save 50 cents. The article "A Tale of Two Pizzas: Building Up from a Basic Product Versus Scaling Down from a Fully Loaded Product" (*Market. Lett.* 2002: 335–344) reported that the mean number of ingredients selected by the scale-down group was significantly greater than the mean number for the build-up group: 5.29 versus 2.71. The calculated value of the appropriate t statistic was 6.07. Would you reject the null hypothesis of equality in favor of inequality at a significance level of .05? .01? .001? Can you think of other products aside from pizza where one could build up or scale down? [*Note*: A separate experiment involved students from the University of Rome, but details were a bit different because there are typically not so many ingredient choices in Italy.]

96. Is the number of export markets in which a firm sells its products related to the firm's return on sales? The article "Technology Industry Success: Strategic Options for Small and Medium Firms" (*Bus. Horizons*, Sept.–Oct. 2003: 41–46) gave the accompanying information on the number of export markets for one group of firms whose return on sales was less than 10% and another group whose return was at least 10%.

Return	Sample size	Sample mean	Sample SD
Less than 10%	36	5.12	.57
At least 10%	47	8.26	1.20

The investigators reported that an appropriate test of hypotheses resulted in a P-value between .01 and .05. What hypotheses do you think were tested, and do you agree with the stated P-value information? What assumptions if any are needed in order to carry out the test? Can the plausibility of these assumptions be investigated based just on the foregoing summary data? Explain.

97. Suppose when using a two-sample t procedure that $m < n$, and show that $v > m - 1$. (This is why some authors suggest using $\min(m - 1, n - 1)$ as df in place of Welch's formula). What impact does this have on the CI and test procedure?

98. The accompanying summary data on compression strength (lb) for $12 \times 10 \times 8$ in. boxes appeared in the article "Compression of Single-Wall Corrugated Shipping

Containers Using Fixed and Floating Test Platens" (*J. Testing Eval.* 1992: 318–320). The authors stated that "the difference between the compression strength using fixed and floating platen method was found to be small compared to normal variation in compression strength between identical boxes." Do you agree?

Method	Sample size	Sample mean	Sample SD
Fixed	10	807	27
Floating	10	757	41

99. The authors of the article "Dynamics of Canopy Structure and Light Interception in *Pinus elliotti*, North Florida" (*Ecol. Monogr.* 1991: 33–51) planned an experiment to determine the effect of fertilizer on a measure of leaf area. A number of plots were available for the study, and half were selected at random to be fertilized. To ensure that the plots to receive the fertilizer and the control plots were similar, before beginning the experiment tree density (the number of trees per hectare) was recorded for eight plots to be fertilized and eight control plots, resulting in the given data. Minitab output follows.

Fertilizer   1024 1216 1312 1280 1216 1312 992   1120
plots
Control   1104 1072 1088 1328 1376 1280 1120 1200
plots

```
Two sample T for fertilize vs. control

            N    Mean   Std. Dev.   SE Mean
Fertilize   8    1184      126         44
Control     8    1196      118         42

95% CI for mu fertilize-mu control:
(-144, 120)
```

a. Construct a comparative boxplot and comment on any interesting features.
b. Would you conclude that there is a significant difference in the mean tree density for fertilizer and control plots? Use $\alpha = .05$.

c. Interpret the given confidence interval.

100. Is the response rate for questionnaires affected by including some sort of incentive to respond along with the questionnaire? In one experiment, 110 questionnaires with no incentive resulted in 75 being returned, whereas 98 questionnaires that included a chance to win a lottery yielded 66 responses ("Charities, No; Lotteries, No; Cash, Yes," *Public Opinion Q.* 1996: 542–562). Does this data suggest that including an incentive increases the likelihood of a response? State and test the relevant hypotheses at significance level .10 by using the *P*-value method.

101. The article "Quantitative MRI and Electrophysiology of Preoperative Carpal Tunnel Syndrome in a Female Population" (*Ergonomics* 1997: 642–649) reported that $(-473.3, 1691.9)$ was a large-sample 95% confidence interval for the difference between true average thenar muscle volume (mm^3) for sufferers of carpal tunnel syndrome and true average volume for non-sufferers. Calculate and interpret a 90% confidence interval for this difference.

102. The following summary data on bending strength (lb-in/in) of joints is taken from the article "Bending Strength of Corner Joints Constructed with Injection Molded Splines" (*Forest Prod. J.* April 1997: 89–92). Assume normal distributions.

Type	Sample size	Sample mean	Sample SD
Without side coating	10	80.95	9.59
With side coating	10	63.23	5.96

a. Calculate a 95% lower confidence bound for true average strength of joints with a side coating.
b. Calculate a 95% lower prediction bound for the strength of a single joint with a side coating.

c. Calculate a 95% confidence interval for the difference between true average strengths for the two types of joints.

103. An experiment was carried out to compare various properties of cotton/polyester spun yarn finished with softener only and yarn finished with softener plus 5% DP-resin ("Properties of a Fabric Made with Tandem Spun Yarns," *Textile Res. J.* 1996: 607–611). One particularly important characteristic of fabric is its durability, that is, its ability to resist wear. For a sample of 40 softener-only specimens, the sample mean stoll-flex abrasion resistance (cycles) in the filling direction of the yarn was 3975.0, with a sample standard deviation of 245.1. Another sample of 40 softener-plus specimens gave a sample mean and sample standard deviation of 2795.0 and 293.7, respectively. Calculate a confidence interval with confidence level 99% for the difference between true average abrasion resistances for the two types of fabrics. Does your interval provide convincing evidence that true average resistances differ for the two types of fabrics? Why or why not?

104. The derailment of a freight train due to the catastrophic failure of a traction motor armature bearing provided the impetus for a study reported in the article "Locomotive Traction Motor Armature Bearing Life Study" (*Lubricat. Engr.* August 1997: 12–19). A sample of 17 high-mileage traction motors was selected, and the amount of cone penetration (mm/10) was determined both for the pinion bearing and for the commutator armature bearing, resulting in the following data:

Motor	1	2	3	4	5	6
Commutator	211	273	305	258	270	209
Pinion	226	278	259	244	273	236
Motor	7	8	9	10	11	12
Commutator	223	288	296	233	262	291
Pinion	290	287	287	242	288	242
Motor	13	14	15	16	17	
Commutator	278	275	210	272	264	
Pinion	278	208	281	274	274	

Calculate an estimate of the population mean difference between penetration for the commutator armature bearing and penetration for the pinion bearing, and do so in a way that conveys information about the reliability and precision of the estimate. [*Note*: A normal probability plot validates the necessary normality assumption.] Would you say that the population mean difference has been precisely estimated? Does it look as though population mean penetration differs for the two types of bearings? Explain.

105. The article "Two Parameters Limiting the Sensitivity of Laboratory Tests of Condoms as Viral Barriers" (*J. Test. Eval.* 1996: 279–286) reported that, in brand A condoms, among 16 tears produced by a puncturing needle, the sample mean tear length was 74.0 μm, whereas for the 14 brand B tears, the sample mean length was 61.0 μm (determined using light microscopy and scanning electron micrographs). Suppose the sample standard deviations are 14.8 and 12.5, respectively (consistent with the sample ranges given in the article). The authors commented that the thicker brand B condom displayed a smaller mean tear length than the thinner brand A condom. Is this difference in fact statistically significant? State the appropriate hypotheses and test at $\alpha = .05$.

106. Information about hand posture and forces generated by the fingers during manipulation of various daily objects is needed for designing high-tech hand prosthetic devices. The article "Grip Posture and Forces During Holding Cylindrical Objects with Circular Grips" (*Ergonomics* 1996: 1163–1176) reported that for a sample of 11 females, the sample mean four-finger pinch strength (N) was 98.1 and the sample standard deviation was 14.2. For a sample of 15 males, the sample mean and sample standard deviation were 129.2 and 39.1, respectively.

a. A test carried out to see whether true average strengths for the two genders were different resulted in $t = 2.51$ and P-value = .019. Does the appropriate test procedure described in this chapter yield this value of t and the stated P-value?

b. Is there substantial evidence for concluding that true average strength for males exceeds that for females by more than 25 N? State and test the relevant hypotheses.

107. After the Enron scandal in the fall of 2001, faculty in accounting began to incorporate ethics more into accounting courses. One study looked at the effectiveness of such educational interventions "pre-Enron" and "post-Enron." The data below shows students' improvement in score on the Accounting Ethical Dilemma Instrument (AEDI) across a one-semester accounting class in Spring 2001 ("pre-Enron") and another in Spring 2002 ("post-Enron"). (From "A Note in Ethics Educational Interventions in an Undergraduate Auditing Course: Is There an 'Enron Effect'?" *Issues Account. Educ.* 2004: 53–71.)

Class	n	Improvement in AEDI score	
		Mean	SD
2001 (pre-Enron)	37	5.48	13.83
2002 (post-Enron)	21	6.31	13.20

a. Test to see whether the 2001 class showed a statistically significant improvement in AEDI score across the semester.

b. Test to see whether the 2002 class showed a statistically significant improvement in AEDI score across the semester.

c. Test to see whether the 2002 class showed a statistically significantly greater improvement in AEDI score than the 2001 class. In this respect, does there appear to be an "Enron effect"?

108. Torsion during hip external rotation (ER) and extension may be responsible for certain kinds of injuries in golfers and other athletes. The article "Hip Rotational Velocities during the Full Golf Swing" (*J. Sport Sci. Med.* 2009: 296–299) reported on a study in which peak ER velocity and peak IR (internal rotation) velocity (both in deg/s) were determined for a sample of 15 female collegiate golfers during their swings. The following data was supplied by the article's authors.

Golfer	ER	IR	Diff.
1	−130.6	−98.9	−31.7
2	−125.1	−115.9	−9.2
3	−51.7	−161.6	109.9
4	−179.7	−196.9	17.2
5	−130.5	−170.7	40.2
6	−101.0	−274.9	173.9
7	−24.4	−275.0	250.6
8	−231.1	−275.7	44.6
9	−186.8	−214.6	27.8
10	−58.5	−117.8	59.3
11	−219.3	−326.7	107.4
12	−113.1	−272.9	159.8
13	−244.3	−429.1	184.8
14	−184.4	−140.6	−43.8
15	−199.2	−345.6	146.4

a. Is it plausible that the differences came from a normally distributed population?

b. The article reported that mean(sd) = −145.3(68.0) for ER velocity and = −227.8(96.6) for IR velocity. Based just on this information, could a test of hypotheses about the difference between true average IR velocity and true average ER velocity be carried out? Explain.

c. Do an appropriate hypothesis test about the difference between true average IR velocity and true average ER velocity and interpret the result.

109. The accompanying summary data on the ratio of strength to cross-sectional area for knee extensors is taken from the article "Knee Extensor and Knee Flexor Strength: Cross-Sectional Area Ratios in Young and Elderly Men" (*J. Gerontol.* 1992: M204–M210).

Group	Sample size	Sample mean	Standard error
Young	13	7.47	.22
Elderly men	12	6.71	.28

Does this data suggest that the true average ratio for young men exceeds that for elderly men? Carry out a test of appropriate hypotheses using $\alpha = .05$. Be sure to state any assumptions necessary for your analysis.

110. The accompanying data on response time appeared in the article "The Extinguishment of Fires Using Low-Flow Water Hose Streams—Part II" (*Fire Tech.* 1991: 291–320). The samples are independent, not paired.

Good visibility	.43	1.17	.37	.47	.68	.58	.50	2.75
Poor visibility	1.47	.80	1.58	1.53	4.33	4.23	3.25	3.22

The authors analyzed the data with the pooled *t* test. Does the use of this test appear justified? [*Hint*: Check for normality.]

111. The accompanying data on the alcohol content of wine is representative of that reported in a study in which wines from the years 1999 and 2000 were randomly selected and the actual content was determined by laboratory analysis (*London Times* August 5, 2001).

Wine	1	2	3	4	5	6
Actual	14.2	14.5	14.0	14.9	13.6	12.6
Label	14.0	14.0	13.5	15.0	13.0	12.5

The two-sample *t* test gives a test statistic value of .62 and a two-tailed *P*-value of .55. Does this convince you that there is no significant difference between true average actual alcohol content and true average content stated on the label? Explain.

112. The article "The Accuracy of Stated Energy Contents of Reduced-Energy, Commercially Prepared Foods" (*J. Am. Diet. Assoc.* 2010: 116–123) presented the accompanying data on vendor-stated gross energy and measured value (both in kcal) for 10 different supermarket convenience meals):

Meal	1	2	3	4	5
Stated	180	220	190	230	200
Measured	212	319	231	306	211
Meal	6	7	8	9	10
Stated	370	250	240	80	180
Measured	431	288	265	145	228

Obtain a 95% confidence interval for the difference of population means. By roughly what percentage are the actual calories higher than the stated value?

Note that the article calls this a convenience sample and suggests that therefore it should have limited value for inference. However, even if the ten meals were a random sample from their local store, there could still be a problem in drawing conclusions about a purchase at your store.

113. How does energy intake compare to energy expenditure? One aspect of this issue was considered in the article "Measurement of Total Energy Expenditure by the Doubly Labelled Water Method in Professional Soccer Players" (*J. Sports Sci.* 2002: 391–397), which contained the accompanying data (MJ/day).

Player	1	2	3	4	5	6	7
Expenditure	14.4	12.1	14.3	14.2	15.2	15.5	17.8
Intake	14.6	9.2	11.8	11.6	12.7	15.0	16.3

Test to see whether there is a significant difference between intake and expenditure. Does the conclusion depend on whether a significance level of .05, .01, or .001 is used?

114. An experimenter wishes to obtain a CI for the difference between true average breaking strength for cables manufactured by company I and by company II. Suppose breaking strength is normally distributed for both types of cable with $\sigma_1 = 30$ psi and $\sigma_2 = 20$ psi.

   a. If costs dictate that the sample size for the type I cable should be three times the sample size for the type II cable,

how many observations are required if the 99% CI is to be no wider than 20 psi?

b. Suppose a total of 400 observations is to be made. How many of the observations should be made on type I cable samples if the width of the resulting interval is to be a minimum?

115. To assess the tendency of people to rationalize poor performance, 246 college students were randomly assigned to one of two groups: a negative feedback group and a positive feedback group. All students took a test which asked them to identify people's emotions based on photographs of their faces. Those in the negative feedback group were all given D grades, while those in the positive feedback group received A's (regardless of how they actually performed). A follow-up questionnaire asked students to assess the validity of the test and the importance of being able to read people's faces. The results of these two follow-up surveys appear below.

Group	n	Test validity rating $\bar{x}$	s	Face reading importance rating $\bar{x}$	s
Positive feedback	123	6.95	1.09	6.62	1.19
Negative feedback	123	5.51	0.79	5.36	1.00

a. Test the hypothesis that negative feedback is associated with a lower average validity rating than positive feedback at the $\alpha = .01$ level.

b. Test the hypothesis that students receiving positive feedback rate face-reading as more important, on average, than do students receiving negative feedback. Again use a 1% significance level.

c. Is it reasonable to conclude that the results seen in parts (a) and (b) are attributable to the different types of feedback? Why or why not?

116. The insulin-binding capacity (pmol/mg protein) was measured for four different groups of rats: (1) nondiabetic, (2) untreated diabetic, (3) diabetic treated with a low dose of insulin, (4) diabetic treated with a high dose of insulin. The accompanying table gives sample sizes and sample standard deviations. Denote the sample size for the ith treatment by n_i and the sample variance by $S_i^2 (i = 1, 2, 3, 4)$. Assuming that the true variance for each treatment is σ^2, construct a pooled estimator of σ^2 that is unbiased, and verify using rules of expected value that it is indeed unbiased. What is your estimate for the following actual data? [*Hint*: Modify the pooled estimator S_p^2 from Section 10.2.]

	Treatment			
	1	2	3	4
Sample size	16	18	8	12
Sample SD	.64	.81	.51	.35

117. Suppose a level .05 test of H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 > 0$ is to be performed, assuming $\sigma_1 = \sigma_2 = 10$ and normality of both distributions, using equal sample sizes ($m = n$). Evaluate the probability of a type II error when $\mu_1 - \mu_2 = 1$ and $n = 25, 100, 2500$, and 10,000. Can you think of real situations in which the difference $\mu_1 - \mu_2 = 1$ has little practical significance? Would sample sizes of $n = 10,000$ be desirable in such situations?

118. Are male college students more easily bored than their female counterparts? This question was examined in the article "Boredom in Young Adults—Gender and Cultural Comparisons" (*J. Cross-Cult. Psych.* 1991: 209–223). The authors administered a scale called the Boredom Proneness Scale to 97 male and 148 female U.S. college students. Does the accompanying data support the research hypothesis that the mean Boredom Proneness Rating is higher for men than for women? Test the appropriate hypotheses using a .05 significance level.

Sex	Sample size	Sample mean	Sample SD
Male	97	10.40	4.83
Female	148	9.26	4.68

119. Researchers sent 5000 resumes in response to job ads that appeared in the *Boston Globe* and *Chicago Tribune*. The resumes were identical except that 2500 of them had "white sounding" first names, such as Brett and Emily, whereas the other 2500 had "black sounding" names such as Tamika and Rasheed. The resumes of the first type elicited 250 responses and the resumes of the second type only 167 responses (these numbers are consistent with information that appeared in a January 15, 2003, report by the Associated Press). Does this data strongly suggest that a resume with a "black" name is less likely to result in a response than is a resume with a "white" name?

120. Is touching by a coworker sexual harassment? This question was included on a survey given to federal employees, who responded on a scale of 1–5, with 1 meaning a strong negative and 5 indicating a strong yes. The table summarizes the results.

Sex	Sample size	Sample mean	Sample SD
Female	4343	4.6056	.8659
Male	3903	4.1709	1.2157

Of course, with 1–5 being the only possible values, the normal distribution does not apply here, but the sample sizes are sufficient that it does not matter. Obtain a two-sided confidence interval for the difference in population means. Does your interval suggest that females are more likely than males to regard touching as harassment? Explain your reasoning.

121. Let $X_1, \ldots, X_m$ be a random sample from a Poisson distribution with parameter μ_1, and let $Y_1, \ldots, Y_n$ be a random sample from another Poisson distribution with parameter μ_2. We wish to test $H_0: \mu_1 - \mu_2 = 0$ against one of the three standard alternatives. When m and n are large, the CLT justifies using a large-sample z test. However, the fact that $V(\overline{X}) = \mu/n$ suggests that a different denominator should be used in

standardizing $\overline{X} - \overline{Y}$. Develop a large-sample test procedure appropriate to this problem, and then apply it to the following data to test whether the plant densities for a particular species are equal in two different regions (where each observation is the number of plants found in a randomly located square sampling quadrat having area 1 m^2, so for region 1, there were 40 quadrats in which one plant was observed, etc.):

	Frequency								
	0	1	2	3	4	5	6	7	
Region 1	28	40	28	17	8	2	1	1	$m = 125$
Region 2	14	25	30	18	49	2	1	1	$n = 140$

122. Referring to the previous exercise, develop a large-sample confidence interval formula for $\mu_1 - \mu_2$. Calculate the interval for the data given there using a confidence level of 95%.

123. Refer back to the pooled t procedures described at the end of Section 10.2. The test statistic for testing $H_0: \mu_1 - \mu_2 = \Delta_0$ is

$$T_p = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{\sqrt{\frac{S_p^2}{m} + \frac{S_p^2}{n}}} = \frac{(\overline{X} - \overline{Y}) - \Delta_0}{S_p \sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

Show that when $\mu_1 - \mu_2 = \Delta'$ (some alternative value for the difference), then T_p has a noncentral t distribution with df $= m + n - 2$ and noncentrality parameter

$$\delta = \frac{\Delta' - \Delta_0}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

[*Hint:* Look back at Exercises 39–40, as well as Chapter 9 Exercise 38.]

124. Let R_1 be a rejection region with significance level α for testing $H_{01}: \theta \in \Omega_1$ versus $H_{a1}: \theta \notin \Omega_1$, and let R_2 be a level α rejection region for testing $H_{02}: \theta \in \Omega_2$ versus $H_{a2}: \theta \notin \Omega_2$, where Ω_1 and Ω_2 are two disjoint sets of possible values of θ. Now

consider testing H_0: $\theta \in \Omega_1 \cup \Omega_2$ versus the alternative H_a: $\theta \notin \Omega_1 \cup \Omega_2$. The proposed rejection region is $R_1 \cap R_2$. That is, H_0 is rejected only if both H_{01} and H_{02} can be rejected. This procedure is called a *union–intersection test* (UIT).

a. Show that the UIT is a level α test.

b. As an example, let μ_T denote the mean value of a particular variable for a generic (test) drug, and μ_R denote the mean value of this variable for a brand-name (reference) drug. In *bioequivalence* testing, the relevant hypotheses are H_0: $\mu_T/\mu_R \leq \delta_L$ or $\mu_T/\mu_R \geq \delta_U$ (the two aren't bioequivalent) versus H_a: $\delta_L < \mu_T/\mu_R < \delta_U$ (bioequivalent). The limits δ_L and δ_U are standards set by regulatory agencies; the FDA often uses .80 and 1.25 = 1/.8, respectively. By

taking logarithms and letting $\eta = \ln(\mu)$, $\tau = \ln(\delta)$, the hypotheses become H_0: either $\eta_T - \eta_R \leq \tau_L$ or $\geq \tau_U$ versus H_a: $\tau_L < \eta_T - \eta_R < \tau_U$. With this setup, a type I error involves saying the drugs are bioequivalent when they are not. The FDA mandates $\alpha = .05$.

Let D be an estimator of $\eta_T - \eta_R$ with standard error S_D such that standardized variable $T = [D - (\eta_T - \eta_R)]/S_D$ has a t distribution with v df. The standard test procedure is referred to as *TOST* for "two one-sided tests" and is based on the two test statistics $T_U = (D - \tau_U)/S_D$ and $T_L = (D - \tau_L)/S_D$. If $v = 20$, state the appropriate conclusion in each of the following cases: (1) $\tau_L = 2.0$, $\tau_U = -1.5$; (2) $\tau_L = 1.5$, $\tau_U = -2.0$; (3) $\tau_L = 2.0$, $\tau_U = -2.0$.

The Analysis of Variance

<div align="right">

11

</div>

Introduction

In studying methods for the analysis of quantitative data, we first focused on problems involving a single sample of numbers and then turned to a comparative analysis of two different samples. Now we are ready for the analysis of several samples.

The **analysis of variance**, or more briefly **ANOVA**, refers broadly to a collection of statistical procedures for the analysis of quantitative responses. The simplest ANOVA problem is referred to variously as a **single-factor**, **single-classification**, or **one-way ANOVA** and involves the analysis of data sampled from two or more numerical populations (distributions). The characteristic that labels the populations is called the **factor** under study, and the populations are referred to as the **levels** of the factor. Examples of such situations include the following:

1. An experiment to study the effects of five different brands of gasoline on automobile engine operating efficiency (mpg)
2. An experiment to study the effects of four different sugar solutions (glucose, sucrose, fructose, and a mixture of the three) on bacterial growth
3. An experiment to investigate whether hardwood concentration in pulp has an effect on tensile strength of bags made from the pulp
4. An experiment to decide whether the color density of fabric specimens depends on the amount of dye used

In (1) the factor of interest is gasoline brand, and there are five different levels of the factor. The factor in (2) is sugar, with four levels (or five, if a control solution containing no sugar is used). The factor in both of these first two examples is categorical in nature, and the levels correspond to possible categories of the factor. In (3) and (4), the factors are concentration of hardwood and amount of dye, respectively; both these factors are quantitative in nature, so the levels identify different settings of the factor. When the factor of interest is quantitative, statistical techniques from regression analysis (discussed in Chapter 12) can also be used to analyze the data.

Here we first introduce single-factor ANOVA. Section 11.1 presents the F test for testing the null hypothesis that the population means are identical. Section 11.2 considers further analysis of the data when H_0 has been rejected. Section 11.3 covers some other aspects of single-factor ANOVA. Many experimental situations involve studying the simultaneous impact of more than one factor. Various aspects of two-factor ANOVA are considered in the last two sections of the chapter.

11.1   Single-Factor ANOVA

Single-factor ANOVA focuses on a comparison of two or more populations. For example, McDonalds may wish to compare the average revenue associated with three different advertising campaigns, or a team of animal nutritionists may carry out an experiment to compare the effect of five different diets on weight gain, or FedEx may want to compare the strengths of cardboard shipping boxes from four different vendors. Let

I = the number of populations/treatments being compared
μ_1 = the mean of population 1 (or the true average response when treatment 1 is applied)
$\vdots$
μ_I = the mean of population I (or the true average response when treatment I is applied)

Then the hypotheses of interest are

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_I$$

versus

$$H_a: \text{at least two of the } \mu_i\text{'s are different}$$

If $I = 4$, H_0 is true only if all four μ_i's are identical. H_a would be true, for example, if $\mu_1 = \mu_2 \neq \mu_3 = \mu_4$, if $\mu_1 = \mu_3 = \mu_4 \neq \mu_2$, or if all four μ_i's differ from each other. A test of these hypotheses requires that we have available a random sample from each population or treatment. Since ANOVA focuses on a comparison of *means*, you may wonder why the method is called analysis of *variance* (actually, analysis of *variability* would be a better name). The following example illustrates why it is appropriate to consider variability.

Example 11.1 The article "Compression of Single-Wall Corrugated Shipping Containers Using Fixed and Floating Test Platens" (*J. Test. Eval.* 1992: 318–320) describes an experiment in which several different types of boxes were compared with respect to compression strength (lb). Table 11.1 presents the results of an experiment involving $I = 4$ types of boxes (the sample means and standard deviations are in good agreement with values given in the article).

Table 11.1 The data and summary quantities for Example 11.1

Type of box	Compression strength (lb)			Sample mean	Sample SD
1	655.5	788.3	734.3	713.00	46.55
	721.4	679.1	699.4		
2	789.2	772.5	786.9	756.93	40.34
	686.1	732.1	774.8		
3	737.1	639.0	696.3	698.07	37.20
	671.7	717.2	727.1		
4	535.1	628.7	542.4	562.02	39.87
	559.0	586.9	520.0		
			Grand mean =	682.50	

With μ_i denoting the true average compression strength for boxes of type i ($i = 1, 2, 3, 4$), the null hypothesis is $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$. Figure 11.1a shows a comparative boxplot for the four samples. There is a substantial amount of overlap among observations on the first three box types, but compression strengths for the fourth type appear considerably smaller than for the others. This suggests that H_0 is not true.
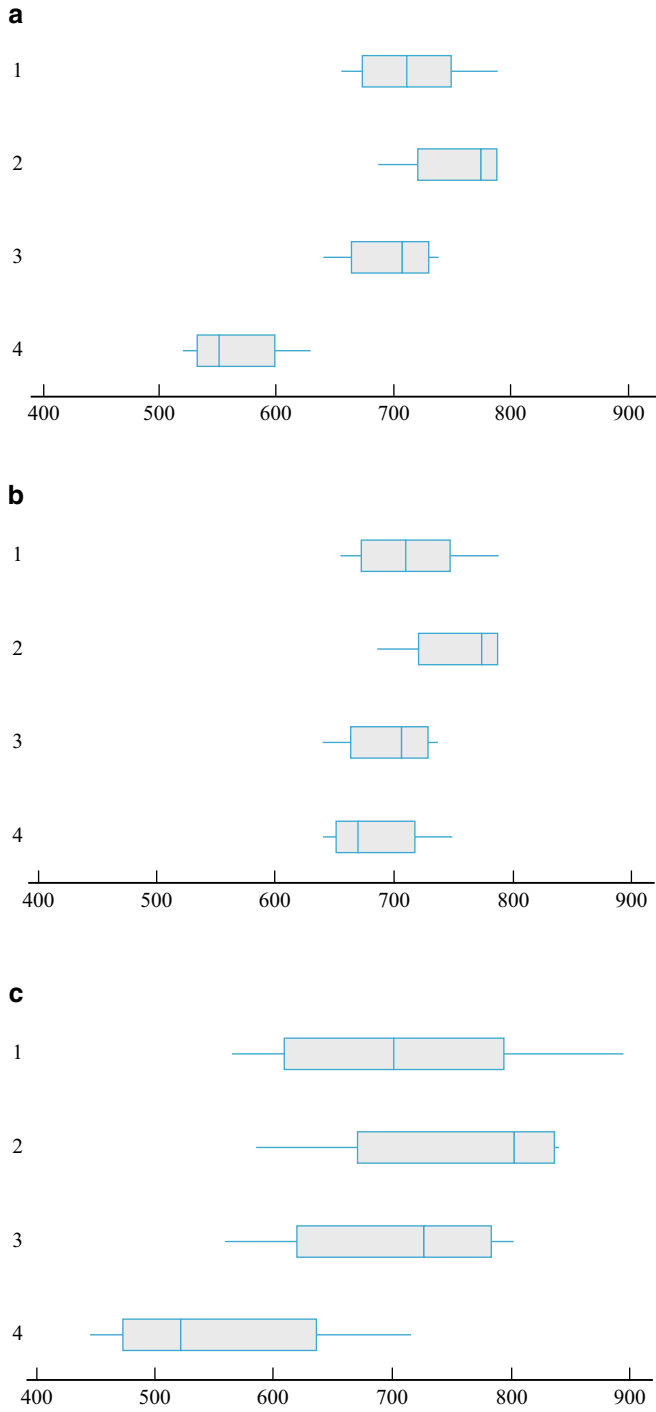
Figure 11.1 Boxplots for Example 11.1: (**a**) original data; (**b**) data with one mean altered; (**c**) data with standard deviations altered

The comparative boxplot in Figure 11.1b is based on adding 120 to each observation in the fourth sample (giving a mean of 682.02 and the same standard deviation) and leaving the other samples unaltered. Because the sample means are now closer together, it is no longer obvious whether H_0 should be rejected.

Lastly, the comparative boxplot in Figure 11.1c is based on inflating the standard deviation of each sample while maintaining the values of the original sample means. Once again, it is unclear from the graph whether H_0 is true, even though now the sample means are separated by the same amounts as they were in Figure 11.1a.

These graphs suggest that it's insufficient to consider only how far apart the sample means are in assessing whether the population means are different; we must also account for the amount of variability *within* each of the I samples.   ∎

Notation and Assumptions

In two-sample problems, we used the letters X and Y to designate the observations in the two samples. Because this is cumbersome for three or more samples, it is customary to use a single letter with two subscripts. The first subscript identifies the sample number, corresponding to the population or treatment being sampled, and the second subscript denotes the position of the observation within that sample. Let

X_{ij} = the random variable denoting the jth measurement from the ith population or treatment
x_{ij} = the observed value of X_{ij} when the experiment is performed

The observed data is often displayed in a rectangular table, such as Table 11.1. There, samples from the different populations appear in different rows of the table, and x_{ij} is the jth number in the ith sample. For example, $x_{23} = 786.9$ (the third observation from the second population), and $x_{41} = 535.1$. When there is potential ambiguity, we will write $x_{i,j}$ rather than x_{ij} (e.g., if there were 15 observations on each of 12 treatments, x_{112} could mean $x_{1,12}$ or $x_{11,2}$). It is assumed that the X_{ij}'s within any particular sample are independent—a random sample from the ith population or treatment distribution—and that different samples are independent of each other.

In some studies, different samples contain different numbers of observations. However, the concepts and methods of single-factor ANOVA are most easily developed for the case of equal sample sizes, known as a **balanced** study design. Unequal sample sizes will be considered in Section 11.3. Restricting ourselves for the moment to balanced designs, let J denote the number of observations in each sample ($J = 6$ in Example 11.1). The data set consists of $n = IJ$ observations. The sample means will be denoted by $\overline{X}_{1.}, \overline{X}_{2.}, \ldots, \overline{X}_{I.}$. That is,

$$\overline{X}_{i.} = \frac{1}{J} \sum_{j=1}^{J} X_{ij} \quad i = 1, 2, \ldots, I$$

Similarly, the average of all IJ observations, called the **grand mean**, is

$$\overline{X}_{..} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} X_{ij} = \frac{1}{I} \sum_{i=1}^{I} \overline{X}_{i.}$$

For the strength data in Table 11.1, $\bar{x}_{1.} = 713.00$, $\bar{x}_{2.} = 756.93$, $\bar{x}_{3.} = 698.07$, $\bar{x}_{4.} = 562.02$, and $\bar{x}_{..} = 682.50$. Additionally, let $S_1^2, S_2^2, \ldots, S_I^2$ represent the sample variances:

$$S_i^2 = \frac{1}{J-1} \sum_{j=1}^{J} \left(X_{ij} - \overline{X}_{i\cdot} \right)^2 \quad i = 1, 2, \ldots, I$$

From Example 11.1, $s_1 = 46.55$, $s_1^2 = 2166.90$, and so on.

ANOVA ASSUMPTIONS   The I population or treatment distributions are all normal with the same variance σ^2. That is, the X_{ij}'s are independent and normally distributed with

$$E\left(X_{ij}\right) = \mu_i \quad V\left(X_{ij}\right) = \sigma^2$$

The plausibility of independence of the samples (and of the individual observations within the samples) stems from a study's design. At the end of this section we discuss methods for checking the plausibility of the normality and equal variance assumptions.

Sums of Squares and Mean Squares

Example 11.1 suggests the need for two distinct measures of variation: *between-samples* variation (i.e., the disparity between the I sample means) and *within-samples* variation (assessing variation separately within each sample and then combining). The test procedure we will develop shortly is based on the following measures of variation in the data.

DEFINITION   A measure of between-samples variation is the **treatment sum of squares SSTr**, given by

$$\text{SSTr} = \sum_i \sum_j \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot} \right)^2 = \sum_i J \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot} \right)^2$$

$$= J \left[\left(\overline{X}_{1\cdot} - \overline{X}_{\cdot\cdot} \right)^2 + \cdots + \left(\overline{X}_{I\cdot} - \overline{X}_{\cdot\cdot} \right)^2 \right]$$

A measure of within-samples variation is the **error sum of squares SSE**, given by

$$\text{SSE} = \sum_i \sum_j \left(X_{ij} - \overline{X}_{i\cdot} \right)^2 = \sum_i \left[(J-1)S_i^2 \right]$$

$$= (J-1) \left[S_1^2 + S_2^2 + \cdots + S_I^2 \right]$$

Thus SSTr assesses variation in the means from the different samples, whereas SSE entails assessing variability within each sample separately (via the sample variance) and then combining these assessments.

Example 11.2 (Example 11.1 continued) For the box compression data displayed in Figure 11.1a,

$$\text{SSTr} = 6(713.00 - 682.50)^2 + 6(756.93 - 682.50)^2 + 6(698.07 - 682.50)^2 + 6(562.02 - 682.50)^2$$
$$= 127,374.7$$

while

$$\text{SSE} = (6 - 1)(46.55)^2 + (6 - 1)(40.34)^2 + (6 - 1)(37.20)^2 + (6 - 1)(39.87)^2$$
$$= 33,838.4$$

Similar computations can be applied to the two modified versions of the original data; the various sums of squares are summarized below.

	Figure 11.1a	Figure 11.1b	Figure 11.1c
SSTr	127,374.7	14,004.6	127,374.7
SSE	33,838.4	33,838.4	211,488.0

For the altered data on which Figure 11.1b is based, $\bar{x}_{4\cdot} = 682.02$ and the revised grand mean is $\bar{x}_{\cdot\cdot} = 712.50$. This greatly reduces SSTr, reflecting the fact that the sample means are less far apart in Figure 11.1b than in Figure 11.1a. Since the standard deviations of the samples were not changed, SSE for the data displayed in Figure 11.1a, b are identical.

For the data used to construct Figure 11.1c, the value of SSTr is unchanged from Figure 11.1a—the $\bar{x}_{i\cdot}$'s were not altered, so this measure of between-samples variation stays the same. On the other hand, SSE for Figure 11.1c is much larger than for the actual data. Since the altered data exhibits much greater within-samples variation than does the actual data, the corresponding SSE should be correspondingly greater. ∎

A descriptive understanding of the treatment and error sums of squares is provided by the following fundamental identity.

THEOREM (Fundamental ANOVA Identity)

$$\text{SSTr} + \text{SSE} = \text{SST}$$

where SST is the **total sum of squares** given by

$$\text{SST} = \sum_i \sum_j (x_{ij} - \bar{x}_{\cdot\cdot})^2$$

The proof of the identity follows from squaring both sides of the relationship

$$x_{ij} - \bar{x}_{\cdot\cdot} = (x_{ij} - \bar{x}_{i\cdot}) + (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot}) \tag{11.1}$$

and summing over all i and j. This gives SST on the left and SSTr and SSE as the two extreme terms on the right; the cross-product term is easily seen to be zero (Exercise 13).

The interpretation of the fundamental identity is an important aid to understanding ANOVA. SST is a measure of the total variation in the data—the sum of all squared deviations about the grand mean. The identity says that this total variation can be partitioned into two pieces. SSTr is the amount of variation (between samples) that can be explained by possible differences in the μ_i's: when the μ_i's

differ substantially, the individual sample means should be further from the grand mean than when the μ_i's are identical or close to one another. SSE measures variation that would be present (within samples) even if H_0 were true; thus, SSE the part of total variation that is *unexplained* by the truth or falsity of H_0. If explained variation is large relative to unexplained variation, then H_0 should be rejected in favor of H_a.

Formal inference will require the sampling distributions of both the statistics SSTr and SSE. Recall a result from Section 6.4: if $X_1, \ldots, X_n$ is a random sample from a normal distribution, then the sample mean $\overline{X}$ and the sample variance S^2 are independent. Also, $\overline{X}$ is normally distributed, and $(n-1)S^2/\sigma^2$ has a chi-squared distribution with $n-1$ df. Similar results hold in our ANOVA situation.

THEOREM   When the ANOVA assumptions are satisfied,

   1. SSE and SSTr are independent random variables.
   2. SSE$/\sigma^2$ has a chi-squared distribution with $IJ - I$ df.

Furthermore, when $H_0: \mu_1 = \cdots = \mu_I$ is true,

   3. SSTr$/\sigma^2$ has a chi-squared distribution with $I - 1$ df.

Proof Independence of SSTr and SSE follows from the fact that SSTr is based on the individual sample means whereas SSE is based on the sample variances, and $\overline{X}_{i\cdot}$ is independent of S_i^2 for each i. Next, SSE$/\sigma^2$ can be expressed as the sum of chi-squared rvs:

$$\frac{\text{SSE}}{\sigma^2} = \frac{(J-1)S_1^2}{\sigma^2} + \cdots + \frac{(J-1)S_I^2}{\sigma^2}$$

Each term in the sum has a χ_{J-1}^2 distribution, and dfs add because the samples are independent. Thus SSE$/\sigma^2$ also has a chi-squared distribution, with df $= (J-1) + \cdots + (J-1) = I(J-1) = IJ - I$.

Now suppose H_0 is true and let $Y_i = \overline{X}_{i\cdot}$ for $i = 1, \ldots, I$. Then $Y_1, Y_2, \ldots, Y_I$ are independent and normally distributed *with the same mean* and with variance σ^2/J. Thus, by the key result from Section 6.4, $(I-1)S_Y^2/(\sigma^2/J)$ has a chi-squared distribution with $I-1$ df when H_0 is true. Furthermore,

$$\frac{(I-1)S_Y^2}{\sigma^2/J} = \frac{(I-1)J}{\sigma^2} \cdot \frac{1}{I-1} \sum_i \left(Y_i - \overline{Y}\right)^2 = \frac{J}{\sigma^2} \sum \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2 = \frac{\text{SSTr}}{\sigma^2},$$

so under H_0, SSTr$/\sigma^2 \sim \chi_{I-1}^2$. ∎

SSTr and SSE provide measures of between- and within-samples variability, respectively, but they are not (yet) directly comparable. Analogous to the definition of sample variance in Chapter 1, wherein a sum of squares was divided by its degrees of freedom, we make the following definitions.

DEFINITION   The **mean square for treatments MSTr** and the **mean square for error MSE** are

$$\text{MSTr} = \frac{\text{SSTr}}{I-1} \qquad \text{MSE} = \frac{\text{SSE}}{IJ - I}$$

The word "mean" is again being used in the sense of average; a mean square is a sum of squares divided by its associated degrees of freedom. The next proposition sets the stage for our ultimate ANOVA hypothesis test.

PROPOSITION   When the ANOVA assumptions are satisfied,
$E(\text{MSE}) = \sigma^2$; that is, MSE is an unbiased estimator of σ^2.
Moreover, when $H_0: \mu_1 = \cdots = \mu_I$ is true,
$E(\text{MSTr}) = \sigma^2$; in this case, MSTr is also an unbiased estimator of σ^2.

Proof The expected value of a chi-squared variable with v df is just v. Thus, from the previous theorem,

$$E\left(\frac{\text{SSE}}{\sigma^2}\right) = IJ - I \Rightarrow E(\text{MSE}) = E\left(\frac{\text{SSE}}{IJ - I}\right) = \sigma^2$$

and

$$H_0 \text{ true} \Rightarrow E\left(\frac{\text{SSTr}}{\sigma^2}\right) = I - 1 \Rightarrow E(\text{MSTr}) = E\left(\frac{\text{SSTr}}{I - 1}\right) = \sigma^2 \qquad \blacksquare$$

MSTr is unbiased for σ^2 when H_0 is true, but what about when H_0 is false? It can be shown (Exercise 14) that in this case, $E(\text{MSTr}) > \sigma^2$. This is because the $\overline{X}_{i\cdot}$'s tend to differ more from each other, and therefore from the grand mean, when the μ_i's are not identical than when they are the same.

The *F* Test

It follows from the preceding discussion that when H_0 is true the values of MSTr and MSE should be close to each other. Equivalently, the *ratio* of these two quantities should be relatively near 1. On the other hand, if H_0 is false then MSTr ought to exceed MSE, so their ratio will tend to exceed 1. This suggests that a sensible test statistic for ANOVA is MSTr/MSE, but how large must this be to provide convincing evidence against H_0? Answering this question requires knowing the sampling distribution of this ratio.

   In Section 6.3 we introduced a family of probability distributions called *F* distributions, which became the basis for inference on the ratio of two variances in Section 10.5. If Y_1 and Y_2 are two independent chi-squared random variables with v_1 and v_2 df, respectively, then the ratio $F = (Y_1/v_1)/(Y_2/v_2)$ has an *F* distribution with v_1 numerator df and v_2 denominator df. Appendix Table A.8 gives *F* critical values for $\alpha = .10, .05, .01,$ and $.001$. Values of v_1 are identified with different columns of the table and the rows are labeled with various values of v_2. For example, the *F* critical value that captures upper-tail area .05 under the *F* curve with $v_1 = 4$ and $v_2 = 6$ is $F_{.05,4,6} = 4.53$, whereas $F_{.05,6,4} = 6.16$ (so don't accidentally switch numerator and denominator df!). The key theoretical result that justifies the ANOVA test procedure is that the test statistic MSTr/MSE has an *F* distribution when H_0 is true.

PROPOSITION When the ANOVA assumptions are satisfied and $H_0: \mu_1 = \cdots = \mu_I$ is true, the test statistic $F = \text{MSTr/MSE}$ has an F distribution with $I - 1$ numerator df and $IJ - I$ denominator df.

This theorem follows immediately from rewriting F as

$$F = \frac{\left[\dfrac{\text{SSTr}}{\sigma^2}\right]/(I-1)}{\left[\dfrac{\text{SSE}}{\sigma^2}\right]/(IJ-I)}$$

and then applying the definition of the F distribution along with properties of SSTr and SSE established earlier in this section. Here, finally, is the test procedure.

F TEST FOR SINGLE-FACTOR ANOVA

Null hypothesis: $H_0: \mu_1 = \cdots = \mu_I$

Alternative hypothesis: H_a: not all of the μ_i's are equal

Test statistic value: $f = \dfrac{\text{MSTr}}{\text{MSE}} = \dfrac{\text{SSTr}/(I-1)}{\text{SSE}/(IJ-I)}$

Rejection region for level α test: $f > F_{\alpha, I-1, IJ-I}$

P-value calculation: area under $F_{I-1, IJ-I}$ curve to the right of f

Refer to Section 10.5 to see how P-value information for F tests can be obtained from the table of F critical values. Alternatively, statistical software packages will automatically include the P-value with ANOVA output.

The computations building up to the test statistic value f are often summarized in a tabular format, called an **ANOVA table**, as displayed in Table 11.2. Tables produced by statistical software customarily include a P-value column to the right of the f column.

Table 11.2 An ANOVA table

Source of variation	df	Sum of squares	Mean square	f
Treatments	$I - 1$	SSTr	MSTr = SSTr/$(I - 1)$	MSTr/MSE
Error	$IJ - I$	SSE	MSE = SSE/$(IJ - I)$	
Total	$IJ - 1$	SST		

Example 11.3 With the ever-increasing power demand driven by everyone's electronic devices, engineers have begun exploring ways to tap into the energy discharge from household items—a hot stove pan, a candle flame, or even a hot soup bowl. The article "Low-Power Energy Harvesting of Thermoelectric Battery Charger with Step-Up DC–DC Converter: Applicable Case Study for Personal Electronic Gadgets" (*J. Energy Engr.* 2017) describes an experiment to compare the charging characteristics of five thermoelectric modules under certain conditions. Consider the accompanying data on the maximum power per unit area (mW/cm^2) for $J = 4$ replications of the experiment on each of the $I = 5$ modules.

Module	Max. power per cm^2				$\bar{x}_{i\cdot}$	s_i
1	98.8	93.1	96.6	91.8	95.08	3.21
2	82.5	87.5	88.8	91.3	87.53	3.70
3	77.7	74.7	76.2	78.8	76.85	1.79
4	82.6	80.5	82.8	84.1	82.50	1.49
5	91.9	87.5	86.9	90.0	89.08	2.31

Let μ_i denote the true mean max power per unit area when module i is used (i = 1, 2, 3, 4, 5). The null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ states that the true average is identical for the five modules. Let's carry out a test at significance level .01 to see whether H_0 should be rejected in favor of the assertion that true average max power per unit area is not the same for all modules. (At this point the plausibility of the normality and equal variance assumptions should be checked; we defer those tasks to later in this section.)

Since $I - 1 = 5 - 1 = 4$ and $IJ - I = 20 - 5 = 15$, the F critical value for the rejection region is $F_{.01,4,15} = 4.89$. The grand mean for the 20 observations is $\bar{x}_{\cdot\cdot} = 86.20\,\text{mW/cm}^2$. The treatment and error sums of squares are

$$\text{SSTr} = 4\left[(95.08 - 86.20)^2 + \cdots + (89.08 - 86.20)^2\right] = 759.6$$

$$\text{SSE} = (4 - 1)\left[(3.21)^2 + \cdots + (2.31)^2\right] = 104.2$$

The remaining computations are summarized in the accompanying ANOVA table. Because $f = 27.33 > F_{.01,4,15} = 4.89$, H_0 is rejected at significance level .01. The P-value is the area under the $F_{4,15}$ curve to the right of 27.33, which is 0 to several decimal places (and, in particular, far less than $\alpha = .01$). The modules clearly do *not* yield the same mean maximum power per cm^2 in size.

Source of variation	df	Sum of squares	Mean square	f	P-value
Treatments	4	759.6	189.90	27.33	<.0001
Error	15	104.2	6.95		
Total	19	863.8			

■

When the F test causes H_0 to be rejected as in Example 11.3, researchers will naturally be interested in further analysis to decide which μ_i's differ from which others. Methods for doing this are called *multiple comparison* procedures and are described in the next two sections.

Checking the ANOVA Assumptions

In previous chapters, a normal probability plot was suggested for checking normality. The individual sample sizes in ANOVA are typically too small for I separate plots to be informative. A single plot can be constructed by first subtracting $\bar{x}_{1\cdot}$ from each observation in the first sample, $\bar{x}_{2\cdot}$ from each observation in the second, and so on. These deviations are called **residuals** and are defined for single-factor ANOVA by

$$\text{residual} = e_{ij} = x_{ij} - \bar{x}_{i\cdot}$$

There are a total of IJ residuals, one for each observation. Table 11.3 shows the residuals for the 24 observations in Example 11.1. For instance, the first residual was computed as $e_{11} = x_{11} - \bar{x}_{1.} = 655.5 - 713.0 = -57.5$. Figure 11.2 displays a normal probability plot of these residuals. The straightness of the pattern gives strong support to the normality assumption. An analogous plot for the data of Example 11.3 conveys the same message.

Table 11.3 Residuals for the data in Example 11.1

Type of box	Residual		
1	−57.50	75.30	21.30
	8.40	−33.90	−13.60
2	32.27	15.57	29.97
	−70.83	−24.83	17.87
3	39.03	−59.07	−1.77
	−26.37	19.13	29.03
4	−26.92	66.68	−19.62
	−3.02	24.88	−42.02
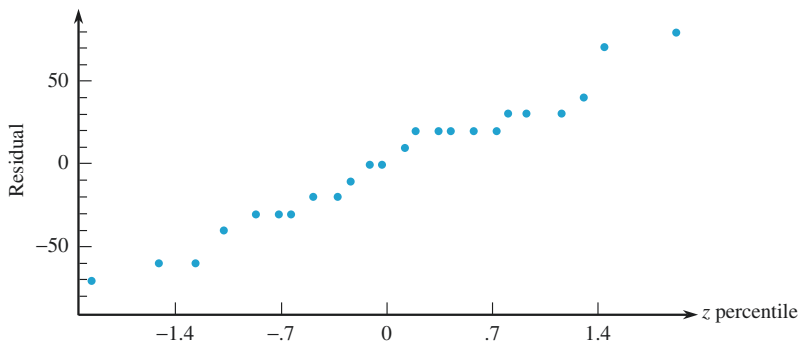
Figure 11.2 A normal probability plot based on the data of Example 11.1

The other ANOVA assumption is that the populations have equal variances. A popular informal rule is that if the largest sample standard deviation is not much more than twice the smallest one, it is permissible to assume equal variances. This is especially true for balanced or nearly-balanced study designs. In Example 11.1, the largest s is only about 1.25 times the smallest. Example 11.3 violates this informal rule slightly—the ratio of the largest s and smallest s is $3.70/1.49 = 2.48$—but, again, balance (i.e., equal sample sizes) makes this disparity somewhat less important.

Several formal tests of equal variance have been devised. If the likelihood ratio principle is applied to the problem of testing for equal variances for normal data, then the result is *Bartlett's test*. This is a generalization of the F test for equal variances given in Section 10.5, and it is very sensitive to the normality assumption. Since the ANOVA F test is robust in the presence of "mild" nonnormality (the significance level is approximately correct), it would be unfortunate to have the equal variances assumption invalidated not because they are different but because of such nonnormality. **Levene's test** is much less sensitive to the assumption of normality. Essentially, this test involves performing an

ANOVA on the absolute values of the residuals. That is, Levene's test performs an ANOVA F test using the absolute residuals $|e_{ij}|$ in place of x_{ij}. The idea is to use absolute residuals to compare the variability of the samples.

Example 11.4 To apply Levene's test to the data from Example 11.1, we first take the absolute values of the 24 residuals in Table 11.3. Then we apply ANOVA to the absolute residuals. With the aid of software,

$$\text{SSTr } = 115.3 \quad \text{MSTr } = 115.3/3 = 38.44 \quad f = 0.08$$
$$\text{SSE} = 9728.7 \quad \text{MSE} = 9728.7/20 = 486.44$$

Compare 0.08 to the critical value $F_{.10,3,20} = 2.38$. Because 0.08 is much smaller than 2.38, there is no reason to doubt that the population variances are equal.

We were somewhat more concerned about the power data from Example 11.3, since the sample standard deviations were rather different. Computing the residuals, taking their absolute values, and applying ANOVA to the results give

$$\text{SSTr } = 7.903 \quad \text{MSTr } = 7.903/4 = 1.976 \quad f = 1.17$$
$$\text{SSE} = 25.265 \quad \text{MSE} = 25.265/15 = 1.684$$

Because $f = 1.17 < F_{.10,4,15} = 2.36$, we do not reject a null hypothesis of equal population variances at the .10 level (in fact, the P-value is .362). There was no need to worry. ∎

Given that the absolute residuals are certainly not normally distributed, it might seem questionable to subject them to ANOVA. Fortunately, Levene's test works in spite of the normality assumption. A common sample size of 10 is sufficient for excellent accuracy in Levene's test, but smaller samples can still give useful results when only approximate P-values are needed (i.e., when the Levene's test P-value falls far above or far below the chosen significance level).

Some software packages perform Levene's test, but they will not necessarily get the same answer because they do not necessarily use absolute deviations from the mean. For example, Minitab uses absolute residuals with respect to the median, an especially good idea in case of skewed data. By default, SAS uses the squared deviations from the mean, although the absolute deviations from the mean can be requested. SAS also allows absolute deviations from the median (as the "BF" test, because Brown and Forsythe studied this procedure).

The ANOVA F test is robust not only to mild departures from normality but also to mild departures from equal variances. When the sample sizes are all the same, as we are assuming so far, the test is especially insensitive to unequal variances. Also, there is a generalization of the two-sample t test of Section 10.2 for more than two samples, and it does not demand equal variances. This test is available in JMP, R, and SAS.

If there is a major violation of assumptions, then the situation can sometimes be corrected by a data transformation, discussed in Section 11.3. Alternatively, the bootstrap can be used, by generalizing the method of Section 10.6 from two groups to several. There is also a nonparametric version of ANOVA (meaning no normality required) called the *Kruskal–Wallis test*, developed in Chapter 14.

Exercises: Section 11.1 (1–14)

1. An experiment to compare $I = 5$ brands of golf balls involved using a robotic driver to hit $J = 7$ balls of each brand. The resulting between-sample and within-sample estimates of σ^2 were MSTr = 123.50 and MSE = 22.16, respectively.

   a. State and test the relevant hypotheses using a significance level of .05.
   b. What can be said about the P-value of the test?

2. The lumen output was determined for each of $I = 3$ different brands of 60-watt soft-white lightbulbs, with $J = 8$ bulbs of each brand tested. The sums of squares were computed as SSE = 4773.3 and SSTr = 591.2. State the hypotheses of interest (including word definitions of parameters), and use the F test of ANOVA ($\alpha = .05$) to decide whether there are any differences in true average lumen outputs among the three brands for this type of bulb by obtaining as much information as possible about the P-value.

3. Freezing and thawing out food can adversely affects its texture. In one experiment described in the article "Effects of Freezing Treatments Before Convective Drying on Quality Parameters" (*J. of Food Engr.* 2019: 15–24), apple pieces were frozen using a –20 °C freezer (F20), a –80 °C freezer (F80), or liquid nitrogen (FLN). After being thawed out, texture tests were performed on each piece. The following information summarizes the elastic modulus (kPa) of the apple pieces. (Elastic modulus measures the apple pieces' resistance to deformation under load.)

Freezing method	J	$\bar{x}_{i\cdot}$	s_i
F20	8	61	10
F80	8	73	20
FLN	8	49	10

   Assuming conditions are met, use the ANOVA F test at level .05 to decide whether there are any differences between true mean elastic modulus of apple pieces using the three freezing methods.

4. The article "Load-Carrying Capacity of Lengthwise Cracked Wood Beams Retrofitted by Self-Tapping Screws" (*J. Struct. Engr.* 2017) provides data on the maximum load (kN) of $J = 5$ specimens of $I = 4$ types of wood beams used in housing: intact beams (L1), long, centered cracks (L2), short, centered cracks (L3), and long, off-center cracks (L4).

Beam type	Maximum load (kN)				
L1	32.53	19.18	7.50	18.18	21.89
L2	13.15	10.62	6.75	16.08	14.12
L3	13.25	18.52	12.02	18.83	12.80
L4	26.95	13.19	11.55	24.63	23.63

   Use a significance level of .05 to test the null hypothesis of no difference in true average maximum load for these four beam types.

5. The article "Differences in Impact Performance of Bicycle Helmets During Oblique Impacts" (*J. Biomech. Engr.* 2018) describes an experiment in which 10 different bicycle helmet brands (4 of each brand, for 40 total helmets) were strapped onto a mannequin head and subjected to a frontal impact at 6.6m/s. At that speed, concussion without a helmet is extremely likely. The peak linear acceleration (PLA, in g) was measured in each test; values over 300 g are associated with a high risk of brain injury.

   a. The sample mean PLAs for the 10 helmet brands are presented below; labels are abbreviations used in the article for the brand names.

BMIPS	BSF	BSP	CW	GMIPS
141	144	122	127	148

GS	NW	SOO	ST	SWE
175	147	117	142	129

   Use these sample means to determine SSTr and MSTr.
   b. The value SSE = 3900 is consistent with information provided in the article.

Construct an ANOVA table, and carry out a hypothesis test at the .01 significance level.

6. In an experiment to investigate the performance of four different brands of spark plugs intended for use on a 125-cc two-stroke motorcycle, five plugs of each brand were tested for the number of miles (at a constant speed) until failure. The partial ANOVA table for the data is given here. Fill in the missing entries, state the relevant hypotheses, and carry out a test by obtaining as much information as you can about the P-value.

Source	df	Sum of squares	Mean square	f
Brand				
Error			14,713.69	
Total		310,500.76		

7. Consider the box compression data presented in Example 11.1. Carry out an analysis of variance F test at significance level .01, and summarize the results in an ANOVA table.

8. Plastic waste, particularly microplastics in oceans and waterways, has become an increasing global environmental concern. The article "Environmentally Relevant Concentrations of Microplastic Particles Influence Larval Fish Ecology" (*Science*, 3 June 2016: 1213–1216) describes an experiment in which fertilized egg strands of European perch were placed in 15 identical tanks. Tanks were then randomly assigned (1) no microplastics, (2) a "typical" microplastic concentration (10,000 particles/m^3), or (3) a high concentration (80,000 particles/m^3). After a three-week period, the successful hatching rates were recorded for every tank; the data appears below.

Microplastic level	Hatching success rate				
None	95	98	96	92	97
Typical	86	93	88	87	91
High	85	74	86	77	83

Does the data provide convincing statistical evidence that microplastic level has an effect on the success rate of eggs hatching for European perch? Test at the .01 significance level.

9. One popular approach to lower back pain is to apply a piece of durable tape along the lower spine. In one study, 108 women with lower back pain were randomly assigned to receive one of four treatments: Kinesio tape applied to a tense back (KTT), Kinesio tape without any back tension (KTNT), Micropore tape (MP), and no tape (CG, a control group). After 10 days of treatment, each woman's lower back extension (degrees) was measured. The accompanying table summarizes the results.

Treatment	J	$\bar{x}_{i.}$	s_i
KTT	27	30°	14°
KTNT	27	29°	15°
MP	27	26°	13°
CG	27	27°	9°

("Kinesio Taping Reduces Pain and Improves Disability in Low Back Pain Patients," *Physiotherapy* 2019: 65–75.)

a. Calculate SSTr, MSTr, SSE, and MSE.
b. Test the null hypothesis that true mean back extension after 10 days is the same for all four treatments, at the .05 level.

10. Does music affect memorization skills, and does it matter if the music includes vocals/lyrics? In one experiment (P. Ramos, "The Impact of Music on Short Term Memory and Cognitive Processes," Univ. of Bridgeport, 2017), subjects were randomly assigned to one of four environments: (A) an instrumental-only version of Adele's "Million Years Ago" playing in background (B) a vocals-only version, (C) a version with both instruments and vocals, and (D) a control group with no ambient music. All subjects were given a list of everyday words to memorize in 90 seconds and then asked to write down as many words as they could remember. Information consistent with that report appears below.

Condition		Number of words remembered	
	J	$\bar{x}_{i\cdot}$	s_i
Instrumental only	26	10.2	2.8
Vocals only	26	7.7	2.1
Instruments & vocals	26	9.0	2.5
Control (no music)	26	10.7	3.0

Test whether music environment affects memorization skills, as measured by population mean number of words remembered using this activity, at the .01 significance level.

11. The article referenced in Example 11.3 also looked at the time (min) to charge a 4.2-V battery.

Module	Time to full charge (min)			
1	200	199	204	208
2	233	229	226	224
3	140	146	146	136
4	169	174	171	166
5	205	212	214	208

a. Check the ANOVA assumptions with a normal plot and a test for equal variances.
b. Does mean charge time differ across these five thermoelectric modules? State and test the relevant hypotheses using $\alpha = .01$.

12. Six samples of each of four types of cereal grain grown in a certain region were analyzed to determine thiamin content, resulting in the following data (μg/g):

Grain	Thiamin content					
Wheat	5.2	4.5	6.0	6.1	6.7	5.8
Barley	6.5	8.0	6.1	7.5	5.9	5.6
Maize	5.8	4.7	6.4	4.9	6.0	5.2
Oats	8.3	6.1	7.8	7.0	5.5	7.2

a. Check the ANOVA assumptions.
b. Test to see if at least two of the grains differ with respect to true average thiamin content. Use an $\alpha = .05$ test based on the P-value method.

13. Derive the fundamental identity SST = SSTr + SSE by squaring both sides of Equation (11.1) and summing over all i and j. [*Hint*: For any particular i, $\sum_j (x_{ij} - \bar{x}_{i\cdot}) = 0$.]

14. In single-factor ANOVA with I treatments and J observations per treatment, let $\mu = (1/I) \sum \mu_i$.

a. Express $E(\overline{X}_{\cdot\cdot})$ in terms of μ. [*Hint*: $\overline{X}_{\cdot\cdot} = (1/I) \sum \overline{X}_{i\cdot}$.]
b. Express $E(\overline{X}_{i\cdot}^2)$ in terms of σ and the μ_i's. [*Hint*: For any rv Y, $E(Y^2) = V(Y) + [E(Y)]^2$.]
c. Express $E(\overline{X}_{\cdot\cdot}^2)$ in terms of μ and σ.
d. Express $E(\text{SSTr})$ in terms of μ, σ, and the μ_i's. Then show that

$$E(\text{MSTr}) = \sigma^2 + \frac{J}{I-1} \sum (\mu_i - \mu)^2$$

e. Using the result of part (d), what is $E(\text{MSTr})$ when H_0 is true? When H_0 is false, how does $E(\text{MSTr})$ compare to σ^2?

11.2 Multiple Comparisons in ANOVA

When the computed value of the F statistic in single-factor ANOVA is not significant, the analysis is terminated because no differences among the μ_i's have been identified. But when H_0 is rejected, the investigator will usually want to know *which* of the μ_i's are different from each other. A method for carrying out this further analysis is called a **multiple comparisons procedure**.

Several of the most frequently used such procedures are based on the following central idea. First calculate a confidence interval for each pairwise difference $\mu_i - \mu_j$ with $i < j$. Thus if $I = 4$, the six required CIs would be for $\mu_1 - \mu_2$ (but not also for $\mu_2 - \mu_1$), $\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_2 - \mu_3$, $\mu_2 - \mu_4$, and $\mu_3 - \mu_4$. Then if the interval for $\mu_1 - \mu_2$ does *not* include 0, conclude that μ_1 and μ_2 differ significantly from each other; if the interval includes 0, the two μ_i's are judged not significantly different. Following the same line of reasoning for each of the other intervals, we end up being able to judge for each pair of μ_i's whether or not they differ significantly from each other.

The procedures based on this idea differ in the method used to calculate the various CIs. Here we present a popular method that controls the *simultaneous* confidence level for all $\binom{I}{2} = I(I-1)/2$ intervals calculated.

Tukey's Procedure
Tukey's procedure involves the use of another probability distribution.

DEFINITION   Let Z_1, Z_2,..., Z_m be m independent standard normal rvs, and let W be a χ_v^2 rv independent of the Z_i's. Then the distribution of

$$Q = \frac{\max|Z_i - Z_j|}{\sqrt{W/v}} = \frac{\max(Z_1, \ldots, Z_m) - \min(Z_1, \ldots, Z_m)}{\sqrt{W/v}}$$

is called the **studentized range distribution**. The distribution has two parameters: $m =$ the number of Z_i's and $v =$ denominator df. We denote the critical value that captures upper-tail area α under the density curve of Q by $Q_{\alpha,m,v}$. A tabulation of these critical values appears in Appendix Table A.9.

The word "range" reflects the fact that the numerator of Q is indeed the range of the Z_i's. Dividing the range by $\sqrt{W/v}$ is the same as dividing each individual Z_i by $\sqrt{W/v}$. But $Z_i/\sqrt{W/v}$ has a (Student) t distribution[1]; "studentizing" refers to the division by $\sqrt{W/v}$. So Q is actually the range of m variables that have the t distribution (but they are not independent because the denominator is the same for each one).

The identification of the quantities in the definition of Q with single-factor ANOVA is as follows:

$$Z_i = \frac{\overline{X}_{i\cdot} - \mu_i}{\sigma/\sqrt{J}} \quad m = I \quad W = \frac{\text{SSE}}{\sigma^2} = \frac{(IJ - I)\text{MSE}}{\sigma^2} \quad v = IJ - I$$

Substituting into Q gives

$$Q = \frac{\max\left|\dfrac{\overline{X}_{i\cdot} - \mu_i}{\sigma/\sqrt{J}} - \dfrac{\overline{X}_{j\cdot} - \mu_j}{\sigma/\sqrt{J}}\right|}{\sqrt{\dfrac{(IJ - I)\text{MSE}}{\sigma^2}/(IJ - I)}} = \frac{\max\left|\overline{X}_{i\cdot} - \overline{X}_{j\cdot} - \left(\mu_i - \mu_j\right)\right|}{\sqrt{\text{MSE}/J}}$$

---

[1]"Student" was the pseudonym used by the statistician Gossett, who derived the t distribution but published his work using the pseudonym "Student" because his employer, the Guinness Brewing Co., would not permit publication under his own name.

In this latter expression for Q, the denominator $\sqrt{\text{MSE}/J}$ is the estimated standard deviation of $\overline{X}_{i\cdot} - \mu_i$. By definition of Q and Q_α, $P(Q \leq Q_\alpha) = 1 - \alpha$, so

$$
\begin{aligned}
1 - \alpha &= P\left(\frac{\max|\overline{X}_{i\cdot} - \overline{X}_{j\cdot} - (\mu_i - \mu_j)|}{\sqrt{\text{MSE}/J}} \leq Q_{\alpha,I,I(J-1)}\right) \\
&= P\left(\frac{|\overline{X}_{i\cdot} - \overline{X}_{j\cdot} - (\mu_i - \mu_j)|}{\sqrt{\text{MSE}/J}} \leq Q_{\alpha,I,I(J-1)} \text{ for all } i,j\right) \\
&= P\left(-Q_\alpha\sqrt{\text{MSE}/J} \leq \overline{X}_{i\cdot} - \overline{X}_{j\cdot} - (\mu_i - \mu_j) \leq Q_\alpha\sqrt{\text{MSE}/J} \text{ for all } i,j\right) \\
&= P\left(\overline{X}_{i\cdot} - \overline{X}_{j\cdot} - Q_\alpha\sqrt{\text{MSE}/J} \leq \mu_i - \mu_j \leq \overline{X}_{i\cdot} - \overline{X}_{j\cdot} + Q_\alpha\sqrt{\text{MSE}/J} \text{ for all } i,j\right)
\end{aligned}
$$

(whew!). Replacing $\overline{X}_{i\cdot}, \overline{X}_{j\cdot}$, and MSE by the values calculated from the data gives the following result.

PROPOSITION   For each $i < j$, form the interval

$$
\bar{x}_{i\cdot} - \bar{x}_{j\cdot} \pm Q_{\alpha,I,IJ-I}\sqrt{\text{MSE}/J} \tag{11.2}
$$

There are $I(I-1)/2$ such intervals: one for $\mu_1 - \mu_2$, another for $\mu_1 - \mu_3$, …, and the last for $\mu_{I-1} - \mu_I$. Then the *simultaneous* confidence level that *every* interval includes the corresponding value of $\mu_i - \mu_j$ is $100(1 - \alpha)\%$.

Notice that the second subscript on Q_α is I, whereas the second subscript on F_α used in the ANOVA F test is $I - 1$. $Q_{\alpha,I,IJ-I}$ can be obtained in R with the command `qtukey(1 - α, I, IJ − I)`.

We will say more about the interpretation of "simultaneous" shortly. Each interval that doesn't include 0 yields the conclusion that the corresponding values of μ_i and μ_j are different—we say that μ_i and μ_j "differ significantly" from each other.

Example 11.5 An experiment was carried out to compare five different brands of automobile oil filters with respect to their ability to capture foreign material. Let μ_i denote the true average amount of material captured by brand i filters ($i = 1, \ldots, 5$) under controlled conditions. A sample of $J = 9$ filters of each brand was used, resulting in the following sample mean amounts: $\bar{x}_{1\cdot} = 14.5$, $\bar{x}_{2\cdot} = 13.8$, $\bar{x}_{3\cdot} = 13.3$, $\bar{x}_{4\cdot} = 14.3$, and $\bar{x}_{5\cdot} = 13.1$. We will assume for this example that the conditions for inference (approximate normality and equal variance) are met by the data set. Table 11.4 is the ANOVA table summarizing the first part of the analysis.

Table 11.4 ANOVA table for Example 11.5

Source of variation	df	Sum of squares	Mean square	f
Treatments (brands)	4	13.32	3.33	37.84
Error	40	3.53	.088	
Total	44	16.85		

Since $f = 37.84 > F_{.001,4,40} = 5.70$, H_0 is decisively rejected.

We now use Tukey's procedure to look for significant differences among the μ_i's. From Appendix Table A.9, $Q_{.05,5,40} = 4.04$ (the second subscript on Q is 5 and not $5 - 1$ as in F). Applying Equation (11.2), the CI for each $\mu_i - \mu_j$ is

$$(\bar{x}_{i\cdot} - \bar{x}_{j\cdot}) \pm 4.04\sqrt{.088/9} = (\bar{x}_{i\cdot} - \bar{x}_{j\cdot}) \pm 0.4$$

Due to the balanced design, the margin of error is the same on all $\binom{5}{2} = 10$ Tukey CIs; if the sample sizes differed, this would not be the case (see Section 11.3). The resulting CIs are displayed in Table 11.5; those marked with an asterisk do *not* include zero:

Table 11.5 Tukey's simultaneous confidence intervals for Example 11.4

i	j	CI for $\mu_i - \mu_j$	i	j	CI for $\mu_i - \mu_j$
1	2	(0.3, 1.1)*	2	4	(−0.9, −0.1)*
1	3	(0.8, 1.6)*	2	5	(0.3, 1.1)*
1	4	(−0.2, 0.6)	3	4	(−1.4, −0.6)*
1	5	(1.0, 1.8)*	3	5	(−0.2, 0.6)
2	3	(0.1, 0.9)*	4	5	(0.8, 1.6)*

Thus brands 1 and 4 are not significantly different from one another, but they are significantly higher than the other three brands in their true average contents. Brand 2 is significantly better than 3 and 5 but worse than 1 and 4, and brands 3 and 5 do not differ significantly. ∎

While the CIs in Table 11.5 correctly indicate which means are believed to be significantly different, this display is rather unwieldy. It is preferable to list out the observed sample means (say, from smallest to largest) and somehow indicate which are "close enough" that the corresponding population means are not judged to be significantly different. The following box describes how nonsignificant differences can be identified visually using an "underscoring pattern."

Tukey's Method for Identifying Significantly Different μ_i's

1. List the sample means in increasing order (make sure to identify the corresponding population/treatment for each $\bar{x}_{i\cdot}$).
2. Starting at the far left, use the Tukey intervals to determine which means are *not* significantly different from the first one in the list. Underscore that set of means with a single line segment.
3. Continue in this fashion for the second mean, third mean, etc., always underscoring in the rightward direction. Duplicate underscorings should only be drawn once.

Any pair of sample means *not* underscored by the same line correspond to a pair of population or treatment means that are judged significantly different.

In fact, it is not necessary to construct the Tukey intervals in order to perform step 2. Rather, the CI for $\mu_i - \mu_j$ will include 0 if and only if $\bar{x}_{i\cdot}$ and $\bar{x}_{j\cdot}$ differ by less than $Q_{\alpha, I, I(J-1)}\sqrt{\text{MSE}/J}$, the margin of error of the confidence interval (11.2). This margin of error is sometimes referred to as Tukey's **honestly significant difference (HSD)**.

As an example, consider $I = 5$ with

$$\bar{x}_{2\cdot} < \bar{x}_{5\cdot} < \bar{x}_{4\cdot} < \bar{x}_{1\cdot} < \bar{x}_{3\cdot}$$

Suppose the Tukey confidence intervals for $\mu_2 - \mu_5$ and $\mu_2 - \mu_4$ include zero, but those for $\mu_2 - \mu_1$ and $\mu_2 - \mu_3$ do not. Then we draw a line segment starting from 2 and extending to 4:

Group: 2   5   4   1   3
Sample mean: $\underline{\bar{x}_{2\cdot} \quad \bar{x}_{5\cdot} \quad \bar{x}_{4\cdot}} \quad \bar{x}_{1\cdot} \quad \bar{x}_{3\cdot}$

Next, suppose the mean for group 5 is not significantly different from that of group 4. Since we have already accounted for that set, no duplicate line is drawn. Finally, if groups 4 and 1 are not significantly different, that pair is underscored:

Group: 2   5   4   1   3
Sample mean: $\bar{x}_{2\cdot} \quad \bar{x}_{5\cdot} \quad \underline{\bar{x}_{4\cdot} \quad \bar{x}_{1\cdot}} \quad \bar{x}_{3\cdot}$

The fact that $\bar{x}_{3\cdot}$ isn't underlined at all indicates that μ_3 is statistically significantly different from *all* other group means.

Example 11.6 (Example 11.5 continued) The five sample means, arranged in order, are

Brand of filter: 5   3   2   4   1
Sample mean: 13.1   13.3   13.8   14.3   14.5

Only two of the Tukey CIs included zero: the interval for $\mu_1 - \mu_4$ and that for $\mu_3 - \mu_5$. Equivalently, only those two pairs of means differ by less than the honestly significant difference $4.04\sqrt{.088/9} = .4$ The resulting underscoring pattern is

Brand of filter: 5   3   2   4   1
Sample mean: $\underline{13.1 \quad 13.3}$ 13.8 $\underline{14.3 \quad 14.5}$

The mean for brand 2 is not underlined at all, since μ_2 was judged to be significantly different from all other means.

   If $\bar{x}_{2\cdot} = 13.6$ rather than 13.8 with HSD = .4, the underscoring configuration would be

Brand of filter: 5   3   2   4   1
Sample mean: 13.1 $\underline{13.3 \quad 13.6}$ $\underline{14.3 \quad 14.5}$

The interpretation of this underscoring must be done with care, since we seem to have concluded that brands 5 and 3 do not differ significantly, 3 and 2 also do not, yet 5 and 2 do differ. One could say here that although evidence allows us to conclude that brands 5 and 2 differ from each other, neither has been shown to be significantly different from brand 3. ∎

Example 11.7 Almost anyone who attends physical therapy receives transcutaneous electrical nerve stimulation (TENS), possibly combined with a cold compress, to address pain. The article "Effect of Burst TENS and Conventional TENS Combined with Cryotherapy on Pressure Pain Threshold" (*Physiotherapy* 2015: 155–160) summarizes an experiment in which 112 healthy women were each

randomly assigned to one of seven treatments listed in the accompanying table (so, $J = 16$ for each treatment). After the treatment, researchers measured each woman's pain threshold, in kg of force applied to the top of the humerus, resulting in the accompanying data.

	Pain threshold (kg of force)	
Treatment	$\bar{x}_{i\cdot}$	s_i
(1) Control	2.8	0.7
(2) Placebo TENS	2.3	0.9
(3) Conventional TENS	3.2	1.0
(4) Burst TENS	4.3	0.9
(5) Cryotherapy	4.4	0.9
(6) Cryotherapy + burst TENS	5.7	0.8
(7) Cryotherapy + conventional TENS	3.0	0.7

Let μ_i = the true mean pain threshold after the ith treatment ($i = 1, \ldots, 7$). We wish to test the null hypothesis H_0: $\mu_1 = \cdots = \mu_7$ against the alternative that not all μ_i's are equal. From the sample means and standard deviations, SSTr = 133.67 and SSE = 75.75, giving the ANOVA table in Table 11.6.

Since $f = 30.88 > F_{.01,6,105} = 2.98$, H_0 is rejected at the .01 level; in fact, P-value ≈ 0. We conclude that the true mean pain threshold differs across the seven treatments.

Table 11.6 ANOVA table for Example 11.7

Source of variation	df	Sum of squares	Mean square	f
Treatments	6	133.67	22.28	30.88
Error	105	75.75	0.72	
Total	111	209.42		

Next, we apply Tukey's method. There are $I = 7$ treatments and 105 df for error, so $Q_{.01,7,105} \approx 5.02$ (interpolating from Table A.9) and Tukey's HSD $= 5.02\sqrt{0.72/16} = 1.06$. Ordering the means and underscoring yields

$$
\begin{array}{ccccccc}
(2) & (1) & (7) & (3) & (4) & (5) & (6) \\
\underline{2.3} & \underline{2.8} & \underline{3.0} & \underline{3.2} & \underline{4.3} & \underline{4.4} & 5.7
\end{array}
$$

Higher pain thresholds are considered better. In that respect, treatment 6 (cryotherapy plus burst TENS) is the clear winner, since the mean pain threshold under that treatment is highest and significantly different than all others. Treatments 4 and 5 (just burst TENS or just cryotherapy) are next-best and are not significantly different from each other. Finally, the other four treatments are not honestly significantly different—and, in particular, treatments 3 and 7 are comparable to the control and placebo groups.

Many research journals and some statistical software packages express the underscoring scheme with letter groupings. Figure 11.3 (p. 659) shows Minitab output from this analysis. The A, B, C letter groupings correspond to the three nonoverlapping sets identified above: treatment 6, treatments 4 and 5, and the rest.

Grouping Information Using the Tukey Method and 99% Confidence

Treatment	N	Mean	Grouping		
6	16	5.700	A		
5	16	4.400		B	
4	16	4.300		B	
3	16	3.200			C
7	16	3.000			C
1	16	2.800			C
2	16	2.300			C

Means that do not share a letter are significantly different.

Figure 11.3 Tukey's method using Minitab   ∎

The Interpretation of α in Tukey's Procedure

We stated previously that the *simultaneous* confidence level is controlled by Tukey's method. So what does "simultaneous" mean here? Consider calculating a 95% CI for a population mean μ based on a sample from that population and then a 95% CI for a population proportion p based on another sample selected independently of the first one. Prior to obtaining data, the probability that the first interval will include μ is .95, and this is also the probability that the second interval will include p. Because the two samples are selected independently of each other, the probability that *both* intervals will include the values of the respective parameters is $(.95)(.95) = (.95)^2 = .9025$. Thus the *simultaneous* or *joint* confidence level for the two intervals is roughly 90%—if pairs of intervals are calculated over and over again from independent samples, in the long run 90.25% of the time the first interval will capture μ *and* the second will include p. Similarly, if three CIs are calculated based on independent samples, the simultaneous confidence level will be $100(.95)^3\% \approx 86\%$. Clearly, as the number of intervals increases, the simultaneous confidence level that all intervals capture their respective parameters will decrease.

Now suppose that we want to maintain the simultaneous confidence level at 95%. Then for two independent samples, the individual confidence level for each would have to be $100\sqrt{.95}\% \approx 97.5\%$. The larger the number of intervals, the higher the individual confidence level would have to be to maintain the 95% simultaneous level.

The tricky thing about the Tukey intervals is that they are not based on independent samples— MSE appears in every one, and various intervals share the same $\bar{x}_{i\cdot}$'s (e.g., in the case $I = 4$, three different intervals all use $\bar{x}_{1\cdot}$). This implies that there is no straightforward probability argument for ascertaining the simultaneous confidence level from the individual confidence levels. Nevertheless, if $Q_{.05}$ is used, the simultaneous confidence level is controlled at 95%, whereas using $Q_{.01}$ gives a simultaneous 99% level. To obtain a 95% simultaneous level, the individual level for each interval must be considerably larger than 95%. Said in a slightly different way, to obtain a 5% *experimentwise* or *family* error rate, the individual or per-comparison error rate for each interval must be considerably smaller than .05.

Confidence Intervals for Other Parametric Functions

In some situations, a CI is desired for a function of the μ_i's more complicated than a difference $\mu_i - \mu_j$. Let $\theta = \sum c_i \mu_i$, where the c_i's are constants. One such function is $\frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{3}(\mu_3 + \mu_4 + \mu_5)$, which in the context of Example 11.5 measures the difference between the group consisting of the first two brands and that of the last three brands. Because the X_{ij}'s are normally

distributed with $E(X_{ij}) = \mu_i$ and $V(X_{ij}) = \sigma^2$, the natural estimator $\hat{\theta} = \sum_i c_i \overline{X}_{i\cdot}$ is normally distributed, unbiased for θ, and

$$V(\hat{\theta}) = V\left(\sum_i c_i \overline{X}_{i\cdot}\right) = \sum_i c_i^2 V(\overline{X}_{i\cdot}) = \frac{\sigma^2}{J} \sum_i c_i^2$$

Estimating σ^2 by MSE and forming $\hat{\sigma}_{\hat{\theta}}$ results in a t variable $(\hat{\theta} - \theta)/\hat{\sigma}_{\hat{\theta}}$, which can be manipulated to obtain the following $100(1 - \alpha)\%$ confidence interval for $\sum c_i \mu_i$:

$$\sum c_i \bar{x}_{i\cdot} \pm t_{\alpha/2, I(J-1)} \sqrt{\text{MSE} \cdot \sum c_i^2 / J} \tag{11.3}$$

Example 11.8 (Example 11.5 continued) The parametric function for comparing the first two (store) brands of oil filter with the last three (national) brands is $\theta = \frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{3}(\mu_3 + \mu_4 + \mu_5)$, from which

$$\sum c_i^2 = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 = \frac{5}{6}$$

With $\hat{\theta} = \frac{1}{2}(\bar{x}_{1\cdot} + \bar{x}_{2\cdot}) - \frac{1}{3}(\bar{x}_{3\cdot} + \bar{x}_{4\cdot} + \bar{x}_{5\cdot}) = .583$ and MSE = .088, a 95% CI for θ is

$$.583 \pm 2.021 \sqrt{(.088) \cdot (5/6)/9} = .583 \pm .182 = (.401, .765) \qquad \blacksquare$$

Notice that in the foregoing example the coefficients $c_1, \ldots, c_5$ satisfy $\sum c_i = \frac{1}{2} + \frac{1}{2} - \frac{1}{3} - \frac{1}{3} - \frac{1}{3} = 0$. When the coefficients sum to 0, the linear combination $\theta = \sum c_i \mu_i$ is called a **contrast** among the means, and the analysis is available in a number of statistical software programs.

Sometimes an experiment is carried out to compare each of several "new" treatments to a control treatment. In such situations, a multiple comparisons technique called *Dunnett's method* is appropriate.

Exercises: Section 11.2 (15–26)

15. An experiment to compare the spreading rates of five different brands of yellow interior latex paint available in a particular area used 4 gallons ($J = 4$) of each paint. The sample average spreading rates (ft^2/gal) for the five brands were $\bar{x}_{1\cdot} = 462.0$, $\bar{x}_{2\cdot} = 512.8$, $\bar{x}_{3\cdot} = 437.5$, $\bar{x}_{4\cdot} = 469.3$, and $\bar{x}_{5\cdot} = 532.1$. The computed value of F was found to be significant at level $\alpha = .05$. With MSE = 272.8, use Tukey's procedure to investigate significant differences in the true average spreading rates between brands.

16. In the previous exercise, suppose $\bar{x}_{3\cdot} = 427.5$. Now which true average spreading rates differ significantly from each

other? Be sure to use the method of underscoring to illustrate your conclusions, and write a paragraph summarizing your results.

17. Repeat the previous exercise supposing that $\bar{x}_{2\cdot} = 502.8$ in addition to $\bar{x}_{3\cdot} = 427.5$.

18. Consider the data on maximum load for cracked wood beams presented in Exercise 4. Would it make sense to apply Tukey's method to this data? Why or why not? [*Hint:* The *P*-value from the analysis of variance is .169.]

19. Use Tukey's procedure on the data of Example 11.1 to identify differences in true average compression strength among the

four box types. Is your answer consistent with the boxplot in Figure 11.1a?

20. Use Tukey's procedure on the data of Example 11.3 to identify differences in true mean maximum power per unit area among the five modules.

21. Of the five modules in Example 11.3, the first two are existing commercial devices and the last three are prototypes constructed by the study's authors. Compute a 95% t CI for the contrast $\theta = \frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{3}(\mu_3 + \mu_4 + \mu_5)$.

22. The article "Iron and Manganese Present in Underground Water Promote Biochemical, Genotoxic, and Behavioral Alterations in Zebrafish" (*Environ. Sci. Pollut. Res.* 2019: 23555–23570) reports the following data on micronucleus frequency in the muscle tissues for zebrafish exposed to varying concentrations of iron and manganese. Micronuclei are a warning sign of possible DNA damage. There were $J = 10$ zebrafish in each group.

Treatment	$\bar{x}_{i \cdot}$	s_i
1. Control	53.72	4.30
2. Fe 0.8	139.11	6.41
3. Fe 1.3	93.62	4.49
4. Mn 0.2	134.66	13.20
5. Mn 0.4	141.12	8.32
6. Fe 0.8/Mn 0.2	101.25	15.41
7. Fe 1.3/Mn 0.4	124.50	9.60

a. Perform an analysis of variance at the .01 significance level. [*Note:* Though unequal variances are a concern here, the balanced study design should at least partially mitigate that issue.]

b. Apply Tukey's method to determine which treatments result in significantly different mean micronucleus frequencies.

c. Suppose that $100(1 - \alpha)\%$ CIs for k different parametric functions are computed from the same ANOVA data set. Then it is easily verified that the simultaneous confidence level is at least $100(1 - k\alpha)\%$. Compute CIs with simultaneous confidence level at least 98% for the contrasts $\mu_1 - \frac{1}{6}(\mu_2 + \mu_3 + \mu_4 + \mu_5 + \mu_6 + \mu_7)$ and $\frac{1}{4}(\mu_2 + \mu_3 + \mu_6 + \mu_7) - \frac{1}{2}(\mu_4 + \mu_5)$.

23. Refer back to the bike helmet data of Exercise 5.

a. Apply Tukey's procedure at the .05 level to determine which bike helmets have honestly significantly different mean peak linear acceleration under the specified experimental conditions. Use SSE = 3900.

b. Seven of the 10 brands are considered road helmets (elongated shape with aerodynamic venting), while three brands—BMIPS (1), GMIPS (5), and NW (7)—are nonroad helmets. Compute a 95% CI for the contrast $\theta = \frac{1}{7}(\mu_2 + \cdots + \mu_{10}) - \frac{1}{3}(\mu_1 + \mu_5 + \mu_7)$, where the first sum spans across the seven road helmet brands.

24. Consider the accompanying data on plant growth after the application of different types of growth hormone.

Hormone	Growth			
1	13	17	7	14
2	21	13	20	17
3	18	15	20	17
4	7	11	18	10
5	6	11	15	8

a. Perform an F test at level $\alpha = .05$.

b. What happens when Tukey's procedure is applied?

25. Consider a single-factor ANOVA experiment in which $I = 3$, $J = 5$, $\bar{x}_{1 \cdot} = 10$, $\bar{x}_{2 \cdot} = 12$, and $\bar{x}_{3 \cdot} = 20$. Determine a value of SSE for which $f > F_{.05, 2, 12}$, so that $H_0: \mu_1 = \mu_2 = \mu_3$ is rejected, yet when Tukey's procedure is applied none of the μ_i's differ significantly from each other.

26. Refer to the previous exercise and suppose $\bar{x}_{1 \cdot} = 10$, $\bar{x}_{2 \cdot} = 15$, and $\bar{x}_{3 \cdot} = 20$. Can you now find a value of SSE that produces such a contradiction between the F test and Tukey's procedure?

11.3   More on Single-Factor ANOVA

In this section, we consider some additional issues relating to single-factor ANOVA. These include an alternative description of the model parameters, power and β for the F test, the relationship of the test to procedures previously considered, data transformation, a random effects model, and formulas for the case of unequal sample sizes.

An Alternative Description of the ANOVA Model

The assumptions of single-factor ANOVA can be described succinctly through the **model equation**

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

where ε_{ij} represents a random deviation from the population or true treatment mean μ_i. The ε_{ij}'s are assumed to be independent, normally distributed rvs (implying that the X_{ij}'s are also) with $E(\varepsilon_{ij}) = 0$ [so that $E(X_{ij}) = \mu_i$] and $V(\varepsilon_{ij}) = \sigma^2$ [from which $V(X_{ij}) = \sigma^2$ for every i and j]. An alternative description of single-factor ANOVA will give added insight and suggest appropriate generalizations to models involving more than one factor. Define a parameter μ by

$$\mu = \frac{1}{I}\sum_{i=1}^{I}\mu_i$$

and parameters $\alpha_1, \ldots, \alpha_I$ by

$$\alpha_i = \mu_i - \mu \quad (i = 1,\ldots,I)$$

Then the treatment mean μ_i can be written as $\mu + \alpha_i$, where μ represents the true average overall response across all populations/treatments, and α_i is the **effect**, measured as a departure from μ, due to the ith treatment. Whereas we initially had I parameters, we now have $I + 1$: $\mu, \alpha_1, \ldots, \alpha_I$. However, because $\sum \alpha_i = 0$ (the average departure from the overall mean response is zero), only I of these new parameters are independently determined, so there are as many independent parameters as there were before. In terms of μ and the α_i's, the model equation becomes

$$X_{ij} = \mu_i + \alpha_i + \varepsilon_{ij} \quad (i = 1,\ldots,I; j = 1,\ldots,J)$$

In the next two sections, we will develop analogous models for two-factor ANOVA. The claim that the μ_i's are identical is equivalent to the equality of the α_i's, and because $\sum \alpha_i = 0$, the null hypothesis becomes

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$

In Section 11.1, it was stated that MSTr is an unbiased estimator of σ^2 when H_0 is true but otherwise tends to overestimate σ^2. More precisely, Exercise 14 established that

$$E(\text{MSTr}) = \sigma^2 + \frac{J}{I-1}\sum (\mu_i - \mu)^2 = \sigma^2 + \frac{J}{I-1}\sum \alpha_i^2$$

When H_0 is true, $\sum \alpha_i^2 = 0$ so $E(\text{MSTr}) = \sigma^2$ (MSE is unbiased whether or not H_0 is true). If $\sum \alpha_i^2$ is used as a measure of the extent to which H_0 is false, then a larger value of $\sum \alpha_i^2$ will result in a greater

tendency for MSTr to overestimate σ^2. More generally, formulas for expected mean squares for multifactor models are used to suggest how to form F ratios to test various hypotheses.

Power and β for the F Test

Consider a set of parameter values α_1, α_2, ..., α_I for which H_0 is not true. The power of the ANOVA F test is the probability that H_0 is rejected when that set is the set of true values, and the probability of a type II error is $\beta = 1 - \text{power}$. One might think that power and β would have to be determined separately for each different configuration of α_i's. Fortunately, power for the F test depends on the α_i's only through $\sum \alpha_i^2$, and so it can be simultaneously evaluated for many different alternatives. For example, $\sum \alpha_i^2 = 4$ for each of the following sets of α_i's, so power is identical for all three alternatives:

1. $\alpha_1 = -1$, $\alpha_2 = -1$, $\alpha_3 = 1$, $\alpha_4 = 1$
2. $\alpha_1 = -\sqrt{2}$, $\alpha_2 = \sqrt{2}$, $\alpha_3 = 0$, $\alpha_4 = 0$
3. $\alpha_1 = -\sqrt{3}$, $\alpha_2 = \sqrt{1/3}$, $\alpha_3 = \sqrt{1/3}$, $\alpha_4 = \sqrt{1/3}$

When H_0 is false, the test statistic MSTr/MSE has a **noncentral F distribution**, a three-parameter family. For one-way ANOVA, the first two parameters are still $v_1 = $ numerator df $= I - 1$ and $v_2 = $ denominator df $= IJ - I$, while the **noncentrality parameter** λ is given by

$$\lambda = \frac{J}{\sigma^2} \sum \alpha_i^2$$

Power is an increasing function of the noncentrality parameter λ (and β is a decreasing function of λ). Thus, for fixed values of σ^2 and J, the null hypothesis is more likely to be rejected for alternatives far from H_0 (large $\sum \alpha_i^2$) than for alternatives close to H_0. For a fixed value of $\sum \alpha_i^2$, power increases and β decreases as the sample size J on each treatment increases, whereas power decreases and β increases as the error variance σ^2 increases (since greater underlying variability makes it more difficult to detect any given departure from H_0).

Because hand computation of power, β, and sample size for the F test are quite difficult (as in the case of t tests), software is typically required. For one-way ANOVA, and with the noncentral F parameters specified as above,

$$\begin{aligned} \beta &= P(H_0 \text{ is not rejected when } H_0 \text{ is false}) \\ &= P\big(F < F_{\alpha, I-1, IJ-I} \text{ when } F \sim \text{noncentral } F\big) \\ &= \text{noncentral cdf evalulated at } F_{\alpha, I-1, IJ-I} \end{aligned}$$

and power $= 1 - \beta$. Many statistical packages (including SAS, JMP, and R) have a function that calculates the cumulative area under a noncentral F curve (required inputs are the critical value F_α, the numerator df, the denominator df, and λ), and this area is β.

Example 11.9 The effects of four different heat treatments on yield point (tons/in^2) of steel ingots are to be investigated. A total of $J = 8$ ingots will be cast using each treatment. Suppose the true standard deviation of yield point for any of the four treatments is $\sigma = 1$. How likely is it that H_0 will not be rejected at level .05 if three of the treatments have the same expected yield point and the other treatment has an expected yield point that is 1 ton/in^2 greater than the common value of the other three (i.e., the fourth yield is on average 1 standard deviation above those for the first three treatments)?

Suppose that $\mu_1 = \mu_2 = \mu_3$ and $\mu_4 = \mu_1 + 1$, so $\mu = (\sum \mu_i)/4 = \mu_1 + \frac{1}{4}$. Then $\alpha_1 = \mu_1 - \mu = -\frac{1}{4}$, $\alpha_2 = -\frac{1}{4}$, $\alpha_3 = -\frac{1}{4}$, $\alpha_4 = \frac{3}{4}$ so

$$\lambda = \frac{8}{1^2} \left[\left(-\frac{1}{4} \right)^2 + \left(-\frac{1}{4} \right)^2 + \left(-\frac{1}{4} \right)^2 + \left(\frac{3}{4} \right)^2 \right] = 6$$

The degrees of freedom are $v_1 = I - 1 = 3$ and $v_2 = IJ - I = 28$, and the F critical value for the .05 test is $F_{.05,3,28} = 2.947$. The probability of a type II error here is $\beta \approx .54$, obtained through the R command `pf(2.947,df1=3,df2=28,ncp=6)`, and power $\approx .46$. This power is rather low, so we might decide to increase the value of J. How many ingots of each type would be required to yield $\beta \approx .05$ (about 95% power) for the alternative under consideration? By trying different values of J, it can be verified that $J = 24$ will meet the requirement, but any smaller J will not.   ∎

In lieu of directly accessing the noncentral F cdf, some software packages will calculate power when the user specifies all the necessary information. For example, R has a function that allows specification of all I of the means, along with any three among J, σ^2, α, and power. The function calculates whichever quantity is unspecified. For example, we might wish to calculate the power of the test with $\alpha = .05$, $\sigma = 1$, $I = 4$, $J = 2$, $\mu_1 = 100$, $\mu_2 = 101$, $\mu_3 = 102$, and $\mu_4 = 106$. The R function calculates power $= .904$ (and so $\beta = .096$).

Minitab v.19 does something rather different. The user is asked to specify the maximum difference between μ_i's rather than the individual means. For example, in the previous scenario the maximum difference is $106 - 100 = 6$. However, power depends not only on this maximum difference but on the values of all the μ_i's. In this situation Minitab calculates the smallest possible value of power subject to $\mu_1 = 100$ and $\mu_4 = 106$, which occurs when the two other μ_i's are both halfway between 100 and 106. This power is .86, so we can say that the power is at least .86 and β is at most .14 when the two most extreme μ_i's are separated by 6. The software will also determine the necessary common sample size if maximum difference and minimum power are specified.

Relationship of the *F* Test to the *t* Test

When the number of populations is just $I = 2$, the ANOVA F test is testing H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 \neq \mu_2$. In this case, a two-tailed, two-sample t test could also be used. In Section 10.2, we mentioned the *pooled t test*, which requires equal variances, as an alternative to the two-sample t procedure. With a little algebra, it can be shown that the single-factor ANOVA F test and the two-tailed pooled t test are equivalent: for any given data set, the P-values for the two tests will be identical, so the same conclusion will be reached by either test. (The test statistic values are related by $f = t^2$.)

The two-sample t test is more flexible than the F test when $I = 2$ for two reasons. First, it is not based on the assumption that $\sigma_1 = \sigma_2$; second, it can be used to test H_a: $\mu_1 > \mu_2$ (an upper-tailed t test) or H_a: $\mu_1 < \mu_2$ (a lower-tailed test) as well as H_a: $\mu_1 \neq \mu_2$.

Single-Factor ANOVA When Sample Sizes Are Unequal

When the sample sizes from each population or treatment are not equal (i.e., an unbalanced study design), let $J_1, J_2, \ldots, J_I$ denote the I sample sizes and let $n = \sum_i J_i$ denote the total number of observations. The accompanying box gives ANOVA formulas and the test procedure.

Grand mean: $\overline{X}_{..} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J_i} X_{ij} = \frac{1}{n}\sum_{i=1}^{I} J_i \overline{X}_{i\cdot}$ (a *weighted* average of the sample means)

Fundamental ANOVA Identity: SST = SSTr + SSE

where the three sums of squares and associated dfs are

$$\text{SSTr} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} \left(\overline{X}_{i\cdot} - \overline{X}_{..}\right)^2 = \sum_{i=1}^{I} J_i\left(\overline{X}_{i\cdot} - \overline{X}_{..}\right)^2 \qquad \text{df} = I - 1$$

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} \left(X_{ij} - \overline{X}_{i\cdot}\right)^2 = \sum_{i=1}^{I} (J_i - 1)S_i^2 \qquad \text{df} = \sum_{i=1}^{I} (J_i - 1) = n - I$$

$$\text{SST} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} \left(X_{ij} - \overline{X}_{..}\right)^2 = \text{SSTr} + \text{SSE} \qquad \text{df} = n - 1$$

Test statistic value:

$$f = \frac{\text{MSTr}}{\text{MSE}} \quad \text{where MSTr} = \frac{\text{SSTr}}{I - 1} \text{ and MSE} = \frac{\text{SSE}}{n - I}$$

Rejection region: $f \geq F_{\alpha, I-1, n-I}$

P-value: area under the $F_{I-1, n-I}$ curve to the right of f

Verification of the fundamental ANOVA identity proceeds as in the case of equal sample sizes. However, it is somewhat trickier here to show that MSTr/MSE has the F distribution under H_0. Validity of the test procedure requires assuming, as before, that the population distributions are all normal with the same variance. The methods described at the end of Section 11.1 for assessing these with the residuals $e_{ij} = x_{ij} - \bar{x}_{i\cdot}$ can also be applied here.

Example 11.10 The article "On the Development of a New Approach for the Determination of Yield Strength in Mg-Based Alloys" (*Light Metal Age*, Oct. 1998: 51–53) presented the following data on elastic modulus (GPa) obtained by a new ultrasonic method for specimens of an alloy produced using three different casting processes.

Process				Observations					J_i	$\bar{x}_{i\cdot}$	s_i
Permanent molding	45.5	45.3	45.4	44.4	44.6	43.9	44.6	44.0	8	44.71	.624
Die casting	44.2	43.9	44.7	44.2	44.0	43.8	44.6	43.1	8	44.06	.501
Plaster molding	46.0	45.9	44.8	46.2	45.1	45.5			6	45.58	.549
									22		

Let μ_1, μ_2, and μ_3 denote the true average elastic moduli for the three different processes under the given circumstances. The relevant hypotheses are H_0: $\mu_1 = \mu_2 = \mu_3$ versus H_a: at least two of the μ_i's are different. The test statistic is, of course, $F = \text{MSTr}/\text{MSE}$, based on $I - 1 = 2$ numerator df and $n - I = 22 - 3 = 19$ denominator df. Relevant quantities include

$$\bar{x}_{\cdot\cdot} = \frac{1}{22}[8(357.7) + 8(352.5) + 6(273.5)] = 44.71$$

$$\text{SSTr} = 8(44.71 - 44.71)^2 + 8(44.06 - 44.71)^2 + 6(45.58 - 44.71)^2 = 7.93$$

$$\text{SSE} = (8-1)(.624)^2 + (8-1)(.501)^2 + (6-1)(.549)^2 = 6.00$$

The remaining computations are displayed in the accompanying ANOVA table. Since $f = 12.56 > F_{.001,2,19} = 10.16$, the P-value is smaller than .001. Thus the null hypothesis should be rejected at any reasonable significance level; there is compelling evidence for concluding that true average elastic modulus somehow depends on which casting process is used.

Source of variation	df	Sum of squares	Mean square	f
Treatments	2	7.93	3.965	12.56
Error	19	6.00	.3158	
Total	21	13.93		

∎

Multiple Comparisons When Sample Sizes Are Unequal

There is more controversy among statisticians regarding which multiple comparisons procedure to use when sample sizes are unequal than there is in the case of equal sample sizes. The procedure that we present here is called the *Tukey–Kramer* procedure for use when the I sample sizes $J_1, J_2, \ldots, J_I$ are reasonably close to each other ("mild imbalance"). It modifies Tukey's method [Equation (11.2)] by using averages of pairs of $1/J_i$'s in place of $1/J$. Let

$$d_{ij} = Q_{\alpha,I,n-I} \cdot \sqrt{\frac{\text{MSE}}{2}\left(\frac{1}{J_i} + \frac{1}{J_j}\right)}$$

Then the probability is *approximately* $1 - \alpha$ that

$$\left(\bar{X}_{i\cdot} - \bar{X}_{j\cdot}\right) - d_{ij} \le \mu_i - \mu_j \le \left(\bar{X}_{i\cdot} - \bar{X}_{j\cdot}\right) + d_{ij}$$

for every i and j ($i = 1, \ldots, I$ and $j = 1, \ldots, I$) with $i \ne j$.

The simultaneous confidence level $100(1 - \alpha)\%$ is now only approximate. The underscoring method can still be used, but now the honestly significant difference d_{ij} used to decide whether $\bar{x}_{i\cdot}$ and $\bar{x}_{j\cdot}$ can be connected by a line segment will depend on J_i and J_j.

Example 11.11 (Example 11.10 continued) The sample sizes for the elastic modulus data were $J_1 = 8$, $J_2 = 8$, $J_3 = 6$, and $I = 3$, $n - I = 19$, MSE = .316. A simultaneous confidence level of approximately 95% requires $Q_{.05,3,19} = 3.59$, from which

$$d_{12} = 3.59\sqrt{\frac{.316}{2}\left(\frac{1}{8} + \frac{1}{8}\right)} = .713 \qquad d_{13} = .771 \qquad d_{23} = .771$$

Since $\bar{x}_{1\cdot} - \bar{x}_{2\cdot} = 44.71 - 44.06 = .65 < d_{12}$, μ_1 and μ_2 are judged not significantly different. The accompanying underscoring scheme shows that μ_1 and μ_3 differ significantly, as do μ_2 and μ_3.

2. Die	1. Permanent	3. Plaster
44.06	44.71	45.58

∎

Data Transformation

The use of ANOVA methods can be invalidated by substantial differences in the variances $\sigma_1^2, \ldots, \sigma_I^2$, which until now have been assumed equal with common value σ^2. It sometimes happens that $V(X_{ij}) = \sigma_i^2 = g(\mu_i)$, a known function of μ_i (so that when H_0 is false, the variances are not equal). For example, if X_{ij} has a Poisson distribution with parameter μ_i (approximately normal if $\mu_i \geq 10$), then $\sigma_i^2 = \mu_i$, so $g(\mu_i) = \mu_i$ is the known function. In such cases, one can often transform the X_{ij}'s to $h(X_{ij})$'s so that they will have approximately equal variances (while hopefully leaving the transformed variables approximately normal), and then the F test can be used on the transformed observations. The basic idea is that, if $h(\cdot)$ is a smooth function, then we can express it approximately using the first terms of a Taylor series: $h(X_{ij}) \approx h(\mu_i) + h'(\mu_i)(X_{ij} - \mu_i)$. Then $V[h(X_{ij})] \approx V(X_{ij}) \cdot [h'(\mu_i)]^2 = g(\mu_i) \cdot [h'(\mu_i)]^2$. We now wish to find the function $h(\cdot)$ for which $g(\mu_i) \cdot [h'(\mu_i)]^2 = c$ (a constant) for every i. Solving this for $h'(\mu_i)$ and integrating give the following result.

PROPOSITION   If $V(X_{ij}) = g(\mu_i)$, a known function of μ_i, then a transformation $h(X_{ij})$ that "stabilizes the variance" so that $V[h(X_{ij})]$ is approximately the same for each i is given by $h(x) \propto \int [g(x)]^{-1/2} dx$.

In the Poisson case, $g(x) = x$, so $h(x)$ should be proportional to $\int x^{-1/2} dx = 2x^{1/2}$. Thus Poisson data should be transformed to $h(x_{ij}) = \sqrt{x_{ij}}$ before the analysis.

A Random Effects Model

The single-factor problems considered so far have all been assumed to be examples of a **fixed effects** ANOVA model. By this we mean that the chosen levels of the factor under study are the *only* ones considered relevant by the experimenter. The single-factor fixed effects model is

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{with} \quad \sum \alpha_i = 0 \tag{11.4}$$

where the ε_{ij}'s are random and both μ and the α_i's are fixed parameters whose values are unknown.

In some single-factor problems, the particular levels studied by the experimenter are chosen, either by design or through sampling, from a large population of levels. For example, to study the effect of using different operators on task performance time for a particular machine, a sample of five operators might be chosen from a large pool of operators. Similarly, the effect of soil pH on the yield of soybean plants might be studied by using soils with four specific pH values chosen from among the many possible pH levels. When the levels used are selected at random from a larger population of possible levels, the factor is said to be random rather than fixed, and the fixed effects model (11.4) is no longer appropriate. An analogous **random effects** model is obtained by replacing the fixed α_i's in (11.4) by random variables. The resulting model description is

$$X_{ij} = \mu + A_i + \varepsilon_{ij} \quad \text{with} \quad E(A_i) = E(\varepsilon_{ij}) = 0$$
$$V(\varepsilon_{ij}) = \sigma^2 \quad V(A_i) = \sigma_A^2 \tag{11.5}$$

with all A_i's and ε_{ij}'s normally distributed and independent of each other.

The condition $E(A_i) = 0$ in (11.5) is similar to the condition $\sum \alpha_i = 0$ in (11.4); it states that the expected or average effect of the ith level measured as a departure from μ is zero.

For the random effects model (11.5), the hypothesis of no effects due to different levels is $H_0: \sigma_A^2 = 0$, which says that different levels of the factor contribute nothing to variability of the response. Critically, although the hypotheses in the single-factor fixed and random effects models are different, *they are tested in exactly the same way*: by forming $F = \text{MSTr}/\text{MSE}$ and rejecting H_0 if $f \geq F_{\alpha, I-1, n-I}$. This can be justified intuitively by noting in the random effects model that $E(\text{MSE}) = \sigma^2$ (as for fixed effects), whereas

$$E(\text{MSTr}) = \sigma^2 + \frac{1}{I-1}\left(n - \frac{\sum J_i^2}{n}\right)\sigma_A^2 \tag{11.6}$$

where again $J_1, J_2, \ldots, J_I$ are the sample sizes and $n = \sum J_i$. The factor in parentheses on the right side of (11.6) is nonnegative, so once again $E(\text{MSTr}) = \sigma^2$ if H_0 is true (i.e., if $\sigma_A^2 = 0$) and $E(\text{MSTr}) > \sigma^2$ if H_0 is false.

Example 11.12 When items are machined out of metal (or plastic or wood) sheets by drills, undesirable burrs form along the edge. The article "Observation of Drilling Burr and Finding out the Condition for Minimum Burr Formation" (*Int. J. Manuf. Engr.* 2014) reports on a study of the effect that cutting speed has on burr size. Eighteen measurements were made at each of three speeds (20, 25, and 31 m/min) that were randomly selected from the range of possible speeds for the particular equipment used in the experiment. Each measurement is the burr height (mm) from drilling into a low-alloy steel specimen. The data is summarized in the accompanying table along with the derived ANOVA table. The very small f statistic and correspondingly large P-value indicates that $H_0: \sigma_A^2 = 0$ should not be rejected. The data does not indicate cutting speed impacts burr size.

Speed (m/min)	$\bar{x}_{i\cdot}$	s_i
20	1.558	2.018
25	1.998	2.415
31	1.867	2.148

Source of variation	df	SS	MS	f	P-value
Cutting speed	2	1.837	0.9186	0.19	.828
Error	51	246.795	4.8931		
Total	53	248.632			

∎

Exercises: Section 11.3 (27–44)

27. The following data refers to yield of tomatoes (kg/plot) for four different levels of salinity; salinity level here refers to electrical conductivity (EC), where the chosen levels were EC = 1.6, 3.8, 6.0, and 10.2 nmhos/cm:

EC			Yield		
1.6	59.5	53.3	56.8	63.1	58.7
3.8	55.2	59.1	52.8	54.5	
6.0	51.7	48.8	53.9	49.0	
10.2	44.6	48.5	41.0	47.3	46.1

Use the F test at level $\alpha = .05$ to test for any differences in true average yield due to the different salinity levels.

28. Apply the modified Tukey's method to the data in the previous exercise to identify significant differences among the μ_i's.

29. A study at Bentley College, a large business school in the eastern United States, examined students' anxiety levels toward the subject of accounting ("Determinants of Accounting Anxiety in Business Students," *J. College Teach. Learn.* 2004). A representative sample of 1020 students completed the Accounting Anxiety Rating Scale (AARS) questionnaire; higher scores (out of 100) indicate greater anxiety. Summary data broken down by grade level appears in the accompanying table.

Class level	n	$\bar{x}$	s
Freshman	86	48.95	9.13
Sophomore	224	51.45	11.29
Junior	225	52.89	11.32
Senior	198	52.92	11.32
Graduate	287	45.55	10.10

a. Comment on the plausibility (or necessity) of the normality and equal variance assumptions for this example.

b. Test at $\alpha = .05$ the hypothesis that the mean accounting anxiety level for business students varies by class level.

c. Apply the Tukey–Kramer method to identify significant differences among the μ_i's.

30. The article "From Dark to Light: Skin Color and Wages among African Americans" (*J. Human Res.* 2007: 701–738) includes the following information on hourly wages ($) for representative samples of the indicated populations.

Skin color	n	$\bar{x}$	s
White	513	15.94	7.73
Light Black	51	14.42	6.05
Medium Black	177	13.23	6.64
Dark Black	207	11.72	5.60

a. Does population mean hourly wage appear to depend on skin color? Carry out an appropriate test of hypotheses.

b. Identify significant differences among the μ_i's at the .05 significance level.

31. The authors of the article "Exploring the Impact of Social Media Practices on Wine Sales in US Wineries" (*J. Direct Data, Digital Market. Pract.* 2016: 272–283) interviewed 361 winery managers. Each manager was asked to report, as a percentage of sales, the impact of social media use on their wine sales. Each winery's social media presence was then categorized by the number of social media platforms it used: 0–2, 3–5, or 6 or more. Summary information appears in the accompanying table. Test to see whether an association exists between social media presence and sales at the .01 significance level.

No. of platforms	n	Mean	SD
Two or fewer	107	12.76	13.00
Three to five	164	17.23	14.07
Six or more	90	21.56	17.35

32. The article "Can You Test Me Now? Equivalence of GMA Tests on Mobile and Non-Mobile Devices" (*Int. J. Select Assess.* 2017: 61–71) describes a study in which 1769 people were recruited through Amazon Mechanical Turk to take a general mental ability test. (Researchers used the WPT-R test, a variation on the Wonderlic Personnel Test used by the NFL to evaluate players.) Participants were randomly assigned to take the test on one of three electronic devices; the data is summarized below.

Device	n	Mean score	SD
Computer	724	33.98	7.65
Tablet	476	34.40	7.75
Smartphone	569	34.26	7.59

The goal of the study was to determine whether the three devices can be considered "equivalent" for the purpose of administering mental ability tests. Perform an ANOVA at the .05 significance level, and explain what you discover.

33. Lipids provide much of the dietary energy in the bodies of infants and young children. There is a growing interest in the quality of the dietary lipid supply during infancy as a major determinant of growth, visual and neural development, and long-term health. The article "Essential Fat Requirements of Preterm Infants" (*Amer. J. Clin. Nutrit.* 2000: 245S–250S) reported the following data on total polyunsaturated fats (%) for infants who were randomized to four different feeding regimens: breast milk, corn-oil-based formula, soy-oil-based formula, or soy-and-marine-oil-based formula:

Regimen	Sample size	Sample mean	Sample SD
Breast milk	8	43.0	1.5
CO	13	42.4	1.3
SO	17	43.1	1.2
SMO	14	43.5	1.2

a. What assumptions must be made about the four total polyunsaturated fat distributions before carrying out a single-factor ANOVA to decide whether there are any differences in true average fat content?

b. Carry out the test suggested in part (a). What can be said about the *P*-value?

34. Samples of six different brands of diet/imitation margarine were analyzed to determine the level of physiologically active polyunsaturated fatty acids (PAPFUA, in percentages). The data below is consistent with a study carried out by *Consumer Reports*:

Brand	PAPFUA				
Imperial	14.1	13.6	14.4	14.3	
Parkay	12.8	12.5	13.4	13.0	12.3
Blue Bonnet	13.5	13.4	14.1	14.3	
Chiffon	13.2	12.7	12.6	13.9	
Mazola	16.8	17.2	16.4	17.3	18.0
Fleischmann's	18.1	17.2	18.7	18.4	

a. Use ANOVA to test for differences among the true average PAPFUA percentages for the different brands.

b. Compute CIs for all $(\mu_i - \mu_j)$'s.

c. Mazola and Fleischmann's are corn-based, whereas the others are soybean-based. Compute a CI for

$$\frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} - \frac{\mu_5 + \mu_6}{2}$$

[*Hint*: Modify the expression for $V(\hat{\theta})$ that led to (11.3) in the previous section.]

35. Subacromial impingement syndrome (SIS) refers to shoulder pain resulting from a particular impingement of the rotator cuff tendon. The article "Evaluation of the Effectiveness of Three Physiotheraputic Treatments for SIS" (*Physiotherapy* 2016: 57–63) reports a study in which 99 SIS sufferers were randomly assigned to receive one of three treatments across 20 sessions. The Constant–Murley score (CMS), a

standard measure of shoulder functionality and pain, was administered to each subject before the experiment and again one month post-treatment. The accompanying table summarizes the change in CMS (higher numbers are better) for the subjects.

Treatment	n	Mean	SD
Ultrasound	32	5.1	10.3
Phonophoresis	33	6.4	9.3
Iontophoresis	34	6.5	18.1

Test to see whether the true mean increase in CMS differs across the three different treatments, at the $\alpha = .05$ significance level.

36. In single-factor ANOVA with sample sizes J_i ($i = 1, \ldots, I$), show that $\text{SSTr} = \sum_i J_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 = \sum_i J_i \overline{X}_{i\cdot}^2 - n\overline{X}_{\cdot\cdot}^2$, where $n = \sum_i J_i$.

37. When sample sizes are equal ($J_i = J$), the parameters $\alpha_1, \alpha_2, \ldots, \alpha_I$ of the alternative parameterization to the model equation are restricted by $\sum \alpha_i = 0$. For unequal sample sizes, the most natural restriction is $\sum J_i \alpha_i = 0$. Use this to show that

$$E(\text{MSTr}) = \sigma^2 + \frac{1}{I-1}\sum J_i \alpha_i^2$$

What is $E(\text{MSTr})$ when H_0 is true? [This expectation is correct if $\sum J_i \alpha_i = 0$ is replaced by the restriction $\sum \alpha_i = 0$, or any other single linear restriction on the α_i's used to reduce the model to I independent parameters, but $\sum J_i \alpha_i = 0$ simplifies the algebra and yields natural estimates for the model parameters.]

38. Reconsider Example 11.9 involving an investigation of the effects of different heat treatments on the yield point of steel ingots.

a. If $J = 8$ and $\sigma = 1$, what is β for a level .05 F test when $\mu_1 = \mu_2$, $\mu_3 = \mu_1 - 1$, and $\mu_4 = \mu_1 + 1$?

b. For the alternative μ_i's of part (a), what value of J is necessary to obtain $\beta = .05$?

c. If there are $I = 5$ heat treatments, $J = 10$, and $\sigma = 1$, what is β for the level .05 F test when four of the μ_i's are equal and the fifth differs by 1 from the other four?

39. For unequal sample sizes, the noncentrality parameter for F test power calculations is $\lambda = \sum J_i \alpha_i^2 / \sigma^2$. Referring to Exercise 27, what is the power of the test when $\mu_2 = \mu_3$, $\mu_1 = \mu_2 - \sigma$, and $\mu_4 = \mu_2 + \sigma$?

40. The following data on number of cycles to failure ($\times 10^6$) appears in the article "An Experimental Study of Influence of Lubrication Methods on Efficiency and Contact Fatigue Life of Spur Gears" (*J. Tribol.* 2018).

Lubrication condition	Cycles to failure ($\times 10^6$)					
D1	18.79	10.44	14.62	12.53	8.35	14.62
J1	16.70	18.79	12.53	26.10	10.44	18.27
J2	10.44	14.62	16.70	29.23	22.97	18.50
J4	6.26	6.26	6.26	6.26	4.70	5.22

a. Calculate the standard deviation of each sample. Why should we be reluctant to proceed with an analysis of variance?

b. Take the logarithm of the observations. Do these transformed values adhere better to the conditions for ANOVA? Explain.

c. Perform one-way ANOVA on these transformed values.

41. Many studies have been conducted to measure mercury (Hg) levels in fish, but little information exists on Hg concentrations in marine mammals. The article "Factors Influencing Exposure of North American River Otters…to Mercury Relative to a Large-Scale Reservoir in Northern British Columbia, Canada" (*Ecotoxicology* 2019: 343–353) describes a study in which river otters living in five Canadian

reservoirs were tested for Hg concentration (mg/kg). The accompanying summary information was extracted from a graph in the article.

Reservoir	n	Hg concentration Mean	Hg concentration SD	ln(Concentration) Mean	ln(Concentration) SD
PW	15	4.1	3.0	1.30	.41
MK	7	9.8	4.5	2.28	.21
NW	20	14.2	6.7	2.62	.18
50K	20	16.1	10.4	2.75	.20
FR	27	18.2	7.1	2.82	.38

Hg concentration distributions at all five reservoirs were heavily right-skewed.

a. Why should we hesitate to perform one-way ANOVA on the Hg concentration data?
b. Consider the transformation $y = \ln(\text{concentration})$. Estimated summary information appears above, and taking the logarithm greatly reduces skewness. Apply the ANOVA F test to the transformed data, and report your findings at the .05 significance level.
c. Apply Tukey's method to the log-transformed data, if appropriate.

42. Simplify $E(\text{MSTr})$ for the random effects model when $J_1 = J_2 = \cdots = J_I = J$.

43. Suppose that X_{ij} is a binomial variable with parameters n and p_i (so it is approximately normal when $np_i \geq 10$ and $nq_i \geq 10$). Then since $\mu_i = np_i$, $V(X_{ij}) = \sigma_i^2 = np_i(1 - p_i) = \mu_i(1 - \mu_i/n)$. How should the X_{ij}'s be transformed so as to stabilize the variance? [Hint: $g(\mu_i) = \mu_i(1 - \mu_i/n)$.]

44. In an experiment to compare the quality of four different brands of magnetic tape (A–D), five 5000-foot reels of each brand were selected and the number of cosmetic flaws on each reel was determined.

Brand	No. of flaws				
A	10	5	12	14	8
B	14	12	17	9	8
C	13	18	10	15	18
D	17	16	12	22	14

It is believed that the number of flaws has approximately a Poisson distribution for each brand. Analyze the data at level .01 to see whether the expected number of flaws per reel is the same for each brand.

11.4 Two-Factor ANOVA without Replication

In many situations there are two factors of simultaneous interest. For example, a baker might experiment with $I = 3$ temperatures (325, 350, 375 °F) and $J = 2$ baking times (45 min, 60 min) to optimize a new cake recipe. Or, an industrial engineer might wish to study the surface roughness resulting from a certain machining process; she might carry out an experiment at various combinations of cutting speed and feed rate.

Call the two factors of interest A and B. When factor A has I levels and factor B has J levels, there are IJ different combinations (pairs) of levels of the two factors, each called a **treatment**. If the data includes multiple observations for each treatment, the study is said to include **replication**. With K_{ij} = the number of observations on the treatment (factor A, factor B) = (i, j), we focus in this section on the case $K_{ij} = 1$ (i.e., no replication), so that the data consists of IJ observations. We will first discuss the fixed effects model, in which the only levels of interest for the two factors are those actually represented in the study. The case in which at least one factor is random is considered briefly at the end of the section.

Example 11.13 Is it really as easy to remove marks on fabrics from erasable pens as the word "erasable" might imply? Consider the following data from an experiment to compare three different brands of pens and four different wash treatments with respect to their ability to remove marks on a particular type of fabric (based on "An Assessment of the Effects of Treatment, Time, and Heat on the Removal of Erasable Pen Marks from Cotton and Cotton/Polyester Blend Fabrics," *J. Test. Eval.* 1991: 394–397). The response variable is a quantitative indicator of overall specimen color change; the lower this value, the more marks were removed.

		Washing treatment				
		1	2	3	4	Average
	1	.97	.48	.48	.46	.598
Brand of pen	2	.77	.14	.22	.25	.345
	3	.67	.39	.57	.19	.455
	Average	.803	.337	.423	.300	

Is there any difference in the true average amount of color change due either to the different brands of pen or to the different washing treatments? ■

As in single-factor ANOVA, double subscripts are used to identify random variables and observed values. Let

X_{ij} = the random variable denoting the measurement when (factor A, factor B) = (i, j)
x_{ij} = the observed value of X_{ij}

The x_{ij}'s are usually presented in a two-way table in which the ith row contains the observed values when factor A is held at level i and the jth column contains the observed values when factor B is held at level j. In the erasable-pen experiment of Example 11.13, the number of levels of factor A (pen brand) is $I = 3$, the number of levels of factor B (washing treatment) is $J = 4$; $x_{13} = .48$, $x_{22} = .14$, and so on.

Whereas in single-factor ANOVA we were interested only in row means and the grand mean, here we are interested also in column means. Let

$$\overline{X}_{i\cdot} = \begin{array}{l}\text{the average of data obtained}\\\text{when factor } A \text{ is held at level } i\end{array} = \frac{1}{J}\sum_{j=1}^{J} X_{ij}$$

$$\overline{X}_{\cdot j} = \begin{array}{l}\text{the average of data obtained}\\\text{when factor } B \text{ is held at level } j\end{array} = \frac{1}{I}\sum_{i=1}^{I} X_{ij}$$

$$\overline{X}_{\cdot\cdot} = \text{the grand mean} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}$$

with observed values $\bar{x}_{i\cdot}$, $\bar{x}_{\cdot j}$, and $\bar{x}_{\cdot\cdot}$. Intuitively, to see whether there is any effect due to the levels of factor A, we should compare the observed $\bar{x}_{i\cdot}$'s with each other, and information about the different levels of factor B should come from the $\bar{x}_{\cdot j}$'s.

A Two-Factor Fixed Effects Model

Proceeding by analogy to single-factor ANOVA, one's first inclination in specifying a model is to let μ_{ij} = the true average response when (factor A, factor B) = (i, j), giving IJ mean parameters. Then let

$$X_{ij} = \mu_{ij} + \varepsilon_{ij}$$

where ε_{ij} is the random amount by which the observed value differs from its expectation, and the ε_{ij}'s are assumed normal and independent with common variance σ^2. Unfortunately, there is no valid test procedure for this choice of parameters. The reason is that under the alternative hypothesis of interest, the μ_{ij}'s are free to take on any values whatsoever and σ^2 can be any value greater than zero, so that there are $IJ + 1$ freely varying parameters. But there are only IJ observations, so after using each x_{ij} as an estimate of μ_{ij}, there is no way to estimate σ^2.

To rectify this problem of a model having more parameters than observed values, we must specify a model that is realistic yet involves relatively few parameters. For the no-replication ($K_{ij} = 1$) scenario, we assume the existence of a parameter μ, I parameters $\alpha_1, \alpha_2, \ldots, \alpha_I$, and J parameters $\beta_1, \beta_2, \ldots, \beta_J$ such that

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (i = 1, \ldots, I; \quad j = 1, \ldots, J) \tag{11.7}$$

Taking expectations on both sides of (11.7) yields

$$\mu_{ij} = \mu + \alpha_i + \beta_j \tag{11.8}$$

The model specified in (11.7) and (11.8) is called an **additive model**, because each mean response μ_{ij} is the sum of a true grand mean (μ), an effect due to factor A at level i (α_i), and an effect due to factor B at level j (β_j). The difference between mean responses for factor A at levels i and i' when B is held at level j is $\mu_{ij} - \mu_{i'j}$. Critically, when the model is additive,

$$\mu_{ij} - \mu_{i'j} = (\mu + \alpha_i + \beta_j) - (\mu + \alpha_{i'} + \beta_j) = \alpha_i - \alpha_{i'}$$

which is independent of the level j of the factor B. A similar result holds for $\mu_{ij} - \mu_{ij'}$. Thus additivity means that *the difference in mean responses for two levels of one of the factors is the same for all levels of the other factor*. Figure 11.4a shows a set of mean responses that satisfy the condition of additivity (which implies parallel lines), and Figure 11.4b shows a nonadditive configuration of mean responses.
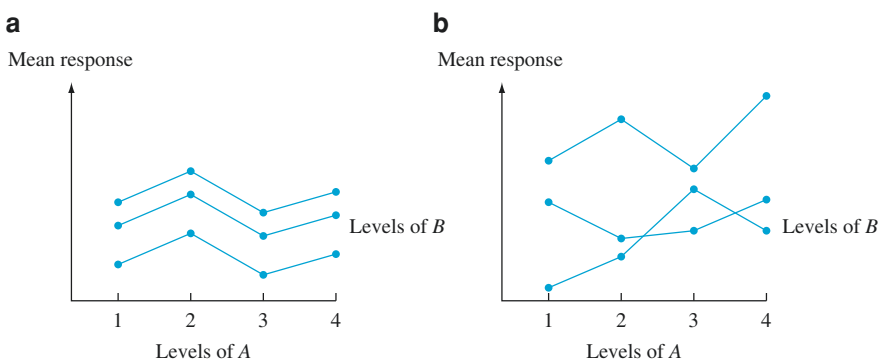
Figure 11.4 Mean responses for two types of model: (**a**) additive; (**b**) nonadditive

If additivity does not hold, we say that **interaction** is present. Factors A and B have an interaction effect on the response variable if the effect of factor A on the (mean) response value depends upon the level of factor B, and vice versa. The foregoing discussion implies that an additive model assumes there is no interaction effect. The graphs in Figure 11.4 are called **interaction plots**; Figure 11.4a displays data consistent with no interaction effect, whereas Figure 11.4b indicates a potentially strong interaction.

When $K_{ij} = 1$, there is insufficient data to estimate any potential interaction effects, and so the additive model specified by (11.7) and (11.8) must be used. In Section 11.5, where $K_{ij} > 1$, we will consider models that include interaction effects.

Example 11.14 (Example 11.13 continued) When the observed x_{ij}'s are plotted in a manner analogous to that of Figure 11.4, we get the result shown in Figure 11.5. Although there is some "crossing over" in the observed x_{ij}'s, the configuration is reasonably representative of what would be expected under additivity with just one observation per treatment.
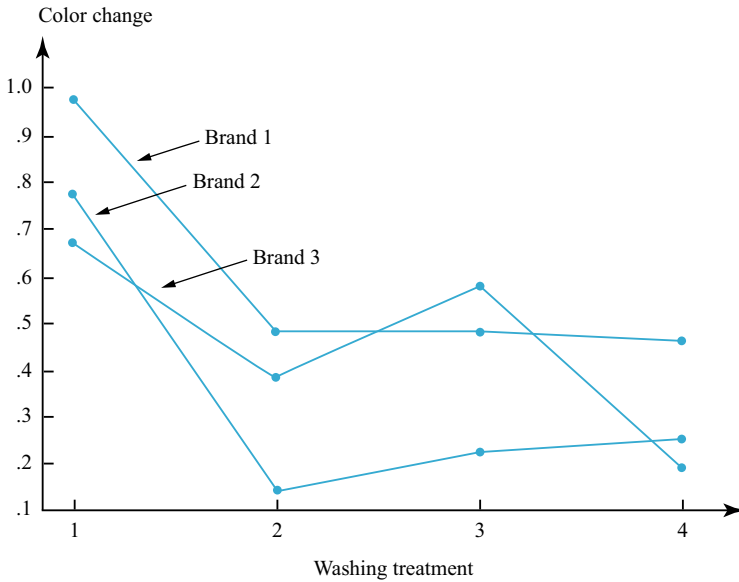
Figure 11.5 Plot of data from Example 11.13   ■

Expression (11.8) is still not quite our final model description, because the α_i's and β_j's are not uniquely determined. Following are two different configurations of the α_i's and β_j's (with $\mu = 0$ for convenience) that yield the same additive μ_{ij}'s.

	$\beta_1 = 1$	$\beta_2 = 4$		$\beta_1 = 2$	$\beta_2 = 5$
$\alpha_1 = 1$	$\mu_{11} = 2$	$\mu_{12} = 5$	$\alpha_1 = 0$	$\mu_{11} = 2$	$\mu_{12} = 5$
$\alpha_2 = 2$	$\mu_{21} = 3$	$\mu_{22} = 6$	$\alpha_2 = 1$	$\mu_{21} = 3$	$\mu_{22} = 6$

By subtracting any constant c from all α_i's and adding c to all β_j's, other configurations corresponding to the same additive model are obtained. This nonuniqueness is eliminated by use of the following model, which imposes an extra constraint on the α_i's and β_j's.

TWO-FACTOR ANOVA ADDITIVE MODEL EQUATION

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \tag{11.9}$$

where $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, and the ε_{ij}'s are assumed to be independent normal rvs with mean 0 and variance σ^2.

This is analogous to the alternative choice of parameters for single-factor ANOVA discussed in Section 11.3. It is not difficult to verify that (11.9) is an additive model in which the parameters are uniquely determined. Notice that there are now only $I - 1$ independently determined α_i's and $J - 1$ independently determined β_j's, so including μ Expression (11.9) specifies $(I - 1) + (J - 1) + 1 = I + J - 1$ parameters.

The interpretation of the parameters of (11.9) is straightforward: μ is the true grand mean response over all levels of both factors; α_i is the effect of factor A at level i measured as a deviation from μ; and β_j is the effect of factor B at level j. Unbiased (and maximum likelihood) estimators for these parameters are

$$\hat{\mu} = \overline{X}_{..} \quad \hat{\alpha}_i = \overline{X}_{i.} - \overline{X}_{..} \quad \hat{\beta}_j = \overline{X}_{.j} - \overline{X}_{..}$$

Test Procedures

There are two different hypotheses of interest in a two-factor experiment with $K_{ij} = 1$. The first, denoted by H_{0A}, states that the different levels of factor A have no effect on true average response. The second, denoted by H_{0B}, asserts that there is no factor B effect.

$$H_{0A}: \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$
$$\text{versus } H_{aA}: \text{ at least one } \alpha_i \neq 0$$
$$H_{0B}: \beta_1 = \beta_2 = \cdots = \beta_J = 0$$
$$\text{versus } H_{aB}: \text{ at least one } \beta_j \neq 0$$

No factor A effect implies that all α_i's are equal, so they must all be 0 since they sum to 0, and similarly for the β_j's. The analysis now follows closely that for single-factor ANOVA. The relevant sums of squares and associated dfs are given in the accompanying box.

$$SSA = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2 = J\sum_{i=1}^{I} \hat{\alpha}_i^2 \quad df = I - 1$$

$$SSB = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot}\right)^2 = I\sum_{j=1}^{J} \hat{\beta}_j^2 \quad df = J - 1$$

$$SSE = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}_{\cdot\cdot}\right)^2 \quad df = (I-1)(J-1)$$

$$SST = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(X_{ij} - \overline{X}_{\cdot\cdot}\right)^2 \quad df = IJ - 1$$

The fundamental ANOVA identity is

$$SST = SSA + SSB + SSE$$

SSA and SSB take the place of SSTr from single-factor ANOVA. The unwieldy expression for SSE results from replacing μ, α_i, and β_j in $\sum \left[X_{ij} - \left(\mu + \alpha_i + \beta_j\right)\right]^2$ by their respective estimators. Error df is IJ – [number of mean parameters estimated] $= IJ - (I + J - 1) = (I - 1)(J - 1)$. Analogous to single-factor ANOVA, total variation is split into a part (SSE) that cannot be attributed to the truth or falsity of H_{0A} and H_{0B} (i.e., unexplained variation) and two parts that can be explained by possible falsity of the two null hypotheses.

Forming F ratios as in single-factor ANOVA, it can be shown as in Section 11.1 that if H_{0A} is true, the corresponding ratio has an F distribution with numerator df $= I - 1$ and denominator df $= (I - 1)(J - 1)$; an analogous result applies when testing H_{0B}.

Hypotheses	Test Statistic Value	Rejection Region
H_{0A} versus H_{aA}	$f_A = MSA/MSE$	$f_A \geq F_{\alpha, I-1, (I-1)(J-1)}$
H_{0B} versus H_{aB}	$f_B = MSB/MSE$	$f_B \geq F_{\alpha, J-1, (I-1)(J-1)}$

The corresponding P-values for the two tests are the areas under the associated F curves to the right of f_A and f_B, respectively.

Example 11.15 (Example 11.13 continued) The $\bar{x}_{i\cdot}$'s (row means) and $\bar{x}_{\cdot j}$'s (column means) for the color change data are displayed along the right and bottom margins of the data table in Example 11.13. In addition, the grand mean is $\bar{x}_{\cdot\cdot} = .466$. Table 11.7 summarizes further calculations.

Table 11.7 ANOVA table for Example 11.15

Source of Variation	df	Sum of squares	Mean square	f	P-value
Factor A (pen brand)	$I - 1 = 2$	SSA = .1282	MSA = .0641	f_A = 4.43	.066
Factor B (wash treatment)	$J - 1 = 3$	SSB = .4797	MSB = .1599	f_B = 11.05	.007
Error	$(I - 1)(J - 1) = 6$	SSE = .0868	MSE = .01447		
Total	$IJ - 1 = 11$	SST = .6947			

The critical value for testing H_{0A} at level of significance .05 is $F_{.05,2,6} = 5.14$. Since $4.43 < 5.14$, H_{0A} cannot be rejected at significance level .05. Based on this (small) data set, we cannot conclude that true average color change depends on brand of pen. Because $F_{.05,3,6} = 4.76$ and $11.05 \geq 4.76$, H_{0B} is rejected at significance level .05 in favor of the assertion that color change varies with washing treatment. The same conclusions result from consideration of the P-values: $.066 > .05$ and $.007 \leq .05$.

The plausibility of the normality and constant variance assumptions can be investigated graphically by first calculating the **predicted values** (also called **fitted values**) $\hat{x}_{ij}$ and the residuals (the differences between the observations and predicted values) e_{ij}:

$$\hat{x}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{x}_{..} + (\bar{x}_{i.} - \bar{x}_{..}) + (\bar{x}_{.j} - \bar{x}_{..}) = \bar{x}_{i.} + \bar{x}_{.j} - \bar{x}_{..}$$
$$e_{ij} = x_{ij} - \hat{x}_{ij} = x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}$$

We can check the normality assumption with a normal plot of the residuals, Figure 11.6a, and then the constant variance assumption with a plot of the residuals against the fitted values, Figure 11.6b.
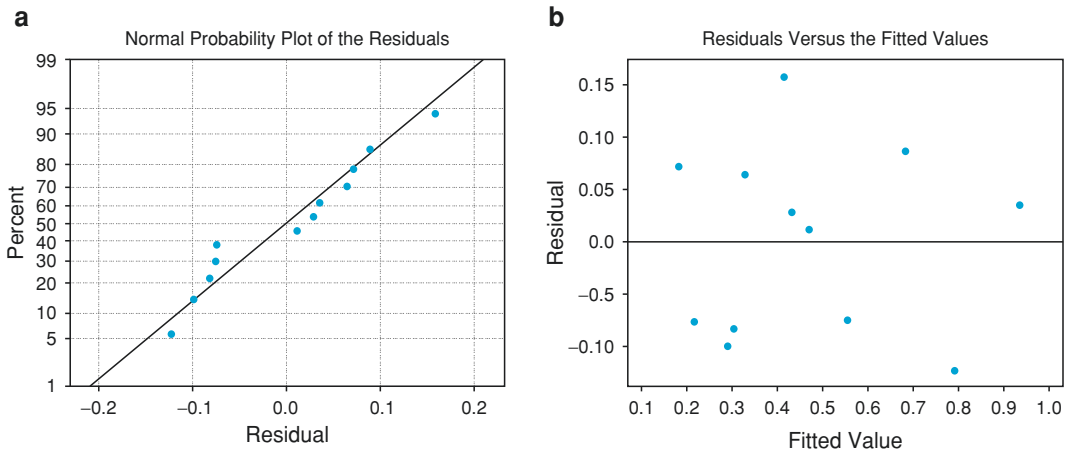
Figure 11.6 Residual plots for Example 11.15

The normal probability plot is reasonably straight, so there is no reason to question normality for this data set. In the plot of the residuals against the fitted values, look for differences in vertical spread as we move horizontally across the graph. For example, if there were a narrow range for small fitted values and a wide range for high fitted values, this would suggest that the variance is higher for larger responses (this happens often, and it can sometimes be cured by transforming via logarithms). No such problem occurs here, so there is no evidence against the constant variance assumption, either. ∎

Expected Mean Squares

The plausibility of using the F tests just described is demonstrated by determining the expected mean squares. After some tedious algebra,

$$E(\text{MSE}) = \sigma^2 \ (\text{when the model is additive})$$

$$E(\text{MSA}) = \sigma^2 + \frac{J}{I-1} \sum_{i=1}^{I} \alpha_i^2$$

$$E(\text{MSB}) = \sigma^2 + \frac{I}{J-1} \sum_{j=1}^{J} \beta_j^2$$

When H_{0A} is true, MSA is an unbiased estimator of σ^2, so F_A is a ratio of two unbiased estimators of σ^2. When H_{0A} is false, MSA tends to overestimate σ^2, so H_{0A} should be rejected when the ratio F_A is too large. Similar comments apply to MSB and H_{0B}.

Multiple Comparisons

When either H_{0A} or H_{0B} has been rejected, Tukey's procedure can be used to identify significant differences between the levels of the factor under investigation. The steps in the analysis are identical to those for a single-factor ANOVA:

1. For comparing levels of factor A, obtain $Q_{\alpha, I, (I-1)(J-1)}$.
   For comparing levels of factor B, obtain $Q_{\alpha, J, (I-1)(J-1)}$.
2. Compute Tukey's honestly significant difference:

$$d = Q \cdot (\text{estimated SD of the sample means being compared})$$

$$= \begin{cases} Q_{\alpha, I, (I-1)(J-1)} \cdot \sqrt{\text{MSE}/J} \text{ for factor } A \text{ comparisons} \\ Q_{\alpha, J, (I-1)(J-1)} \cdot \sqrt{MSE/I} \text{ for factor } B \text{ comparisons} \end{cases}$$

(because, e.g., the standard deviation of $\overline{X}_{i\cdot} = (1/J) \sum X_{ij}$ is $\sigma/\sqrt{J}$).
3. Arrange the sample means in increasing order, then underscore those pairs differing by less than d. Pairs not underscored by the same line correspond to significantly different levels of the given factor.

Example 11.16 (Example 11.15 continued) Identification of significant differences among the four washing treatments requires $Q_{.05,4,6} = 4.90$ and $d = 4.90\sqrt{.01447/3} = .340$. The four factor B sample means (column averages) are now listed in increasing order, and any pair differing by less than .340 is underscored by a line segment:

$$\begin{array}{cccc} \overline{x}_{4\cdot} & \overline{x}_{2\cdot} & \overline{x}_{3\cdot} & \overline{x}_{1\cdot} \\ .300 & .337 & .423 & .803 \end{array}$$

Washing treatment 1 is significantly worse than the other three treatments, but no other significant differences are identified. In particular, it is not apparent which among treatments 2, 3, and 4 is best at removing marks.

Notice that Tukey's HSD is not required for comparing the levels of factor A, since the ANOVA F test for that factor did not reveal any statistically significant effect. ∎

Randomized Block Experiments

In using single-factor ANOVA to test for the presence of effects due to the I different treatments under study, once the IJ subjects or experimental units have been chosen, treatments should be allocated in a completely random fashion. That is, J subjects should be chosen at random for the first treatment, then another sample of J chosen at random from the remaining subjects for the second treatment, and so on.

It frequently happens, though, that subjects or experimental units exhibit differences with respect to other characteristics that may affect the observed responses. For example, some patients might be healthier than others. When this is the case, the presence or absence of a significant F value may be due to these other differences rather than to the presence or absence of factor effects. This was the reason for introducing paired experiments in Chapter 10. The generalization of the paired experiment to $I > 2$ is called a **randomized block** design. An extraneous factor, "blocks," is constructed by dividing the IJ units into J groups (with I units in each group) in such a way that *within each block*, the I units are homogeneous with respect to other factors thought to affect the responses. Then within each homogeneous block, the I treatments are randomly assigned to the I units or subjects in the block.

Example 11.17 A consumer product-testing organization wished to compare the annual power consumption for five different brands of dehumidifier. Because power consumption depends on the prevailing humidity level, it was decided to monitor each brand at four different levels ranging from moderate to heavy humidity (thus blocking on humidity level). Within each humidity level, brands were randomly assigned to the five selected locations. The resulting amount of power consumption (annual kWh) appears in Table 11.8, and the ANOVA calculations are summarized in Table 11.9.

Table 11.8 Power consumption data for Example 11.17

| Treatments (brands) | Blocks (humidity level) | | | | |
	1	2	3	4	$\bar{x}_{i\cdot}$
1	685	792	838	875	797.50
2	722	806	893	953	843.50
3	733	802	880	941	839.00
4	811	888	952	1005	914.00
5	828	920	978	1023	937.25

Table 11.9 ANOVA table for Example 11.17

Source of variation	df	Sum of squares	Mean square	f
Treatments (brands)	4	53,231.00	13,307.75	f_A = 95.57
Blocks	3	116,217.75	38,739.25	f_B = 278.20
Error	12	1671.00	139.25	
Total	19	171,119.75		

Since $f_A = 95.57 \geq F_{.05,4,12} = 3.26$, H_0 is rejected in favor of H_a. We conclude that power consumption does depend on the brand of humidifier. To identify significantly different brands, we use Tukey's procedure; $Q_{.05,5,12} = 4.51$ and $d = 4.51\sqrt{139.25/4} = 26.6$.

$$\begin{array}{ccccc} \bar{x}_{1.} & \bar{x}_{3.} & \bar{x}_{2.} & \bar{x}_{4.} & \bar{x}_{5.} \\ 797.50 & \underline{839.00} & \underline{843.50} & \underline{914.00} & \underline{937.25} \end{array}$$

The underscoring indicates that the brands can be divided into three groups with respect to power consumption.

Because the blocking factor is of secondary interest, $F_{.05,3,12}$ is not needed, though the computed value of F_B is clearly highly significant. Figure 11.7 shows SAS output for this data. Notice that in the first part of the ANOVA table, the sums of squares (SS's) for treatments (brands) and blocks (humidity levels) are combined into a single "model" SS.

```
                         Analysis of Variance Procedure

    Dependent Variable: POWERUSE

                              Sum of       Mean
     Source           DF      Squares      Square      F Value      Pr > F

     Model            7     169448.750   24206.964     173.84       0.0001

     Error           12       1671.000     139.250

     Corrected Total 19     171119.750

                        R-Square  C.V.       Root MSE     POWERUSE Mean

                        0.990235  1.362242   11.8004      866.25000

     Source           DF      Anova SS     Mean Square  F Value      Pr > F

     BRAND            4       53231.000   13307.750      95.57       0.0001

     HUMIDITY         3      116217.750   38739.250     278.20       0.0001

            Alpha = 0.05 df = 12 MSE = 139.25
            Critical Value of Studentized Range = 4.508
            Minimum Significant Difference = 26.597

    Means with the same letter are not significantly different.

      Tukey Grouping     Mean      N      BRAND
               A       937.250     4       5

               A

               A       914.000     4       4

               B       843.500     4       2

               B

               B       839.000     4       3

               C       797.500     4       1
```

Figure 11.7 SAS output for consumption data   ■

In some experimental situations in which treatments are to be applied to subjects, a single subject can receive all I of the treatments. Blocking is then often done on the subjects themselves to control for variability between subjects, typically in random order; each subject is then said to act as its own control. Social scientists sometimes refer to such experiments as **repeated-measures** designs. The "units" within a block are then the different "instances" of treatment application. Similarly, blocks are often taken as different time periods, locations, or observers.

In most randomized block experiments in which subjects serve as blocks, the subjects actually participating in the experiment are selected from a large population. The subjects then contribute random rather than fixed effects. This does not impact the procedure for comparing treatments when $K_{ij} = 1$ (one observation per "cell," as in this section), but the procedure is altered if $K_{ij} > 1$. We will shortly consider two-factor models in which effects are random.

More on Blocking When $I = 2$, either the F test above or the paired differences t test can be used to analyze the data. The resulting conclusion will not depend on which procedure is used, since $T^2 = F$ and $t^2_{\alpha/2,v} = F_{\alpha,1,v}$.

Just as with pairing, blocking entails both a potential gain and a potential loss in precision. If there is a great deal of heterogeneity in experimental units, the value of the variance parameter σ^2 in the one-way model will be large. The effect of blocking is to filter out the variation represented by σ^2 in the two-way model appropriate for a randomized block experiment. Other things being equal, a smaller value of σ^2 results in a test that is more likely to detect departures from H_0 (i.e., a test with greater power).

However, other things are not equal here, since the single-factor F test is based on $I(J - 1)$ degrees of freedom (df) for error, whereas the two-factor F test is based on $(I - 1)(J - 1)$ df for error. Fewer degrees of freedom for error results in a decrease in power, essentially because the denominator estimator of σ^2 is not as precise. This loss in degrees of freedom can be especially serious if the experimenter can afford only a small number of observations. Nevertheless, if it appears that blocking will significantly reduce variability, it is probably worth the loss in degrees of freedom.

Models for Random Effects

In many experiments, the actual levels of a factor used in the experiment, rather than being the only ones of interest to the experimenter, have been selected from a much larger population of possible levels of the factor. In a two-factor situation, when this is the case for both factors, a **random effects model** is appropriate. The case in which the levels of one factor are the only ones of interest while the levels of the other factor are selected from a population of levels leads to a **mixed effects model**. The two-factor random effects model when $K_{ij} = 1$ is

$$X_{ij} = \mu + A_i + B_j + \varepsilon_{ij} \quad (i = 1, \ldots, I; \quad j = 1, \ldots, J)$$

where the A_i's, B_j's, and ε_{ij}'s are all independent, normally distributed rvs with mean 0 and variances σ_A^2, σ_B^2, and σ^2, respectively.

The hypotheses of interest are then $H_{0A}: \sigma_A^2 = 0$ (level of factor A does not contribute to variation in the response) versus $H_{aA}: \sigma_A^2 > 0$ and $H_{0B}: \sigma_B^2 = 0$ versus $H_{aB}: \sigma_B^2 > 0$. Whereas $E(\text{MSE}) = \sigma^2$ as before, the expected mean squares for factors A and B are now

$$E(\text{MSA}) = \sigma^2 + J\sigma_A^2 \quad \text{and} \quad E(\text{MSB}) = \sigma^2 + I\sigma_B^2$$

Thus when H_{0A} (H_{0B}) is true, F_A (F_B) is still a ratio of two unbiased estimators of σ^2. It can be shown that a test with significance level α for H_{0A} versus H_{aA} still rejects H_{0A} if $f_A \geq F_{\alpha, I-1, (I-1)(J-1)}$, and, similarly, the same procedure as before is used to decide between H_{0B} and H_{aB}.

For the case in which factor A is fixed and factor B is random, the mixed model is

$$X_{ij} = \mu + \alpha_i + B_j + \varepsilon_{ij} \quad (i = 1, \ldots, I; \quad j = 1, \ldots, J)$$

where $\sum \alpha_i = 0$, and the B_j's and ε_{ij}'s are all independent, normally distributed rvs with mean 0 and variances σ_B^2 and σ^2, respectively. Now the two null hypotheses are

$$H_{0A}: \alpha_1 = \cdots = \alpha_I = 0 \quad \text{and} \quad H_{0B}: \sigma_B^2 = 0$$

Expected mean squares are

$$E(\text{MSA}) = \sigma^2 + \frac{J}{I-1}\sum \alpha_i^2 \quad \text{and} \quad E(\text{MSB}) = \sigma^2 + I\sigma_B^2$$

The test procedures for H_{0A} versus H_{aA} and H_{0B} versus H_{aB} are exactly as before. For example, in the analysis of the color change data in Example 11.13, if the four wash treatments were randomly selected, then because $f_B = 11.05 > F_{.05,3,6} = 4.76$, $H_{0B}: \sigma_B^2 = 0$ is rejected in favor of $H_{aB}: \sigma_B^2 > 0$.

Summarizing, when $K_{ij} = 1$, although the hypotheses and expected mean squares differ from the case of both effects fixed, the test procedures are identical.

Exercises: Section 11.4 (45–60)

45. The number of miles of useful tread wear (in 1000's) was determined for tires of each of five different makes of subcompact car (factor A, with $I = 5$) in combination with each of four different brands of radial tires (factor B, with $J = 4$), resulting in $IJ = 20$ observations. The values SSA = 30.6, SSB = 44.1, and SSE = 59.2 were then computed. Assume that an additive model is appropriate.

   a. Test H_0: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ (no differences in true average tire lifetime due to makes of cars) versus H_a: at least one $\alpha_i \neq 0$ using a level .05 test.
   b. Test H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (no differences in true average tire lifetime due to brands of tires) versus H_a: at least one $\beta_j \neq 0$ using a level .05 test.

46. Four different coatings are being considered for corrosion protection of metal pipe. The pipe will be buried in three different types of soil. To investigate whether the amount of corrosion depends either on the coating or on the type of soil, 12 pieces of pipe are selected. Each piece is coated with one of the four coatings and buried in one of the three types of soil for a fixed time, after which the amount of corrosion (depth of

maximum pits, in .0001 in.) is determined. The depths are shown in this table:

		Soil type (B)		
		1	2	3
	1	64	49	50
Coating (A)	2	53	51	48
	3	47	45	50
	4	51	43	52

   a. Assuming the validity of the additive model, carry out the ANOVA analysis using an ANOVA table to see whether the amount of corrosion depends on either the type of coating used or the type of soil. Use $\alpha = .05$.
   b. Compute $\hat{\mu}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$.

47. The article "Step-Counting Accuracy of Activity Monitors in Persons with Down Syndrome" (*J. Intellect. Disabil. Res.* 2019: 21–30) describes a study in which 17 people with DS walked for a set time period with multiple step-counting devices attached to them. (Walking is a common form of exercise for people with DS, and clinicians want to insure that step counts for them are accurate.) The accompanying table summarizes the different step-counting methods and the number of steps recorded for each

participant. (LFE refers to a low frequency extension filter applied to the device.)

Step-counting method	Mean	SD
Hand tally	668	70
Pedometer	557	202
Hip accelerometer	466	159
Hip accelerometer + LFE	606	93
Wrist accelerometer	449	89
Wrist accelerometer + LFE	579	85

Sums of squares consistent with this and other information in the article include SS(Method) = 596,748, SSE = 987,380, and SST = 2,113,228.

a. Determine the sum of squares associated with the blocking variable (subject), and then construct an ANOVA table.
b. Assuming that model assumptions are plausible, test the null hypothesis of "no method effect" at the .01 significance level.
c. Apply Tukey's procedure to these six step-counting methods. Are any of the five device-based methods *not* significantly different from hand tally (considered by the researchers to be the most correct)?

48. In an experiment to see whether the amount of coverage of light-blue interior latex paint depends either on the brand of paint or on the brand of roller used, 1 gallon of each of four brands of paint was applied using each of three brands of roller, resulting in the following data (number of square feet covered).

		Roller brand		
		1	2	3
	1	454	446	451
Paint	2	446	444	447
brand	3	439	442	444
	4	444	437	443

a. Construct the ANOVA table. [*Hint*: The computations can be expedited by subtracting 400 (or any other convenient number) from each observation. This does not affect the final results.]
b. Check the normality and constant variance assumptions graphically.

c. State and test hypotheses appropriate for deciding whether paint brand has any effect on coverage. Use $\alpha = .05$.
d. Repeat part (c) for brand of roller.
e. Use Tukey's method to identify significant differences among brands. Is there one brand that seems clearly preferable to the others?

49. The following data is presented in the article "Influence of Cutting Parameters on Drill Bit Temperature" (*Ind. Lubr. Tribol.* 2007: 186–193); values in the table are temperatures in °C.

		Feed rate		
		1	2	3
	1	275	325	365
Spindle speed	2	380	415	420
	3	425	420	405

a. Construct an ANOVA table from this data.
b. Test whether spindle speed impacts drill bit temperature at the .05 significance level.
c. Test whether feed rate impacts drill bit temperature at the .05 significance level.

50. A particular county employs three assessors who are responsible for determining the value of residential property in the county. To see whether these assessors differ systematically in their assessments, 5 houses are selected, and each assessor is asked to determine the market value of each house. With factor A denoting assessors ($I = 3$) and factor B denoting houses ($J = 5$), suppose SSA = 11.7, SSB = 113.5, and SSE = 25.6.

a. Test H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ at level .05. (H_0 states that there are no systematic differences among assessors.)
b. Explain why a randomized block experiment with only 5 houses was used rather than a one-way ANOVA experiment involving a total of 15 different houses with each assessor asked to assess 5 different houses (a different group of 5 for each assessor).

51. In a 2018 class activity, 54 students measured how much time (sec) it took to melt each of the following in their mouths: (1) a butterscotch chip, (2) a chocolate chip, (3) a white chip (yes, white is a chip flavor). Each student rolled a die to determine the order in which to melt the chips.

    a. Why was it important to randomize the chip order?
    b. Besides not having to recruit as many students to get the same number of observations, what is the advantage of using blocking here, versus randomly assigning one chip to each student?
    c. Summary quantities include $\bar{x}_{1.} = 88.15$, $\bar{x}_{2.} = 60.49$, $\bar{x}_{3.} = 72.35$, SS(Subject) = 135,833, and SSE = 31,506. Construct an ANOVA table and test at significance level .01 to see whether mean melting time varies by chip flavor.
    d. Judging from the F ratio for subjects, do you think that blocking on subjects was effective in this experiment? Explain.

52. The efficiency (%) of a 0.7 L Daihatsu diesel engine at 3100 rpm was determined at various torques (N-m) and coolant temperatures (°C), resulting in the following data kindly provided by the study's authors ("Performance of a Diesel Engine at High Coolant Temperatures, *J. Energy Resour. Technol.* 2017).

		Coolant Temp. (°C)					
		90	100	125	150	175	200
	12	24.07	23.81	23.55	22.84	24.34	24.62
	15	26.52	26.00	26.06	25.33	26.92	27.08
Torque	18	28.50	28.23	26.67	26.42	28.94	28.74
	21	28.61	29.96	28.38	27.16	29.35	29.16
	24	28.47	28.34	28.20	26.55	28.87	27.27

    a. Test at the .01 significance level whether mean engine efficiency differs with torque.
    b. Test at the .01 significance level whether mean engine efficiency differs with coolant temperature.
    c. Apply Tukey's procedure as appropriate to the results of (a) and (b).

53. An experiment was conducted to assess the effect of current and voltage on the tensile strength (ksi) of welds made using a tungsten inert gas (TIG) welding tool, which resulted in the following data.

		Voltage		
		10	12	14
	130	197.62	200.35	199.40
Current	135	185.90	215.56	179.36
	140	203.23	174.81	194.47

Use two-factor ANOVA to determine whether current or voltage impacts the tensile strength of welds under these experimental conditions at the .10 significance level. (Data is from "To Investigate the [E]ffect of Process Parameters on Mechanical Properties of TIG Welded 6351 Aluminum Alloy by ANOVA," GE-*Int. J. Engr. Res.* 2014: 50–62.)

54. The article "Effect of Face Value on Product Valuation in Foreign Currencies" (*J. Consum. Res.* 2002) describes a classroom experiment involving 97 business students to see whether they could adjust for exchange rates when deciding how much to spend a product. The students acted as buyers in a mock World Garment Expo at which each would be purchasing silk ties from six different nations. They were provided pictures of the ties and the exchange rates for each national currency into US$; the pictures were randomly permuted to reduce any perceived quality differences. Students then had to report how much, in the foreign currencies, they would pay for one silk tie. The average prices students were willing to pay, converted back into dollars, appear in the accompanying table; exchange rates are the number of foreign currency units (e.g., Norwegian krone or Japanese yen) equaling $1.

Country (Exch. rate)	Mean	SD
Norway (9.5)	$15.85	$15.36
Luxembourg (48)	$15.45	$13.47
Japan (110)	$15.91	$11.26
Korea (1100)	$12.42	$10.17
Romania (24,500)	$11.33	$12.05
Turkey (685,000)	$10.77	$9.01

Relevant sums of squares include SS(Country) = 2752, SSE = 45,086, and SST = 86,653.

a. What experimental design was used in this study? What was the advantage of using this method?

b. Construct an ANOVA table from the information provided, then test the null hypothesis of "no currency exchange effect" at the .01 level.

c. The exchange rates vary by orders of magnitude (this was deliberate). As the exchange rate increases, what happens to the average amount students are willing to pay for a silk tie in that currency? (The article's authors note that "th[is] evidence is against the common wisdom that when the home currency is perceived to go a long way in foreign currency terms, the foreign currency will be treated as play money and that people will overspend.")

55. The strength of concrete used in commercial construction tends to vary from one batch to another. Consequently, small test cylinders of concrete sampled from a batch are "cured" for periods up to about 28 days in temperature- and moisture-controlled environments before strength measurements are made. Concrete is then "bought and sold on the basis of strength test cylinders" (ASTM C 31 Standard Test Method for Making and Curing Concrete Test Specimens in the Field). The accompanying data resulted from an experiment carried out to compare three different curing methods with respect to compressive strength (MPa). Analyze this data.

Batch	Method A	Method B	Method C
1	30.7	33.7	30.5
2	29.1	30.6	32.6
3	30.0	32.2	30.5
4	31.9	34.6	33.5
5	30.5	33.0	32.4
6	26.9	29.3	27.8

(continued)

Batch	Method A	Method B	Method C
7	28.2	28.4	30.7
8	32.4	32.4	33.6
9	26.6	29.5	29.2
10	28.6	29.4	33.2

56. Check the normality and constant variance assumptions graphically for the data of Example 11.17.

57. Suppose that in the experiment described in Exercise 50 the five houses had actually been selected at random from among those of a certain age and size, so that factor B is random rather than fixed. Test $H_0 : \sigma_B^2 = 0$ versus $H_a : \sigma_B^2 > 0$ using a level .01 test.

58. a. Show that a constant d can be added to (or subtracted from) each x_{ij} without affecting any of the ANOVA sums of squares.

b. Suppose that each x_{ij} is multiplied by a nonzero constant c. How does this affect the ANOVA sums of squares? How does this affect the values of the F statistics F_A and F_B ? What effect does "coding" the data by $y_{ij} = cx_{ij} + d$ have on the conclusions resulting from the ANOVA procedures?

59. Use the fact that $E(X_{ij}) = \mu + \alpha_i + \beta_j$ with $\sum \alpha_i = \sum \beta_j = 0$ to show that $E(\overline{X}_{i.} - \overline{X}_{..}) = \alpha_i$, so that $\hat{\alpha}_i = \overline{X}_{i.} - \overline{X}_{..}$ is an unbiased estimator for α_i.

60. Power for the F test in two-factor ANOVA is calculated using a similar method to the one shown in Section 11.3. For fixed values of $\alpha_1, \alpha_2, \ldots, \alpha_I$, power calculations are based on the noncentral F distributions with parameters $v_1 = I - 1$, $v_2 = (I - 1)(J - 1)$, and noncentrality parameter $\lambda = J \sum \alpha_i^2 / \sigma^2$.

a. For the corrosion experiment described in Exercise 46, determine power when $\alpha_1 = 4$, $\alpha_2 = 0$, $\alpha_3 = \alpha_4 = -2$, and $\sigma = 4$. Repeat for $\alpha_1 = 6$, $\alpha_2 = 0$, $\alpha_3 = \alpha_4 = -3$, and $\sigma = 4$.

b. By symmetry, what is the power for the test of H_{0B} versus H_{aB} in Example 11.13 when $\beta_1 = .3$, $\beta_2 = \beta_3 = \beta_4 = -.1$, and $\sigma = .3$?

11.5 Two-Factor ANOVA with Replication

In Section 11.4, we analyzed data from a two-factor experiment in which there was one observation for each of the *IJ* combinations of levels of the two factors (i.e., no replication). To obtain valid test procedures in that situation, the μ_{ij}'s were assumed to have an *additive* structure, meaning that the difference in true average responses for any two levels of the factors is the same for each level of the other factor. This was shown in Figure 11.4a, in which the lines connecting true average responses are parallel.

Figure 11.4b depicted a set of true average responses that does not have additive structure. The lines connecting these μ_{ij}'s are not parallel, which means that the difference in true mean responses for different levels of one factor does depend on the level of the other factor—what's known as an *interaction*. When $K_{ij} > 1$ for at least one (i, j) pair, we can estimate the interaction effect and formally test for whether interaction is present.

In specifying the appropriate model and deriving test procedures, we will focus on the case $K_{ij} = K > 1$, so the number of observations per "cell" (i.e., for each combination of levels) is constant. That is, throughout this section we will assume a *balanced* study design.

A Two-Factor Model With Interaction

Again μ_{ij} will denote the true mean response when factor *A* is at level *i* and factor *B* is at level *j*. Expressions (11.7)–(11.9) show the development of a model equation that assumes additivity (i.e., no interaction). To extend this model, first let

$$\mu = \frac{1}{IJ} \sum_i \sum_j \mu_{ij} \quad \mu_{i\cdot} = \frac{1}{J} \sum_j \mu_{ij} \quad \mu_{\cdot j} = \frac{1}{I} \sum_i \mu_{ij} \tag{11.10}$$

Thus μ is the expected response averaged over all levels of both factors (the true grand mean), $\mu_{i\cdot}$ is the expected response averaged over levels of factor *B* when factor *A* is held at level *i*, and similarly for $\mu_{\cdot j}$. Now define three sets of parameters by

$$\begin{aligned}
\alpha_i &= \mu_{i\cdot} - \mu = \text{ the effect of factor } A \text{ at level } i \\
\beta_j &= \mu_{\cdot j} - \mu = \text{ the effect of factor } B \text{ at level } j \\
\gamma_{ij} &= \mu_{ij} - (\mu + \alpha_i + \beta_j)
\end{aligned} \tag{11.11}$$

from which

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

The α_i's and β_j's are the same as those from Section 11.4. The α_i's are called the **main effects for factor A**, and the β_j's are the **main effects for factor B**. The new parameters, the γ_{ij}'s, measure the difference between the true treatment means μ_{ij} and the means assumed under the additive model (11.8). The γ_{ij}'s are referred to as the **interaction parameters**, and the model is additive if and only if all γ_{ij}'s = 0.

Although there are I α_i's, J β_j's, and IJ γ_{ij}'s in addition to μ, the conditions $\sum \alpha_i = 0$, $\sum \beta_j = 0$, $\sum_j \gamma_{ij} = 0$ for every i, and $\sum_i \gamma_{ij} = 0$ for any j—all true by virtue of (11.10) and (11.11)—imply that only IJ of these new parameters are independently determined: μ, $I-1$ of the α_i's, $J-1$ of the β_j's, and $(I-1)(J-1)$ of the γ_{ij}'s.

We now must use triple subscripts for both random variables and observed values: X_{ijk} and x_{ijk} denote the kth observation (replication) when factor A is at level i and factor B is at level j.

TWO-FACTOR ANOVA GENERAL MODEL EQUATION

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$
$$i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K$$

(11.12)

where the ε_{ijk}'s are independent and normally distributed, each with mean 0 and variance σ^2.

Example 11.18 Three different varieties of tomato (Harvester, Ife No. 1, and Pusa Early Dwarf) and four different plant densities (10, 20, 30, and 40 thousand plants per hectare) are being considered for planting in a particular region. To see whether either variety or plant density affects yield, each combination of variety and plant density is used in three different plots, resulting in the data on yields in Table 11.10.

Table 11.10   Yield data for Example 11.18

Variety	10,000			20,000			30,000			40,000		
H	10.5	9.2	7.9	12.8	11.2	13.3	12.1	12.6	14.0	10.8	9.1	12.5
Ife	8.1	8.6	10.1	12.7	13.7	11.5	14.4	15.4	13.7	11.3	12.5	14.5
P	16.1	15.3	17.5	16.6	19.2	18.5	20.8	18.0	21.0	18.4	18.9	17.2

(with "Planting density" spanning the four density columns)

Here, $I = 3$, $J = 4$, and $K = 3$, for a total of $IJK = 36$ observations. If we identify factor A = variety and B = density, then the observations across the first row of Table 11.10 are $x_{111} = 10.5$, $x_{112} = 9.2$, $x_{113} = 7.9$, $x_{121} = 12.8$, and so on. Some of the parameters specified in the model equation (11.12) include

μ = true average yield of all tomato plants in this population
$\mu_{1.}$ = true average yield of all Harvester plants ($i = 1$) in this population
β_2 = the effect of 20,000 plants/hectare density ($j = 2$) on average yield

To check the normality and constant variance assumptions, we can make plots similar to those of Section 11.4. Define the predicted/fitted values to be the cell means, $\hat{x}_{ijk} = \bar{x}_{ij\cdot}$, so the residuals are $e_{ijk} = x_{ijk} - \hat{x}_{ijk} = x_{ijk} - \bar{x}_{ij\cdot}$. For example, the mean of the three observations in the top-left cell of Table 11.10 is $\bar{x}_{11\cdot} = (10.5 + 9.2 + 7.9)/3 = 9.2$, and the residual for the very first observation is $e_{111} = x_{111} - \bar{x}_{11\cdot} = 10.5 - 9.2 = 1.3$. The normal probability plot of the 36 residuals is Figure 11.8a, and the plot of the residuals against the fitted values is Figure 11.8b. The normal plot is sufficiently straight that there should be no concern about the normality assumption. The plot of residuals against predicted values has a fairly uniform vertical spread, so there is no cause for concern about the constant variance assumption.

a

Normal Probability Plot of the Residuals
(response is Yield)

b

Residuals Versus the Fitted Values
(response is Yield)
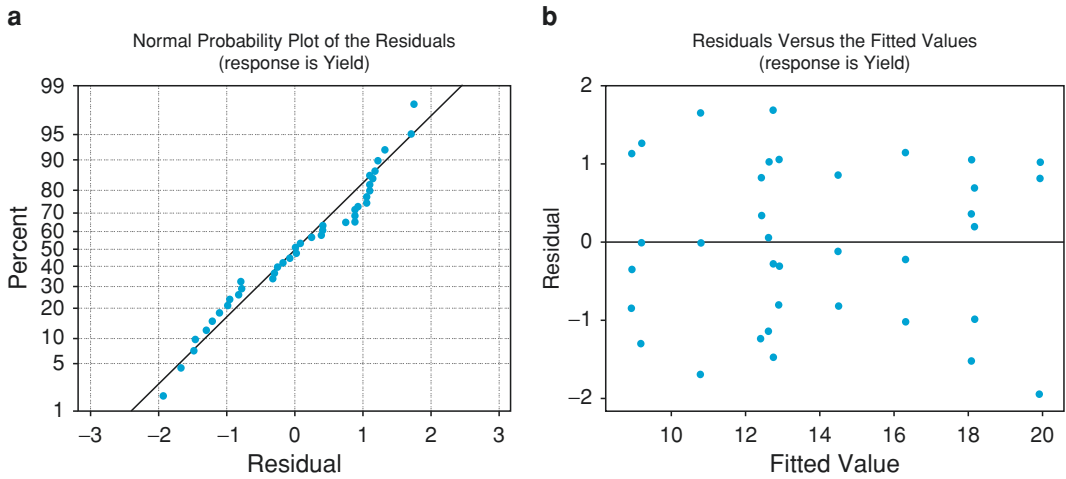
Figure 11.8 Plots from Minitab to verify assumptions for Example 11.18   ∎

Sums of Squares and Test Procedures

There are now three relevant pairs of hypotheses:

$$H_{0AB}: \quad \gamma_{ij} = 0 \text{ for all } i,j \quad \text{versus} \quad H_{aAB}: \text{at least one } \gamma_{ij} \neq 0$$

$$H_{0A}: \quad \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0 \quad \text{versus} \quad H_{aA}: \text{at least one } \alpha_i \neq 0$$

$$H_{0B}: \quad \beta_1 = \beta_2 = \cdots = \beta_J = 0 \quad \text{versus} \quad H_{aB}: \text{at least one } \beta_j \neq 0$$

The no-interaction hypothesis H_{0AB} is usually tested first. If H_{0AB} is *not* rejected, then the other two hypotheses can be tested to see whether the main effects are significant. But once H_{0AB} is rejected, we believe that the effect of factor A at any particular level depends on the level of B (and vice versa). It then does not make sense to test H_{0A} or H_{0B}. In this case, an interaction plot similar to that of Figure 11.4b is helpful in visualizing the way the factors interact.

To test the hypotheses of interest, we again define sums of squares and indicate their corresponding degrees of freedom. Again, a dot in place of a subscript means that we have summed over all values of that subscript, and a horizontal bar denotes averaging. So, for example, $\overline{X}_{ij\cdot}$ denotes the mean of the K observations in the (i, j)th cell of the data table, while $\overline{X}_{i\cdot\cdot}$ represents the average of all JK values in the ith row.

$$SSA = \sum_i \sum_j \sum_k (\overline{X}_{i\cdot\cdot} - \overline{X}_{\cdots})^2 \qquad\qquad \text{df} = I - 1$$

$$SSB = \sum_i \sum_j \sum_k (\overline{X}_{\cdot j\cdot} - \overline{X}_{\cdots})^2 \qquad\qquad \text{df} = J - 1$$

$$SSAB = \sum_i \sum_j \sum_k (\overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} + \overline{X}_{\cdots})^2 \quad \text{df} = (I-1)(J-1)$$

$$SSE = \sum_i \sum_j \sum_k (X_{ijk} - \overline{X}_{ij\cdot})^2 \qquad\qquad \text{df} = IJK - IJ$$

$$SST = \sum_i \sum_j \sum_k (X_{ijk} - \overline{X}_{\cdots})^2 \qquad\qquad \text{df} = IJK - 1$$

SSAB is called the **interaction sum of squares**. Mean squares are, as always, defined by (sum of squares)/df.

The fundamental ANOVA identity is

$$SST = SSA + SSB + SSAB + SSE$$

According to the fundamental identity, variation is partitioned into four pieces: unexplained (SSE—which would be present whether or not any of the three null hypotheses was true) and three pieces that may be explained by the truth or falsity of the three H_0's.

The expected mean squares suggest how each set of hypotheses should be tested using the appropriate ratio of mean squares with MSE in the denominator:

$$E(\text{MSE}) = \sigma^2$$

$$E(\text{MSA}) = \sigma^2 + \frac{JK}{I-1} \sum_{i=1}^{I} \alpha_i^2$$

$$E(\text{MSB}) = \sigma^2 + \frac{IK}{J-1} \sum_{j=1}^{J} \beta_j^2$$

$$E(\text{MSAB}) = \sigma^2 + \frac{K}{(I-1)(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \gamma_{ij}^2$$

Each of the three mean-square ratios can be shown to have an F distribution when the associated H_0 is true, which yields the following level α test procedures.

Hypotheses	Test Statistic Value	Rejection Region
H_{0A} versus H_{aA}	$f_A = MSA/MSE$	$f_A \geq F_{\alpha, I-1, IJ(K-1)}$
H_{0B} versus H_{aB}	$f_A = MSB/MSE$	$f_B \geq F_{\alpha, J-1, IJ(K-1)}$
H_{0AB} versus H_{aAB}	$f_{AB} = MSAB/MSE$	$f_{AB} \geq F_{\alpha, (I-1)(J-1), IJ(K-1)}$

As before, the results of the analysis are summarized in an ANOVA table.

Example 11.19 (Example 11.18 continued) The cell, row, column, and grand means for the given data are

	10,000	20,000	30,000	40,000	$\bar{x}_{i..}$
H	9.20	12.43	12.90	10.80	11.33
Ife	8.93	12.63	14.50	12.77	12.21
P	16.30	18.10	19.93	18.17	18.13
$\bar{x}_{.j.}$	11.48	14.39	15.78	13.91	$\bar{x}_{...} = 13.89$

Table 11.11 summarizes the resulting ANOVA computations.

Table 11.11 ANOVA table for Example 11.19

Source of variation	df	Sum of squares	Mean square	f		P-value
Varieties	2	327.60	163.80	f_A	= 103.02	<.0001
Density	3	86.69	28.90	f_B	= 18.18	<.0001
Interaction	6	8.03	1.34	f_{AB} =	.84	.551
Error	24	38.04	1.59			
Total	35	460.36				

Since $f_{AB} = .84 < F_{.01,6,24} = 3.67$, H_{0AB} cannot be rejected at level .01, so we conclude that the interaction effects are not significant. Now the presence or absence of main effects can be investigated. Since $f_A = 103.02 \geq F_{.01,2,24} = 5.61$, H_{0A} is rejected at level .01 in favor of the conclusion that different varieties do affect the true average yields. Similarly, $f_B = 18.18 \geq 4.72 = F_{.01,3,24}$, so we conclude that true average yield also depends on plant density.

Figure 11.9 shows the interaction plot. Notice the nearly parallel lines for the three tomato varieties, in agreement with the F test showing no significant interaction. The yield for Pusa Early Dwarf appears to be significantly above the yields for the other two varieties, and this is in accord with the highly significant F for varieties. Furthermore, all three varieties show the same pattern in which yield increases as the density goes up, but decreases beyond 30,000 per hectare. This suggests that planting more seed will increase the yield, but eventually overcrowding causes the yield to drop.
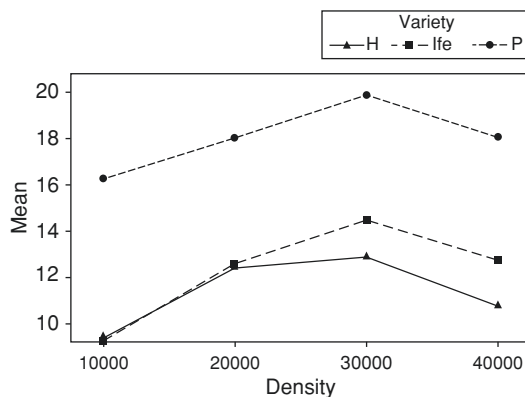
Figure 11.9 Interaction plot for the tomato yield data   ■

Multiple Comparisons

When the no-interaction hypothesis H_{0AB} is not rejected and at least one of the two main-effect null hypotheses is rejected, Tukey's method can be used to identify significant differences in levels. To identify differences among the α_i's when H_{0A} is rejected:

1. Obtain $Q_{\alpha,I,IJ(K-1)}$; the second subscript I identifies the number of levels being compared and the third subscript refers to the error df.

2. Compute $d = Q \cdot \sqrt{MSE/JK}$; JK is the number of observations averaged to obtain each of the $\bar{x}_{i..}$'s to be compared in step 3.
3. Order the $\bar{x}_{i..}$'s from smallest to largest and, as before, underscore all pairs that differ by less than d. Pairs not underscored correspond to significantly different levels of factor A.

To identify different levels of factor B when H_{0B} is rejected, replace the second subscript in Q by J, replace JK by IK in d, and replace $\bar{x}_{i..}$ by $\bar{x}_{.j.}$.

Example 11.20 (Example 11.19 continued) For factor A (varieties), $I = 3$, so with $\alpha = .01$ and $IJ(K - 1) = 24$, $Q_{.01,3,24} = 4.55$. Then $d = 4.55\sqrt{1.59/12} = 1.66$, so ordering and underscoring give

$$
\begin{array}{ccc}
\bar{x}_{1..} & \bar{x}_{2..} & \bar{x}_{3..} \\
11.33 & 12.21 & 18.13
\end{array}
$$

The Harvester and Ife varieties do not differ significantly from each other in effect on true average yield, but both differ from the Pusa variety.

For factor B (density), $J = 4$ so $Q_{.01,4,24} = 4.91$ and $d = 4.91\sqrt{1.59/9} = 2.06$.

$$
\begin{array}{cccc}
\bar{x}_{.1.} & \bar{x}_{.4.} & \bar{x}_{.2.} & \bar{x}_{.3.} \\
11.48 & 13.91 & 14.39 & 15.78
\end{array}
$$

Thus with experimentwise error rate .01, only the lowest density differs significantly from all others. Even with $\alpha = .05$ (so that $d = 1.64$), densities 2 and 3 cannot be judged significantly different from each other in their effect on yield. ∎

Models with Mixed and Random Effects

In some situations, the levels of either factor may have been chosen from a large population of possible levels, so that the effects contributed by the factor are random rather than fixed. As in Section 11.4, if both factors contribute random effects, the model is referred to as a random effects model, whereas if one factor is fixed and the other is random, a mixed effects model results. We will now consider the analysis for a mixed effects model in which factor A (rows) is the fixed factor and factor B (columns) is the random factor; the case in which both factors are random is dealt with in Exercise 73. When either factor is random, interaction effects will also be random, and the mixed effects model is

$$
X_{ij} = \mu + \alpha_i + B_j + G_{ij} + \varepsilon_{ijk}
$$
$$
i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K
$$

Here μ and the α_i's are constants with $\sum \alpha_i = 0$ and the B_j's, G_{ij}'s, and ε_{ijk}'s are independent, normally distributed random variables with expected value 0 and variances σ_B^2, σ_G^2, and σ^2, respectively.[2] The three hypotheses of interest are

$$
\begin{aligned}
&H_{0A}: &&\alpha_1 = \cdots = \alpha_I = 0 &&\text{versus} &&H_{aA}: \text{at least one } \alpha_i \neq 0 \\
&H_{0B}: &&\sigma_B^2 = 0 &&\text{versus} &&H_{aB}: \sigma_B^2 > 0 \\
&H_{0G}: &&\sigma_G^2 = 0 &&\text{versus} &&H_{aG}: \sigma_G^2 > 0
\end{aligned}
$$

It is customary to test H_{0A} and H_{0B} only if the no-interaction hypothesis H_{0G} cannot be rejected.

The relevant sums of squares and mean squares needed for the test procedures are defined and computed exactly as in the fixed effects case. The expected mean squares for the mixed model are

$$
\begin{aligned}
E(\text{MSE}) &= \sigma^2 \\
E(\text{MSA}) &= \sigma^2 + K\sigma_G^2 + \frac{JK}{I-1}\sum \alpha_i^2 \\
E(\text{MSB}) &= \sigma^2 + K\sigma_G^2 + IK\sigma_B^2 \\
E(\text{MSAB}) &= \sigma^2 + K\sigma_G^2
\end{aligned}
$$

Thus, to test the no-interaction hypothesis, the ratio $f_{AB} = \text{MSAB/MSE}$ is again appropriate, with H_{0G} rejected if $f_{AB} \geq F_{\alpha,(I-1)(J-1),IJ(K-1)}$. However, for testing H_{0A} versus H_{aA}, the expected mean squares suggest that although the numerator of the F ratio should still be MSA, the denominator should be MSAB rather than MSE. MSAB is also the denominator of the F ratio for testing H_{0B}.

For testing H_{0A} versus H_{aA} (factor A fixed, B random), the test statistic value is $f_A = \text{MSA/MSAB}$, and the rejection region is $f_A \geq F_{\alpha,I-1,(I-1)(J-1)}$. The test of H_{0B} versus H_{aB} utilizes $f_B = \text{MSB/MSAB}$, and the rejection region is $f_B \geq F_{\alpha,J-1,(I-1)(J-1)}$.

Example 11.21 A process engineer has identified two potential causes of electric motor vibration, the material used for the motor casing (factor A) and the supply source of bearings used in the motor (factor B). The accompanying data on the amount of vibration (microns) resulted from an experiment in which motors with casings made of steel, aluminum, and plastic were constructed using bearings supplied by five randomly selected sources.

Material	Supply source									
	1		2		3		4		5	
Steel	13.1	13.2	16.3	15.8	13.7	14.3	15.7	15.8	13.5	12.5
Aluminum	15.0	14.8	15.7	16.4	13.9	14.3	13.7	14.2	13.4	13.8
Plastic	14.0	14.3	17.2	16.7	12.4	12.3	14.4	13.9	13.2	13.1

---

[2]This is referred to as an "unrestricted" model. An alternative "restricted" model requires that $\sum_i G_{ij} = 0$, so the G_{ij}'s are no longer independent. Expected mean squares and F ratios appropriate for testing certain hypotheses depend on the choice of model.

Only the three casing materials used in the experiment are under consideration for use in production, so factor A is fixed. However, the five supply sources were randomly selected from a much larger population, so factor B is random. The relevant null hypotheses are

$$H_{0A}: \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad H_{0B}: \sigma_B^2 = 0 \quad H_{0G}: \sigma_G^2 = 0$$

Minitab output appears in Figure 11.10. Notice that $f_A = 0.24 = 0.3523/1.4507 = MSA/MSAB$, not MSA/MSE as in a test with all fixed effects.

Factor Information

Factor	Type	Levels	Values
Material	Fixed	3	Aluminum, Plastic, Steel
Supplier	Random	5	1, 2, 3, 4, 5

Analysis of Variance

Source	DF	SS	MS	F-Value	P-Value
Material	2	0.7047	0.3523	0.24	0.790
Supplier	4	36.6747	9.1687	6.32	0.013
Material*Supplier	8	11.6053	1.4507	13.03	0.000
Error	15	1.6700	0.1113		
Total	29	50.6547			

Figure 11.10   Minitab output for the data of Example 11.21

The included 0.000 P-value for interaction means that it is less than .0005 (the actual value is .000018). To interpret the significant interaction we use the interaction plot, Figure 11.11, which has both versions, one with source on the x-axis and one with material on the x-axis. Interaction is evident, because the best material (the one with the least vibration) depends strongly on source. For source 1 the best material is steel, for source 3 the best material is plastic, and for source 4 the best material is aluminum. Because of this interaction, we ordinarily would not interpret the main effects, but one cannot help noticing that there is strong dependence of vibration on source. Source 2 is bad for all three materials and source 3 is pretty good for all three materials. When one-way ANOVA analyses are done to compare the five sources for each of the three materials, all three show highly significant differences. This is consistent with the P-value of 0.013 for supplier in Figure 11.10. We can conclude that, although the interaction causes the best material to depend on the supply source, the source also makes a difference of its own.

Interaction Plot(data means)for vibration
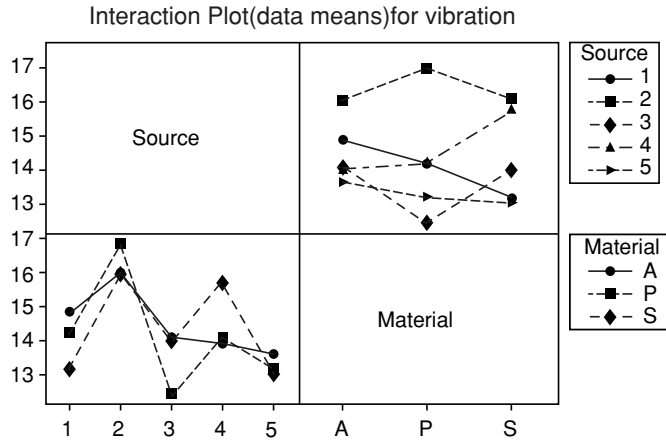
Figure 11.11 Minitab interaction plot for the data of Example 11.21   ∎

Final Comments on Two-Factor ANOVA

Power and sample size calculations for two-factor ANOVA are even more unwieldy than those for the single-factor case—software is essential for such computations. The pwr2 package in R includes graphical and computational tools for balanced two-way designs, and PROC GLMPOWER in SAS has similar functionality.

When at least two of the K_{ij}'s are unequal, the ANOVA computations are much more complex than for the case $K_{ij} = K$, and there are no nice formulas for the appropriate test statistics. Most software packages analyze unbalanced data by using a broader framework called the *general linear model*, which also encompasses the methods of Chapter 12. The references by Kutner et al., Miller, Montgomery, or Ott and Longnecker in the bibliography can be consulted for more information.

Exercises: Section 11.5 (61–73)

61. In an experiment to assess the effects of curing time (factor A) and type of mix (factor B) on the compressive strength of hardened cement cubes, three different curing times were used in combination with four different mixes, with three observations obtained for each of the 12 curing time–mix combinations. The resulting sums of squares were computed to be SSA = 30,763.0, SSB = 34,185.6, SSE = 97,436.8, and SST = 205,966.6.

    a. Construct an ANOVA table.
    b. Test at level .05 the null hypothesis H_{0AB}: all γ_{ij}'s = 0 (no interaction of factors) against H_{0AB}: at least one $\gamma_{ij} \neq 0$.
    c. Test at level .05 the null hypothesis H_{0A}: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ (factor A main effects are absent) against H_{aA}: at least one $\alpha_i \neq 0$.

    d. Test H_{0B}: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ versus H_{aB}: at least one $\beta_j \neq 0$ using a level .05 test.
    e. The values of the $\bar{x}_{i..}$'s were $\bar{x}_{1..} = 4010.88$, $\bar{x}_{2..} = 4029.10$, and $\bar{x}_{3..} = 3960.02$. Use Tukey's procedure to investigate significant differences among the three curing times.

62. In an experiment described in the article "The Impact of SMS Advertising on Members of a Virtual Community" (*J. Advert. Res.* 2008: 363–374), researchers worked with an online gaming forum to send out messages advertising a deal on Subway sandwiches. The message was varied in two ways: the apparent spokesperson (either Subway or "Nik," a made-up forum member) and the language

used (full English or texting shorthand). Forum members were randomly assigned to receive one of the four possible messages (spokesperson-language pairs); recipients were later surveyed about their views on the advertisement (attractiveness, credibility, etc.). Assume the researchers obtained $K = 12$ survey responses for each message (which approximates the actual study results).

a. Credibility ratings were translated into normalized scores for each participant, with positive values meaning the message was more credible (zero neutral, negative less credible). Mean normalized credibility scores appear below.

Credibility	Full English	Texting shorthand
Subway	.2729	.3465
Nik	−.4076	−.1689

Construct an interaction plot, and describe what you see.

b. With A = spokesperson and B = language, sums of squares include SSA = 4.2905, SSB = 0.2926, SSAB = 0.0818, and SSE = 36.5930. Perform a complete two-factor ANOVA, testing each of the three possible effects hypotheses at the .05 level. Explain what you discover.

c. Researchers also the measured "purchase intention" of each subject, with higher numbers indicating a greater likelihood of buying a Subway sandwich. Use the information below to create an interaction plot and perform a two-factor ANOVA as in parts (a)–(b). Again, explain your findings.

Purchase intention	Full English	Texting shorthand
Subway	3.949	5.417
Nik	4.389	4.056

SSA = 2.545   SSB = 3.865
SSAB = 9.731   SSE = 112.409

63. The accompanying data resulted from an experiment to investigate whether yield from a chemical process depended either on the formulation of a particular input or on mixer speed.

		Speed		
		60	70	80
Formulation	1	189.7	185.1	189.0
		188.6	179.4	193.0
		190.1	177.3	191.1
	2	165.1	161.7	163.3
		165.9	159.8	166.6
		167.6	161.6	170.3

A statistical computer package gave SS(Form) = 2253.44, SS(Speed) = 230.81, SS(Form*Speed) = 18.58, and SSE = 71.87.

a. Does there appear to be interaction between the factors?

b. Does yield appear to depend on either formulation or speed?

c. Calculate estimates of the main effects.

d. Verify that the residuals are 0.23, −0.87, 0.63, 4.50, −1.20, −3.30, −2.03, 1.97, 0.07, −1.10, −0.30, 1.40, 0.67, −1.23, 0.57, −3.43, −0.13, 3.57.

e. Construct a normal plot from the residuals given in part (d). Do the ε_{ijk}'s appear to be normally distributed?

f. Plot the residuals against the predicted values (cell means) to see if the population variance appears reasonably constant.

64. Artificial human joints are usually secured with acrylic bone cement. The following data on the force (in Newtons) required to break an acrylic cement bond under different temperatures and media is consistent with the information in the article "Validation of Small-Punch Test as a Technique for Characterizing the Mechanical Properties of Acrylic Bone Cement" (*J. Engr. Med.* 2006: 11–21):

Temp. (°C)	Medium	Breaking force data (N)
22	Dry	100.8, 141.9, 194.8, 118.4, 176.1, 213.1
37	Dry	302.1, 339.2, 288.8, 306.8, 305.2, 327.5
22	Wet	385.3, 368.3, 322.6, 307.4, 357.9, 321.4
37	Wet	363.5, 377.7, 327.7, 331.9, 338.1, 394.6

a. Identify the factors, levels, and treatments in this experiment.

b. Create an interaction plot using the breaking force data, and comment on what you see.

c. Test for the presence of main and interaction effects at the $\alpha = .05$ level. Are the results consistent with the interaction plot?

65. Students Philip Hurst and Yuan Chao Jiang investigated the accuracy of three different brands of .22 caliber ammunition at two different distances. The brands were Remington, Winchester, and Federal, and the two designated distances were 25 and 50 yards. "Accuracy" was measured by distance from the bull's-eye, in centimeters. All bullets were shot from the same rifle, and the order of the bullets was randomized. The accompanying table shows the mean accuracy at each combination based on $K = 75$ bullets.

		Bullet brand		
		Fed.	Rem.	Win.
Firing	25	3.027	3.360	3.280
distance	50	5.440	5.413	5.560

a. Identify the factors, levels, and treatments in this experiment.

b. Create an interaction plot for the accuracy data, and comment on what you see.

c. From software, SS(Dist.) = 568.97, SS(Brand) = 2.97, SS(Dist.*Brand) = 2.48, and SSE = 1041.49. Test for the

presence of main and interaction effects for the accuracy data at the $\alpha = .05$ level. Are the results consistent with your interaction plot?

66. In a study reported in the article "Can 'Low-Fat' Nutrition Labels Lead to Obesity?" (*J. Market. Res.* 2006: 605–617), students and parents attending a university open house were offered one of two candy bowls: one labeled "New Colors of Regular M&M's" or another labeled "New 'Low-Fat' M&M's." (The latter product does not really exist; the candies were just regular M&M's.) The researchers asked each person to fill out a questionnaire (including height and weight information) and recorded how much candy s/he took. The two factors of interest are A = how the candy was labeled (regular, low-fat) and B = the person's weight status (defined as "normal weight" for a body mass index below 25, "overweight" otherwise). The response variable used for the analysis was the amount of calories in the M&M's taken by the subject.

a. The accompanying table shows the average calorie consumption for each "treatment." Construct an interaction plot, and describe what you discover.

		Subject's weight	
		Normal	Overweight
Food	Regular	189	192
label	Low-fat	219	281

b. The article includes the following F-values and P-values for the various effects (error and total df do not follow our formulas because the study design was not balanced, but this does not affect the interpretation of the results). Are the results of the F tests consistent with the interaction plot? Explain.

Source of variation	df	F	P-value
Food label	1	13.1	0.000
Weight	1	4.3	0.039
Interaction	1	3.9	0.049
Error	251		
Total	254		

Source of variation	df	F	P-value
CBL Method	1	0.50	.291
Prior Knowledge	1	33.73	.000
Interaction	1	1.60	.105
Error	89		
Total	92		

67. A study was carried out to compare the writing lifetimes of four premium brands of pens. It was thought that the writing surface might affect lifetime, so three different surfaces were randomly selected. A writing machine was used to ensure that conditions were otherwise homogeneous (e.g., constant pressure and a fixed angle). The accompanying table shows the two lifetimes (min) obtained for each brand–surface combination.

Brand of pen	Writing surface		
	1	2	3
1	709, 659	713, 726	660, 645
2	668, 685	722, 740	692, 720
3	659, 685	666, 684	678, 750
4	698, 650	704, 666	686, 733

Carry out an appropriate ANOVA, and state your conclusions.

68. A 2005 article in *Issues in Accounting Education* described an experiment in which students in an introductory accounting course were randomly assigned to one of two computer-based learning (CBL) methods: one based on problem-solving and another using worked examples. Students in each group were also classified according to prior accounting knowledge (yes or no). A 15-point diagnostic exam was then administered to all students; average scores for the four groups, as well as a partial F table, appear below.

CBL method	Prior accounting knowledge?	
	Yes	No
Problem-solving	10.45	7.59
Worked examples	10.17	8.32

a. Create an interaction plot, and comment on what you see.
b. State the formal hypotheses being tested in the ANOVA table (there are three sets of hypotheses), and test each at the $\alpha = .05$ level. Assume the conditions required for this inference procedure are met.
c. Explain in practical terms what the tests in part (b) say about the effects of different computer-based learning methods and/or prior accounting knowledge on accounting diagnostic exam performance.

69. Several factors can impact the structural soundness of 3D-printed objects, including the "struts" that connect various pieces. The following data appears in the article "Analyzing the Effects of Temperature, Nozzle-Bed Distance, and Their Interactions on the Width of Fused Deposition Modeled Struts Using Statistical Techniques Toward Precision Scaffold Fabrication"(*J. Manuf. Sci. Engr.* 2017). Response values are strut widths, in microns.

Temp.	Nozzle-bed distance								
	0.2 mm			0.3 mm			0.4 mm		
180 °C	845	850	885	600	605	625	495	495	525
200 °C	770	800	850	650	690	690	490	520	525
220 °C	900	910	995	630	645	655	510	545	560

Perform a complete two-factor ANOVA, and report your findings.

70. The article "Is It Really Good to Talk? Testing the Impact of Providing Concurrent Verbal Protocols on Driving Performance" (*Ergonomics* 2017: 770–779) reported an experiment in which 20 drivers drove four

laps on a fixed course. During two of the laps, drivers remained silent; on the other two, drivers were instructed to "think aloud" about their driving as they proceeded along the course. Lap order was randomized for each driver, and each driver's average speed throughout the lap was recorded.

With A = protocol (silent or thinking aloud) and B = driver, sums of squares consistent with information in the article include SSA = 6.272, SSB = 343.975, SSAB = 46.733, and SSE = 138.571. Protocol is a fixed factor, while driver is a random factor. Perform the appropriate two-factor ANOVA (complete with ANOVA table), testing each effect at the .05 significance level. Explain what each F test tells you.

71. a. Show that $E(\overline{X}_{i\cdot\cdot} - \overline{X}_{\cdots}) = \alpha_i$, so that $\overline{X}_{i\cdot\cdot} - \overline{X}_{\cdots}$ is an unbiased estimator for α_i in the fixed effects model.
    b. With $\hat{\gamma}_{ij} = \overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} + \overline{X}_{\cdots}$, show that $\hat{\gamma}_{ij}$ is an unbiased estimator for γ_{ij} in the fixed effects model.

72. Refer back to the previous exercise. Show how a $100(1 - \alpha)\%$ t CI for $\alpha_i - \alpha_{i'}$ can be obtained. Then compute a 95% interval for $\alpha_2 - \alpha_3$ using the data from Example 11.18. [*Hint:* With $\theta = \alpha_2 - \alpha_3$, the result of the previous exercise indicates how to obtain $\hat{\theta}$. Then compute $V(\hat{\theta})$ and $\sigma_{\hat{\theta}}$ and obtain an estimate of $\sigma_{\hat{\theta}}$ by using $\sqrt{\text{MSE}}$ to estimate σ, which identifies the appropriate number of df.]

73. When both factors are random in a two-way ANOVA experiment with K replications per combination of factor levels, the expected mean squares are $E(\text{MSE}) = \sigma^2$, $E(\text{MSA}) = \sigma^2 + K\sigma_G^2 + JK\sigma_A^2$, $E(\text{MSB}) = \sigma^2 + K\sigma_G^2 + IK\sigma_B^2$, and $E(\text{MSAB}) = \sigma^2 + K\sigma_G^2$.

a. What F ratio is appropriate for testing $H_{0G}: \sigma_G^2 = 0$ versus $H_{aG}: \sigma_G^2 > 0$?
b. What F ratio is appropriate for testing $H_{0A}: \sigma_A^2 = 0$ versus $H_{aA}: \sigma_A^2 > 0$? Testing $H_{0B}: \sigma_B^2 = 0$ versus $H_{aB}: \sigma_B^2 > 0$?

Supplementary Exercises: (74–84)

74. Consider the following summary data on the modulus of elasticity ($\times 10^6$ psi) for lumber of three different grades (in close agreement with values in the article "Bending Strength and Stiffness of Second-Growth Douglas-Fir Dimension Lumber" (*For. Prod. J.* 1991: 35–43), except that the sample sizes there were larger):

Grade	J	$\overline{x}_{i\cdot}$	s_i
1	10	1.63	.27
2	10	1.56	.24
3	10	1.42	.26

Use this data and a significance level of .01 to test the null hypothesis of no difference in mean modulus of elasticity for the three grades.

75. The article "The Effects of a Pneumatic Stool and a One-Legged Stool on Lower Limb Joint Load and Muscular Activity During Sitting and Rising" (*Ergonomics* 1993: 519–535) gives the accompanying data on the effort required of a subject to arise from four different types of stools (Borg scale). Perform an analysis of variance using $\alpha = .05$, and follow this with a multiple comparisons analysis if appropriate.

		Subject								
		1	2	3	4	5	6	7	8	9	$\overline{x}_{i\cdot}$
Type of stool	1	12	10	7	7	8	9	8	7	9	8.56
	2	15	14	14	11	11	11	12	11	13	12.44
	3	12	13	13	10	8	11	12	8	10	10.78
	4	10	12	9	9	7	10	11	7	8	9.22

76. The article "Antimicrobial Activities of Essential Oil of Eight Plant Species from Different Families Against Some Pathogenic Microorganisms" (*Res. J. Microbiol.* 2016: 28–34) reported on an experiment in which various concentrations of eight essential oils were applied to active bacterial cultures. The accompanying table shows the inhibition percentage of *E. coli* for each oil-concentration combination.

Oil Type	Concentration (µL/mL)				
	2	4	6	8	10
Ginger	60	72	80	92	99
Thyme	58	64	78	86	96
Coriander	57	63	81	89	99
Marjoram	17	34	49	51	67
Mustard	14	31	45	63	70
Chamomile	22	36	54	63	72
Licorice	10	14	23	28	33
Nigella	15	29	42	48	57

   a. Perform a two-factor ANOVA, testing both main effects at the .01 level.
   b. Apply Tukey's method to the eight essential oils, and describe what you find.

77. An experiment was carried out to compare flow rates for four different types of nozzle.
   a. Sample sizes were 5, 6, 7, and 6, respectively, and calculations gave $f = 3.68$. State and test the relevant hypotheses using $\alpha = .01$.
   b. Analysis of the data using a statistical computer package yielded *P*-value = .029. At level .01, what would you conclude, and why?

78. The article "Towards Improving the Properties of Plaster Moulds and Castings" (*J. Engr. Manuf.* 1991: 265–269) describes several ANOVAs carried out to study how the amount of carbon fiber and sand additions affect various characteristics of the molding process. Here we give data on casting hardness and on wet-mold strength.

Sand addition (%)	Carbon fiber addition (%)	Casting hardness	Wet-mold strength
0	0	61.0	34.0
0	0	63.0	16.0
15	0	67.0	36.0
15	0	69.0	19.0
30	0	65.0	28.0
30	0	74.0	17.0
0	.25	69.0	49.0
0	.25	69.0	48.0
15	.25	69.0	43.0
15	.25	74.0	29.0
30	.25	74.0	31.0
30	.25	72.0	24.0
0	.50	67.0	55.0
0	.50	69.0	60.0
15	.50	69.0	45.0
15	.50	74.0	43.0
30	.50	74.0	22.0
30	.50	74.0	48.0

   a. An ANOVA for wet-mold strength gives SS(Sand) = 705, SS(Fiber) = 1278, SSE = 843, and SST = 3105. Test for the presence of any effects using $\alpha = .05$.
   b. Carry out an ANOVA on the casting hardness observations using $\alpha = .05$.
   c. Construct an interaction plot with sand percentage on the horizontal axis, and discuss the results of part (b) in terms of what the plot shows.

79. The article "The Effectiveness of Virtual and Augmented Reality in Health Science and Medical Anatomy" (*Anatom. Sc. Educ.* 2017: 549–559) describes an experiment in which 59 health science students received an identical, electronic 10-minute lesson on skull anatomy (complete with 3D graphics models) through one of three devices: a virtual reality (VR) system, an augmented reality (AR) system, or a 3D-capable tablet. After the lesson, all students took a 20-question test on skull anatomy. The accompanying table summarizes the students' exam scores.

Lesson delivery	n	Mean	SD
VR	20	12.9	4.3
AR	17	12.5	4.5
3D Tablet	22	13.3	4.2

Does the data suggest that there is a difference among the three lesson delivery methods with respect to true mean exam score? Use $\alpha = .05$.

80. Numerous factors contribute to the smooth running of an electric motor ("Increasing Market Share Through Improved Product and Process Design: An Experimental Approach," *Qual. Engr.* 1991: 361–369). In particular, it is desirable to keep motor noise and vibration to a minimum. To study the effect that the brand of bearing has on motor vibration, five different motor bearing brands were examined by installing each type of bearing on different random samples of six motors. The amount of motor vibration (measured in microns) was recorded when each of the 30 motors was running. The data for this study follows. State and test the relevant hypotheses at significance level .05, and then carry out a multiple comparisons analysis if appropriate.

							Mean
Brand 1:	13.1	15.0	14.0	14.4	14.0	11.6	13.68
Brand 2:	16.3	15.7	17.2	14.9	14.4	17.2	15.95
Brand 3:	13.7	13.9	12.4	13.8	14.9	13.3	13.67
Brand 4:	15.7	13.7	14.4	16.0	13.9	14.7	14.73
Brand 5:	13.5	13.4	13.2	12.7	13.4	12.3	13.08

81. An article in the British scientific journal *Nature* reported on an experiment in which each of five groups consisting of six rats was put on a diet with a different carbohydrate. At the conclusion of the experiment, the DNA content of the liver of each rat was determined (mg/g liver), with the following results:

Carbohydrate	$\bar{x}_{i\cdot}$
Starch	2.58
Sucrose	2.63
Fructose	2.13

(continued)

Carbohydrate	$\bar{x}_{i\cdot}$
Glucose	2.41
Maltose	2.49

a. Assuming also that SST = 3.62, is the true average DNA content affected by the type of carbohydrate in the diet? Construct an ANOVA table and use a .05 level of significance.

b. Construct a t CI for the contrast

$$\theta = \mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4$$

which measures the difference between the average DNA content for the starch diet and the combined average for the four other diets. Does the resulting interval include zero?

c. What is β for the test when true average DNA content is identical for three of the diets and falls below this common value by 1 standard deviation (σ) for the other two diets?

82. Four laboratories (1–4) are randomly selected from a large population, and each is asked to make three determinations of the percentage of methyl alcohol in specimens of a compound taken from a single batch. Based on the accompanying data, are differences among laboratories a source of variation in the percentage of methyl alcohol? State and test the relevant hypotheses using significance level .05.

1:	85.06	85.25	84.87
2:	84.99	84.28	84.88
3:	84.48	84.72	85.10
4:	84.10	84.55	84.05

83. The article "Effects of Pulmonary Rehabilitation on Exercise Capacity and Disease Impact in Patients with Chronic Obstructive Pulmonary Disease and Obesity" (*Physiotherapy* 2018: 248–250) reports a study in which 155 COPD sufferers completed an eight-week pulmonary rehabilitation program at St. James' Hospital in Dublin,

Ireland. Before and after the program, subjects performed the Six-Minute Walk Test (6MWT), which simply measures the distance (in meters) patients can walk in six minutes. The accompanying table summarizes the *increase* in 6MWT distance for the participants (i.e., post-rehab distance minus pre-rehab distance), separated by their weight category.

Weight category	n	Mean	SD
Underweight/normal	53	61	80
Overweight	39	67	86
Obese	63	41	87

a. Does the data suggest that the pulmonary rehab program is not equally effective for COPD patients of all weight categories? State and test the relevant hypotheses at significance level .05.

b. Investigate differences between weight categories with respect to mean increase in 6MWT distance.

84. Recall from Section 11.2 that if $c_1, c_2, \ldots, c_I$ are numbers satisfying $\sum c_i = 0$ then $\theta = \sum c_i \mu_i$ is called a *contrast* in the μ_i's. Notice that with $c_1 = 1$, $c_2 = -1$, $c_3 = \cdots = c_I = 0$, $\sum c_i \mu_i = \mu_1 - \mu_2$, which implies that every pairwise difference between μ_i's is a contrast (and so is, e.g., $\mu_1 - .5\mu_2 - .5\mu_3$). A method attributed to Scheffé gives simultaneous CIs with simultaneous confidence level $100(1 - \alpha)\%$ for *all* possible contrasts (an infinite number of them!). The interval for $\sum c_i \mu_i$ is

$$\sum c_i \bar{x}_{i\cdot} \pm \sqrt{(I-1)F_{\alpha, I-1, n-I} \text{MSE} \sum c_i^2 / J_i}$$

Using the data from the previous exercise, calculate the 95% confidence Scheffé intervals for the contrasts $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, $\mu_2 - \mu_3$, and $.5\mu_1 + .5\mu_2 - \mu_3$ (the last contrast compares obese patients to the average of normal and overweight). Which contrasts differ significantly from 0, and why?

Regression and Correlation

12

Introduction

The general objective of a *regression analysis* is to investigate the relationship between two (or more) variables so that we can gain information about one of them through knowing values of the other(s). Much of mathematics is devoted to studying variables that are *deterministically* related, meaning that once we are told the value of x, the value of y is completely specified. For example, suppose we decide to rent a van for a day and that the rental cost is \$25.00 plus \$.30 per mile driven. Letting x = the number of miles driven and y = the rental charge, then $y = 25 + .3x$. If the van is driven 100 miles ($x = 100$), then $y = 25 + .3(100) = \$55$. As another example, suppose the initial velocity of a particle is v_0 and it undergoes constant acceleration a. Then distance traveled $= y = v_0 x + \frac{1}{2}ax^2$, where x = time.

There are many variables x and y that would appear to be related to each other, but not in a deterministic fashion. A familiar example to many students is given by variables x = high school grade point average (GPA) and y = college GPA. The value of y cannot be determined completely from knowledge of x, as two different students could have the same x value but very different y values. Yet there is a tendency for those students who have high (low) high school GPAs also to have high (low) college GPAs. Knowledge of a student's high school GPA should help us predict how that person will do in college. Other examples of variables related in a nondeterministic fashion include x = applied tensile force and y = amount of elongation in a metal strip, x = age of a child and y = size of that child's vocabulary, and x = size of an engine and y = fuel efficiency for an automobile equipped with that engine.

In this chapter, we generalize a deterministic linear relationship to obtain a *probabilistic* linear model for relating two variables x and y. We then develop procedures for making inferences based on data obtained from the model and obtain a quantitative measure (the *correlation coefficient*) of the extent to which the two variables are related. Techniques for assessing the adequacy of any particular regression model are then considered. *Multiple regression analysis* is introduced next as a way of relating y to two or more variables—for example, relating fuel efficiency of an automobile to weight, engine size, number of cylinders, and transmission type. The penultimate section of the chapter shows how matrix algebra techniques can be used to facilitate a concise and elegant development of regression procedures. The final section explains *logistic regression*, a method devised to predict a categorical variable y (e.g., absence or presence of lung cancer) from one or more x variables (amount of nicotine smoked/vaped per day, age, and so on).

12.1   The Simple Linear Regression Model

The key idea in developing a probabilistic relationship between a **response** (or **dependent**) variable y and an **explanatory** (or **predictor** or **independent**) variable x is to realize that once the value of x has been fixed, there is still uncertainty in what the resulting y value will be. That is, for a fixed value of x, we think of the response variable as being random. This random variable will be denoted by Y and its observed value by y. For example, suppose an investigator plans a study to relate y = yearly energy usage of an industrial building (1000's of BTUs) to x = the shell area of the building (ft^2). If one of the buildings selected for the study has a shell area of 25,000 ft^2, the resulting energy usage might be 2,215,000 or 2,348,000 or any one of a number of other possibilities. Since we don't know a priori what the value of energy usage will be—because usage is determined partly by factors other than shell area—usage is regarded as a random variable Y.

We typically relate the explanatory and response variables by an **additive model equation**:

$$Y = (\text{some particular deterministic function of } x) + (\text{a random deviation})$$
$$= f(x) + \varepsilon \tag{12.1}$$

The symbol ε represents a random deviation or random "error" (i.e., a random variable), which is assumed to have mean value 0. This rv incorporates all variation in the response variable due to factors *other than* x. Without the random deviation ε, whenever x is fixed prior to making an observation on the response variable, the resulting (x, y) point would fall exactly on the graph of $y = f(x)$, i.e., y would be entirely determined by x. The role of the random deviation ε is to allow a nondeterministic relationship. The assumption that ε has mean value 0 implies that, at any fixed x value, the *mean* (or *expected*) Y value is given by the function $f(x)$. In other words, we regard $f(x)$ in (12.1) as the *mean response* for a given x value.

How should the deterministic part of the model equation be selected? Occasionally some sort of theoretical argument will suggest an appropriate choice of $f(x)$. However, in practice the specification of $f(x)$ is almost always made by obtaining sample data consisting of (x, y) pairs. A picture of the resulting observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, called a **scatterplot**, is then constructed. In this scatterplot each (x_i, y_i) is represented as a point in a two-dimensional coordinate system. The pattern of points in the plot should suggest an appropriate $f(x)$.

Example 12.1   Troops deployed in active conflict areas worldwide depend on their body armor for protection. In conjunction with the US Army, the National Research Council developed the 2012 report "Testing of Body Armor Materials—Phase III." In one test, specimens of UHMWPE body armor were shot with a 7.62 mm round at different firing velocities. The accompanying data on x = velocity (m/s) and y = penetration area (mm^2, a proxy for amount of damage) appears in a graph in the report.

i	1	2	3	4	5	6	7	8	9	10
x_i	670	675	679	681	694	699	699	708	726	732
y_i	66.4	64.5	63.6	72.9	79.1	76.7	65.5	68.0	57.8	72.4

i	11	12	13	14	15	16	17	18	19	20
x_i	738	740	762	762	768	780	792	786	790	787
y_i	78.6	87.9	92.6	83.0	79.0	75.3	83.4	100.7	106.6	112.8

Thus $(x_1, y_1) = (670, 66.4)$, $(x_2, y_2) = (675, 64.5)$, and so on. A scatterplot is shown in Figure 12.1. Here are some things to notice about the data and plot:

- Several observations have identical x values yet different y values (e.g., $x_6 = x_7 = 699$, but $y_6 = 76.7$ and $y_7 = 65.5$). Thus x and y are not deterministically related.
- There is a strong tendency for y to increase as x increases. That is, higher firing velocities tend, not surprisingly, to be associated with larger penetration areas—a *positive relationship* between the variables.
- It appears that the value of y could be predicted from x by finding a line that "cuts through the heart" of the points in the plot; in fact, the authors of the report superimposed such a line on their plot. In other words, there is evidence of a substantial, though certainly not perfect, linear relationship between the two variables.
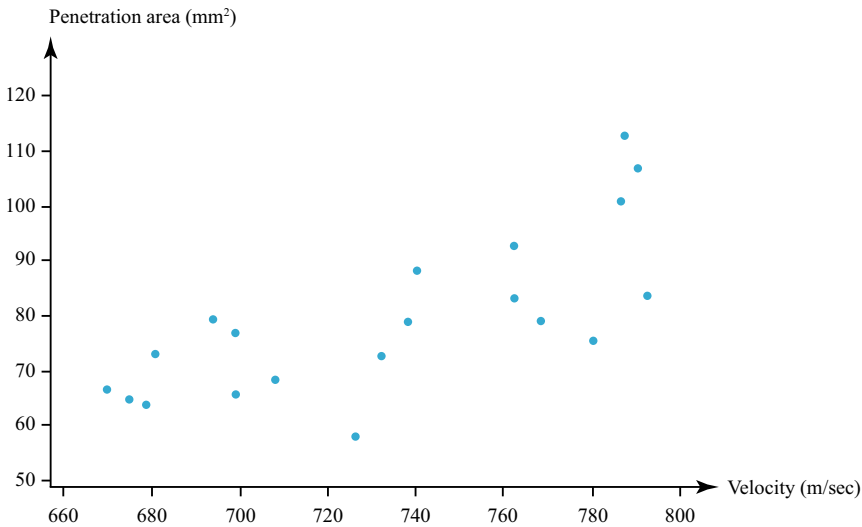
Figure 12.1   Scatterplot for the data from Example 12.1   ■

Notice that the axes in Figure 12.1 do not meet at $(0, 0)$; rather, the lower-left corner is roughly at $(50, 660)$. In most data sets, the values of x and/or y differ considerably from zero, and it makes better visual sense to adjust the axis boundaries to reflect the ranges of the variables.

Example 12.2   As demand for renewable energy such as solar and wind power increases, companies are spending more research money to develop more efficient methods for producing such energy. The scatterplot in Figure 12.2 shows the efficiency of a solar cell (y, measured as a percentage of the theoretical maximum efficiency) and the "sheet resistance" of the cell (x, measured in ohms) for a random sample of 132 prototype solar cells manufactured by a certain energy company. (Data provided by John Coleman; efficiency in the 8–15% range may seem low, but these were typical values in the solar energy industry at the time the data was collected.)
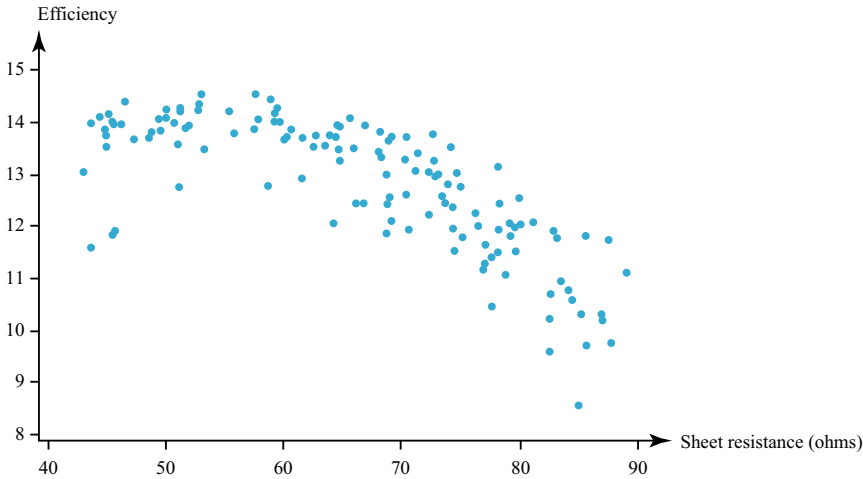
Figure 12.2 Scatterplot for the data from Example 12.2

As in the previous example, Figure 12.2 suggests a probabilistic relationship between the two variables: it appears that two solar cells with the same sheet resistance will not necessarily have the same efficiency. But the curvature in the scatterplot implies that a nonlinear relationship exists between x and y. A quadratic function $f(x)$ would be more appropriate here if we wished to apply the model equation (12.1) to this scenario. ∎

Throughout the next several sections, we will concentrate on situations for which a linear relationship, such as in Example 12.1, is reasonable. Quadratic and other more sophisticated models to accommodate data such as Example 12.2 are considered in Section 12.8.

A Linear Probabilistic Model

For the deterministic linear relationship $y = \beta_0 + \beta_1 x$, the **slope coefficient** β_1 is the guaranteed increase in y when x increases by one unit, and the **intercept coefficient** β_0 is the value of y when $x = 0$. When a scatterplot of bivariate data consisting of (x_i, y_i) pairs shows a reasonably substantial linear pattern, it is natural to specify $f(x)$ in the model equation (12.1) to be a linear function. Rather than assuming that the response variable itself is a linear function of x, the model assumes that the *expected* value of Y is a linear function of x. For each data point, the observed value of Y will deviate by a random amount from its expected value.

THE SIMPLE LINEAR REGRESSION MODEL

There are parameters β_0, β_1, and σ such that for any fixed value of the explanatory variable x, the response variable is related to x through the model equation

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Moreover, regardless of the fixed x value, the random variable ε is assumed to follow a $N(0, \sigma)$ distribution.

The term "simple" here refers to the use of a single explanatory variable; in Section 12.7, we will consider models with multiple x variables. The n observed pairs (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) are

regarded as having been generated independently of one another from the model equation: first fix $x = x_1$ and observe $Y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$, then fix $x = x_2$ and observe $Y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$, and so on. Assuming that the ε_i's are independent of each other implies that the Y_i's are also.

Figure 12.3 gives an illustration of data resulting from the simple linear regression model.
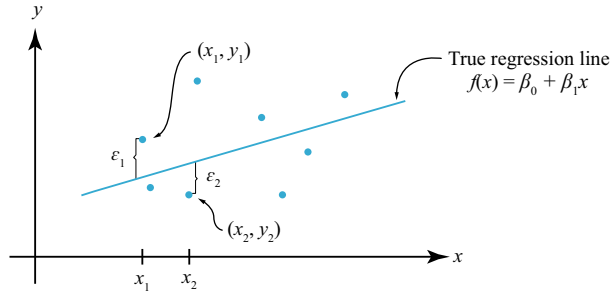
Figure 12.3 Points corresponding to observations from the simple linear regression model

The first two model parameters β_0 and β_1 are the coefficients of the **population** (or **true**) **regression line** $f(x) = \beta_0 + \beta_1 x$. The slope parameter β_1 is now interpreted as the *change in the expectation of Y* associated with a one-unit increase in x. As an example, if x = size of a house (sq. ft.), y = amount of natural gas used (therms) during a specified period, and $\beta_1 = .017$, then the change in expected gas usage associated with a one-sq-ft increase in house size is .017 therms. The standard deviation parameter σ controls the inherent amount of variability in the data. When σ is very close to 0, virtually all of the (x_i, y_i) pairs in the sample should correspond to points quite close to the population regression line. But if σ is relatively large, a number of points in the scatterplot are likely to fall far from the line. Roughly speaking, the magnitude of σ is the size of a "typical" deviation from the population line.

The following notation will help clarify implications of the model relationship. Let x^* denote a particular value of the explanatory variable x, and

$$\mu_{Y|x^*} = E(Y|x^*) = \text{the expected (i.e., mean) value of } Y \text{ when } x = x^*$$
$$\sigma^2_{Y|x^*} = V(Y|x^*) = \text{the variance of } Y \text{ when } x = x^*$$

For example, if x = applied stress (kg/mm^2) and y = time to fracture (h), then $\mu_{Y|20}$ denotes the expected time to fracture when applied stress is 20 kg/mm^2. If we conceptualize an entire population of (x, y) pairs resulting from applying stress to specimens, then $\mu_{Y|20}$ is the average of all values of the response variable for which $x = 20$. The variance $\sigma^2_{Y|20}$ describes the spread in the distribution of all y values for which applied stress is 20.

When the value $x = x^*$ is fixed, the only randomness on the right-hand side of the model equation is from the random deviation ε. Recalling that the mean value of a numerical constant is itself and that adding a constant does not affect variance, we have that

$$\mu_{Y|x^*} = E(\beta_0 + \beta_1 x^* + \varepsilon) = \beta_0 + \beta_1 x^* + E(\varepsilon) = \beta_0 + \beta_1 x^*$$
$$\sigma^2_{Y|x^*} = V(\beta_0 + \beta_1 x^* + \varepsilon) = V(\varepsilon) = \sigma^2$$

The first sequence of equalities says that the mean value of Y when $x = x^*$ is the height of the population regression line above the value x^*. That is, *the population regression line is the line of mean Y values*—the *mean response* is a linear function of the explanatory variable. The second sequence of equalities tells us that the amount of variability in the distribution of Y is the same at every x value—this "constant variance" assumption is part of the simple linear regression model.

The constant variance property implies that points should spread out about the population regression line to the same extent throughout the range of x values in the sample, rather than fanning out more as x increases or as x decreases. If x = age of a preschool child and Y = the child's vocabulary size, data suggests that mean vocabulary size increases linearly with age. However, there is more variability in vocabulary size for four-year-olds than for two-year-olds, so there is *not* constant variation in Y about the population line, and the simple linear regression model is therefore *not* appropriate. In Section 12.6, we will briefly discuss possible remedies to this assumption violation.

Finally, the sum of a constant and a normally distributed variable is itself normally distributed, and the addition of the constant affects only the mean value and not the standard deviation. So for any fixed value x^*, Y $(= \beta_0 + \beta_1 x^* + \varepsilon)$ has a normal distribution. The foregoing properties are summarized in Figure 12.4.
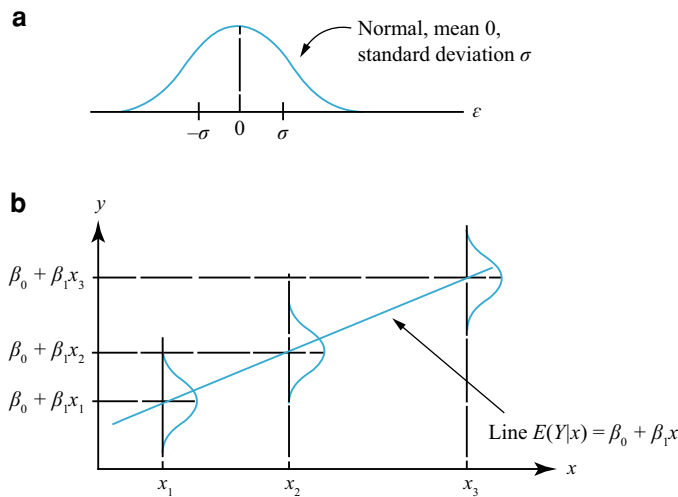
Figure 12.4 (**a**) Distribution of ε, (**b**) distribution of Y for different values of x

Example 12.3 Suppose the relationship between applied stress x and time to fracture y is described by the simple linear regression model with $\beta_0 = 65$, $\beta_1 = -1.2$, and $\sigma = 8$. Then there is a 1.2-h decrease in average (or expected) fracture time associated with an increase of 1 kg/mm² in applied stress. For any fixed value of x^* of stress, time to fracture is normally distributed with mean value $65 - 1.2x^*$ and standard deviation 8. Roughly speaking, in the population consisting of all (x, y) points, the magnitude of a typical deviation from the true regression line is about 8.

For $x = 20$, Y has mean value $\mu_{Y|20} = 65 - 1.2(20) = 41$, so

$$P(Y > 50 \text{ when } x = 20) = P\left(Z > \frac{50 - 41}{8}\right) = 1 - \Phi(1.13) = .1292$$

When applied stress is 25, $\mu_{Y|25} = 35$, so the probability that time to fracture exceeds 50 is

$$P(Y > 50 \text{ when } x = 25) = P\left(Z > \frac{50 - 35}{8}\right) = 1 - \Phi(1.88) = .0301$$

These probabilities are illustrated as the shaded areas in Figure 12.5.
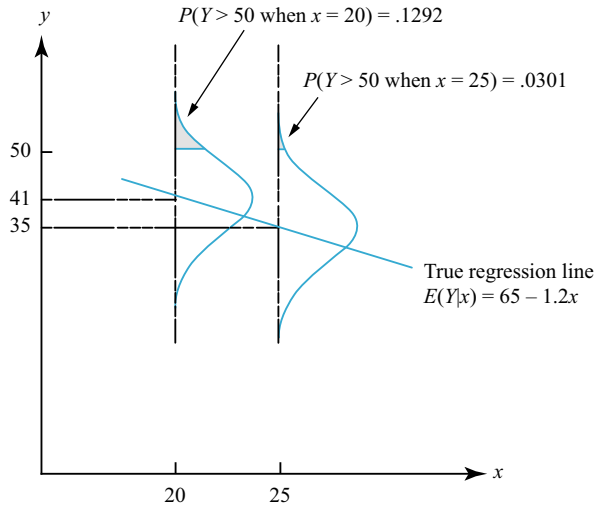
Figure 12.5 Probabilities based on the simple linear regression model

Suppose that Y_1 denotes an observation on time to fracture made with $x = 25$ and Y_2 denotes an independent observation made with $x = 24$. Then the difference $Y_1 - Y_2$ is normally distributed with mean value $E(Y_1 - Y_2) = \beta_1 = -1.2$, variance $V(Y_1 - Y_2) = \sigma^2 + \sigma^2 = 128$, and standard deviation $\sqrt{128} = 11.314$. The probability that Y_1 exceeds Y_2 is

$$P(Y_1 - Y_2 > 0) = P\left(Z > \frac{0 - (-1.2)}{11.314}\right) = P(Z > .11) = .4562$$

That is, even though we expect Y to decrease when x increases by one unit, the probability is fairly high (but less than .5) that the observed Y at $x + 1$ will be larger than the observed Y at x. ■

Our discussion thus far has presumed that the explanatory variable is under the control of the investigator, so that only the response variable Y is random. This will not always be the case: if we take a random sample of college students and record the height and weight of each, neither variable is preselected, so both x and y could be considered random. Methods and conclusions of the next several sections can be applied both when the values of the explanatory variable are fixed in advance and when they are random, but because the derivations and interpretations are more straightforward in the former case, we will continue to work explicitly with it. For more commentary, see the excellent book by Michael Kutner et al. listed in the bibliography.

Exercises: Section 12.1 (1–12)

1. Obesity is associated with higher foot load that can potentially increase pain and discomfort, but little research has been done on this relationship in children and its possible effects. A graph in the article "Childhood Obesity is Associated with Altered Plantar Pressure Distribution During Running" (*Gait and Posture* 2018: 202–205) gave the accompanying data on x = body mass index (kg/m^2) and y = peak foot pressure (kPa) while running for a sample of 42 children.

x	12.8	13.0	13.0	13.5	13.8	13.8	14.2	14.4	14.5
y	340	346	641	572	360	334	366	538	360

x	14.6	14.6	14.8	14.9	15.0	15.0	15.0	15.5	15.6
y	417	627	609	552	414	575	578	546	314

x	15.9	16.6	16.9	17.0	17.1	17.1	17.5	17.6	18.6
y	466	572	494	454	305	368	494	322	494

x	18.7	18.7	20.0	20.1	20.5	20.6	21.0	21.1	21.2
y	589	305	664	368	362	474	486	351	382

x	21.6	22.4	23.1	24.2	24.7	26.5
y	491	893	741	850	815	376

   a. Construct stem-and-leaf displays of both BMI and peak foot pressure, and comment on interesting features.
   b. Is the value of peak foot pressure completely and uniquely determined by BMI? Explain your reasoning.
   c. Construct a scatterplot of the data. Does it appear that peak foot pressure could be predicted by a child's body mass index? Explain your reasoning.

2. Verapamil is used to treat certain heart conditions, including high blood pressure and arrhythmia. Studies continue on the factors that affect the drug's absorption into the body. The article "The Effect of Non-ionic Surfactant Brij 35 on Solubility and Acid-Base Equilibria of Verapamil" (*J. Chem. Engr. Data* 2017: 1776–1781) includes the following data on x = pH and y = Verapamil solubility (10^{-5} mol/L) at 25 °C for one such study.

pH	8.12	8.32	8.41	8.62	8.70	8.84	8.88	9.09
sol.	53.4	32.3	22.2	14.6	13.9	8.76	5.06	5.57

   a. Construct a scatterplot of solubility versus pH, and describe what you see. Does it appear that a linear model would be appropriate here?
   b. Hydrogen ion concentration [H$^+$] is related to pH by pH = $-\log_{10}$([H$^+$]). Use this to calculate the hydrogen ion concentrations for each observation, then make a scatterplot of solubility versus [H$^+$]. Does it appear that a linear model would fit this data well?
   c. Would a linear function fit the data in part (b) perfectly? That is, is it reasonable to assume a completely deterministic relationship here? Explain your reasoning.

3. Bivariate data often arises from the use of two different techniques to measure the same quantity. As an example, the accompanying observations on x = hydrogen concentration (ppm) using a gas chromatography method and y = concentration using a new sensor method were read from a graph in the article "A New Method to Measure the Diffusible Hydrogen Content in Steel Weldments Using a Polymer Electrolyte-Based Hydrogen Sensor" (*Welding Res.*, July 1997: 251s–256s).

x	47	62	65	70	70	78	95	100	114	118
y	38	62	53	67	84	79	93	106	117	116

x	124	127	140	140	140	150	152	164	198	221
y	127	114	134	139	142	170	149	154	200	215

   Construct a scatterplot. Does there appear to be a very strong relationship between the two types of concentration measurements? Do the two methods appear to be measuring roughly the same quantity? Explain your reasoning.

4. A study to assess the capability of subsurface flow wetland systems to remove biochemical oxygen demand (BOD, a measure of organic matter in sewage) and various other chemical constituents resulted in the accompanying data on x = BOD mass loading (kg/ha/d) and y = BOD mass removal (kg/ha/d) ("Subsurface Flow

Wetlands—A Performance Evaluation," *Water Environ. Res.* 1995: 244–247).

x	3	8	10	11	13	16	27
y	4	7	8	8	10	11	16

x	30	35	37	38	44	103	142
y	26	21	9	31	30	75	90

a. Construct boxplots of both mass loading and mass removal, and comment on any interesting features.
b. Construct a scatterplot of the data, and comment on any interesting features.

5. The article "Objective Measurement of the Stretchability of Mozzarella Cheese" (*J. Texture Stud.* 1992: 185–194) reported on an experiment to investigate how the behavior of mozzarella cheese varied with temperature. Consider the accompanying data on x = temperature and y = elongation (%) at failure of the cheese. [*Note*: The researchers were Italian and used *real* mozzarella cheese, not the poor cousin widely available in the USA.]

x	59	63	68	72	74	78	83
y	118	182	247	208	197	135	132

a. Construct a scatterplot in which the axes intersect at (0, 0). Mark 0, 20, 40, 60, 80, and 100 on the horizontal axis and 0, 50, 100, 150, 200, and 250 on the vertical axis.
b. Construct a scatterplot in which the axes intersect at (55, 100), as was done in the cited article. Does this plot seem preferable to the one in part (a)? Explain your reasoning.
c. What do the plots of parts (a) and (b) suggest about the nature of the relationship between the two variables?

6. One factor in the development of tennis elbow, a malady that strikes fear in the hearts of all serious tennis players, is the impact-induced vibration of the racket-and-arm system at ball contact. It is well known that the likelihood of getting tennis elbow depends on various properties of the racket

used. Consider the scatterplot of x = racket resonance frequency (Hz) and y = sum of peak-to-peak acceleration (a characteristic of arm vibration, in m/s/s) for $n = 23$ different rackets ("Transfer of Tennis Racket Vibrations into the Human Forearm," *Med. Sci. Sports Exercise* 1992: 1134–1140). Discuss interesting features of the data and scatterplot.
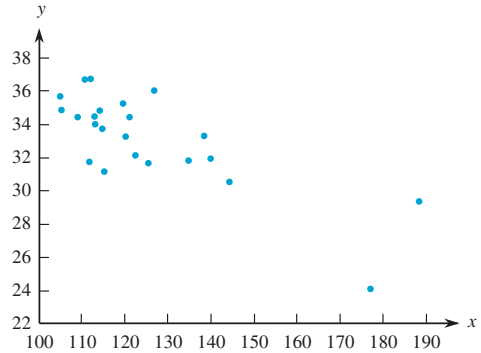
7. Data from the EPA's Fuel Efficiency Guide suggests an approximate linear relationship between y = highway fuel efficiency (mpg) and x = weight (lbs) for midsize cars. Suppose the equation of the true regression line is $f(x) = 70 - .0085x$.

a. What is the expected value of highway fuel efficiency when weight = 2500 lbs?
b. By how much can we expect highway fuel efficiency to change when weight increases by 1 lb?
c. Answer part (b) for an increase of 500 lbs.
d. Answer part (b) for a decrease of 500 lbs.

8. Referring to the previous exercise, suppose that the random deviation ε is normally distributed with standard deviation 4.6 mpg.

a. What is the probability that the observed value of highway fuel efficiency will exceed 30 mpg when the car's weight is 4000 lbs?
b. Repeat part (a) with 5000 in place of 4000.

c. Consider making two independent observations on highway fuel efficiency, the first for a car weighing 4000 lbs and the second for $x = 5000$. What is the probability that the first observation will exceed the second by more than 5 mpg?

d. Let Y_1 and Y_2 denote observations on highway fuel efficiency when $x = x_1$ and $x = x_2$, respectively. By how much would x_2 have to exceed x_1 in order that $P(Y_1 > Y_2) = .95$?

9. The flow rate y (m^3/min) in a device used for air-quality measurement depends on the pressure drop x (in. of water) across the device's filter. Suppose that for x values between 5 and 20, the two variables are related according to the simple linear regression model with true regression line $E(Y|x) = -.12 + .095x$.

   a. What is the expected change in flow rate associated with a 1-in. increase in pressure drop? Explain.

   b. What change in flow rate can be expected when pressure drop decreases by 5 in.?

   c. What is the expected flow rate for a pressure drop of 10 in.? A drop of 15 in.?

   d. Suppose $\sigma = .025$ and consider a pressure drop of 10 in. What is the probability that the observed value of flow rate will exceed .835? That observed flow rate will exceed .840?

   e. What is the probability that an observation on flow rate when pressure drop is 10 in. will exceed an observation on flow rate made when pressure drop is 11 in.?

10. Suppose the expected cost of a production run is related to the size of the run by the equation $E(Y|x) = 4000 + 10x$. Let Y denote an observation on the cost of a run. Assuming that the variables *size* and *cost* are related according to the simple linear regression model, could it be the case that

$P(Y > 5500$ when $x = 100) = .05$ and $P(Y > 6500$ when $x = 200) = .10$? Explain.

11. Suppose that in a certain chemical process the reaction time y (hr) is related to the temperature (°F) in the chamber in which the reaction takes place according to the simple linear regression model with equation $E(Y|x) = 5.00 - .01x$ and $\sigma = .075$.

   a. What is the expected change in reaction time for a 1 °F increase in temperature? For a 10 °F increase in temperature?

   b. What is the expected reaction time when temperature is 200 °F? When temperature is 250 °F?

   c. Suppose five observations are made independently on reaction time, each one for a temperature of 250 °F. What is the probability that all five times are between 2.4 and 2.6 h?

   d. What is the probability that two independently observed reaction times for temperatures 1° apart are such that the time at the higher temperature exceeds the time at the lower temperature?

12. The article "On the Theoretical Velocity Distribution and Flow Resistance in Natural Channels" (*J. Hydrol.* 2017: 777–785) suggests a quadratic relationship between $x = $ flow depth (m) and $y = $ water surface slope at certain points along the Tiber River in Italy. Suppose the variables are related by Equation (12.1) with $f(x) = -0.6x^2 + 5x + 1$ (similar to the equation suggested in the article).

   a. What is the expected water surface slope when the flow depth is 2.0 m? 2.5 m? 3.0 m?

   b. Does the expected water surface slope change by a fixed amount for each 1-m increase in flow depth? Explain.

   c. Determine a flow depth for which the expected surface slope is the same as the expectation for $x = 2.0$ m obtained in part (a). [*Note:* Your answer should be something other than 2.0.]

d. For what depth is expected water sur-
    face slope maximized?
e. Assume the rv ε in Equation (12.1) has
    a standard normal distribution; this is
    consistent with information in the

article. At a flow depth of 3.0 m,
what's the probability the water surface
slope Y is greater than 10? Less than 6?

12.2   Estimating Model Parameters

We will assume in this and the next several sections that the variables x and y are related according to the simple linear regression model. The values of the parameters β_0, β_1, and σ will almost never be known to an investigator. Instead, sample data consisting of n observed pairs $(x_1, y_1), \ldots, (x_n, y_n)$ will be available, from which the model parameters and the true regression line itself can be estimated. These observations are assumed to have been obtained independently of each other.

According to the model, the observed points will be distributed about the true regression line $f(x) = \beta_0 + \beta_1 x$ in a random manner. Figure 12.6 shows a scatterplot of observed pairs along with two candidates for the estimated regression line, $y = a_0 + a_1 x$ and $y = b_0 + b_1 x$. Intuitively, the line $y = a_0 + a_1 x$ is not a reasonable estimate of the true line because, if $y = a_0 + a_1 x$ were the true line, the observed points would almost surely have been closer to this line. The line $y = b_0 + b_1 x$ is a more plausible estimate because the observed points are scattered rather closely about this line.
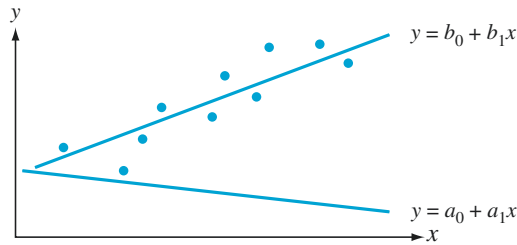
Figure 12.6   Two different estimates of the true regression line: one good and one bad

Figure 12.6 and the foregoing discussion suggest that our estimate of $\beta_0 + \beta_1 x$ should be a line that provides, in some sense, a "best fit" to the observed data points. This is what motivates the **principle of least squares**, which can be traced back to the mathematicians Gauss and Legendre around the year 1800. According to this principle, a line provides a good fit to the data if the vertical distances or deviations from the observed points to the line (see Figure 12.7) are small. The proposed measure of the goodness-of-fit is the sum of the squares of these deviations; the best-fit line is then the one having the smallest possible sum of squared deviations.
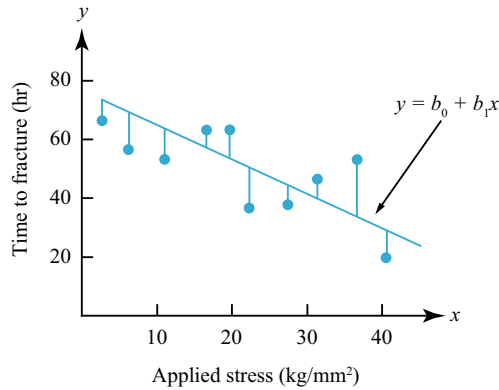
Figure 12.7 Deviations of observed data from line $y = b_0 + b_1 x$

PRINCIPLE OF LEAST SQUARES	The vertical deviation of the point (x_i, y_i) from a line $y = b_0 + b_1 x$ is

$$\text{height of point} - \text{height of line} = y_i - (b_0 + b_1 x_i)$$

The sum of squared vertical deviations from the points (x_1, y_1), …, (x_n, y_n) to the line is then

$$g(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least squares estimates**, are those values that minimize $g(b_0, b_1)$. The **estimated regression line** or **least squares regression line (LSRL)** is then the line whose equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

The minimizing values of b_0 and b_1 are found by taking partial derivatives of $g(b_0, b_1)$ with respect to both b_0 and b_1, equating them both to zero, and solving the equations

$$\frac{\partial g(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$
$$\frac{\partial g(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

Cancellation of the factor 2 and re-arrangement gives the following system of equations, called the **normal equations**:

$$n b_0 + \left(\sum x_i \right) b_1 = \sum y_i$$
$$\left(\sum x_i \right) b_0 + \left(\sum x_i^2 \right) b_1 = \sum x_i y_i$$

The normal equations are linear in the two unknowns b_0 and b_1. Provided that at least two of the x_i's are different, the least squares estimates are the unique solution to this linear system.

PROPOSITION   The least squares estimate of the slope coefficient β_1 of the true regression line is

$$b_1 = \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{12.2}$$

The least squares estimate of the intercept β_0 of the true regression line is

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \tag{12.3}$$

Moreover, under the normality assumption of the simple linear regression model, $\hat{\beta}_0$ and $\hat{\beta}_1$ are also the maximum likelihood estimates (see Exercise 23).

Because they will feature prominently here and in subsequent sections, we define the following notation for certain sums:

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The S_{xx} formula was presented in Chapter 1 in connection with the sample variance: $s_x^2 = S_{xx}/(n-1)$ and similarly for y. The least squares estimates of the regression coefficients can then be written as

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

We emphasize that *before $\hat{\beta}_1$ and $\hat{\beta}_0$ are computed, a scatterplot should be examined to see whether a linear probabilistic model is plausible*. If the points do not tend to cluster about a straight line with roughly the same degree of spread for all x (e.g., Figure 12.2), then other models should be investigated.

Example 12.4 As brick-and-mortar shops decline and online retailers like Amazon and Wayfair ascend, demand for warehouse storage space has steadily increased. Despite effectively being empty shells, warehouses still require professional appraisal. The following data on x = truss height (ft), which determines how high stored goods can be stacked, and y = sale price ($) per square foot appeared in the article "Challenges in Appraising 'Simple' Warehouse Properties" (*The Appraisal J.* 2001: 174–178).

Warehouse	1	2	3	4	5	6	7	8	9	10
x	12	14	14	15	15	16	18	22	22	24
y	35.53	37.82	36.90	40.00	38.00	37.50	41.00	48.50	47.00	47.50

Warehouse	11	12	13	14	15	16	17	18	19
x	24	26	26	27	28	30	30	33	36
y	46.20	50.35	49.13	48.07	50.90	54.78	54.32	57.17	57.45

From the sample data,

$$\bar{x} = \frac{1}{19}\sum x_i = 22.737 \text{ ft}, \qquad\qquad \bar{y} = \frac{1}{19}\sum y_i = 46.217 \, \$/\text{ft}^2$$

$$S_{xx} = \sum (x_i - 22.737)^2 = 913.684 \qquad S_{yy} = \sum (y_i - 46.217)^2 = 924.436$$

$$S_{xy} = \sum (x_i - 22.737)(y_i - 46.217) = 901.944$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{901.944}{913.684} = 0.987$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 46.217 - 0.987(22.737) = 23.8$$

The equation of the LSRL is $y = 23.8 + .987x$. We estimate that the change in expected sale price associated with a 1-ft increase in truss height is .987, or about 99 cents per square foot. The intercept of 23.8, while important for correctly summarizing the data, does not have a direct interpretation— after all, it doesn't make sense for a warehouse to have a truss height of $x = 0$ feet (how would you store anything?). Figure 12.8, generated by the statistical software package R, shows that the least squares line provides an excellent summary of the relationship between the two variables.
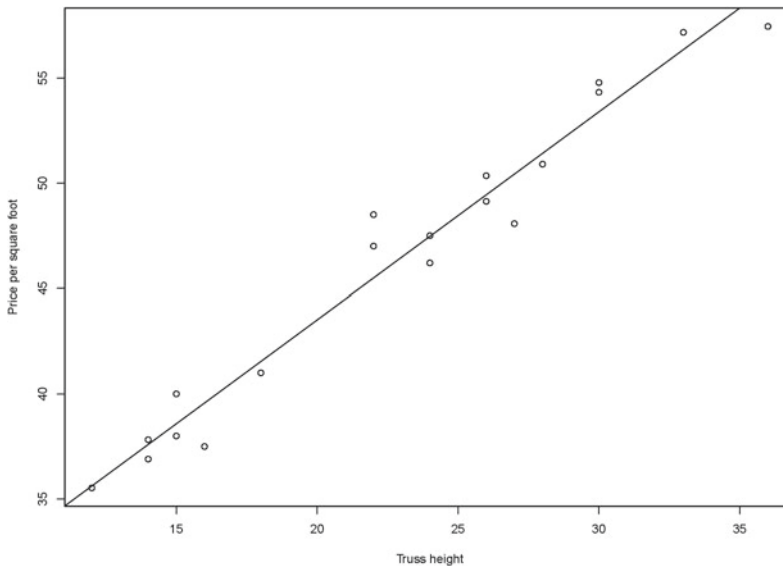
Figure 12.8   A scatterplot of the data in Example 12.4 with the LSRL superimposed, from R   ■

The LSRL can immediately be used for two different purposes. For a fixed x value x^*, $\hat{\beta}_0 + \hat{\beta}_1 x^*$ (the height of the line above x^*) gives both (1) a point estimate of the mean value of Y when $x = x^*$ and (2) a point prediction of the Y value that will result from a single new observation made at $x = x^*$.

The least squares line should *not* be used to make a prediction for an x value much beyond the range of the data, such as $x = 5$ or $x = 45$ in Example 12.4. The **danger of extrapolation** is that the fitted relationship (a line here) may not be valid for such x values.

Example 12.5 (Example 12.4 continued) A point estimate for the true average price for all warehouses with 25-ft truss height is

$$\hat{\mu}_{Y|25} = \hat{\beta}_0 + \hat{\beta}_1(25) = 23.8 + .987(25) = \$48.48/\text{ft}^2$$

This also represents a point prediction for the price of a single warehouse with 25-ft truss height. Notice that although no sample observations had $x = 25$, this value lies in the "middle" of the set of x values (see Figure 12.8). This is an example of **interpolation**: using the LSRL for x values that were unseen but are consistent with the sample data.

A point estimate for the true average price for all warehouses with 50-ft truss height is

$$\hat{\mu}_{Y|50} = \hat{\beta}_0 + \hat{\beta}_1(50) = 23.8 + .987(50) = \$73.15/\text{ft}^2$$

However, because this calculation involves an extrapolation—the value $x = 50$ is well outside the bounds of the available data—we have much less faith that this estimated cost is accurate. ∎

Residuals and Estimating σ

The parameter σ determines the amount of variability inherent in the regression model. A large value of σ will lead to observed (x_i, y_i)'s that are typically quite spread out about the true regression line, whereas when σ is small the observed points will tend to fall very close to the true line (see Figure 12.9). An estimate of σ will be used in confidence interval formulas and hypothesis-testing procedures presented in the next two sections. Because the equation of the true line is unknown, the estimate is based on the extent to which the sample observations deviate from the estimated line.
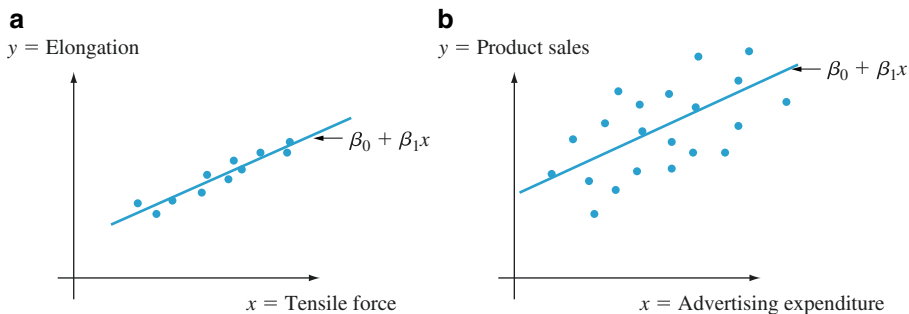
Figure 12.9 Typical sample for σ: (**a**) small; (**b**) large

---

DEFINITION   The **fitted** (or **predicted**) **values** $\hat{y}_1, \ldots, \hat{y}_n$ are obtained by successively substituting the x values $x_1, \ldots, x_n$ into the equation of the LSRL: the ith fitted value is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$$

The **residuals** $e_1, \ldots, e_n$ are the vertical deviations from the LSRL: the ith residual is

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = (y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x}) \tag{12.4}$$

In words, the predicted value $\hat{y}_i$ is the value of y that we would predict or expect when using the estimated regression line with $x = x_i$; $\hat{y}_i$ is the height of the estimated regression line above x_i. The residual e_i is the difference between the observed y_i and the predicted $\hat{y}_i$.

Assuming that the line in Figure 12.7 is the least squares line, the residuals are identified by the vertical line segments from the observed points to the line. In fact, the principle of least squares is equivalent to determining the line for which the sum of squared residuals is minimized. If the residuals are all small in magnitude, then much of the variability in observed y values appears to be due to the linear relationship between x and y, whereas many large residuals suggest quite a bit of inherent variability in y relative to the amount due to the linear relation. The residuals from the LSRL always satisfy $\sum e_i = 0$ and so $\bar{e} = 0$ (see Exercise 24; in practice, the sum may deviate a bit from zero due to rounding).

The ith residual e_i may also be regarded as a proxy for the unobservable "true" error ε_i for the ith observation:

$$\text{true error:} \varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$
$$\text{estimated error:} e_i = \hat{\varepsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

The sum of squared residuals is used here to estimate the standard deviation σ of the ε_i's in the same way that the sum of squares S_{xx} was previously used to estimate a population sd.

DEFINITION   The **error sum of squares** (or **residual sum of squares**), denoted by SSE, is

$$\text{SSE} = \sum (e_i - \bar{e})^2 = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

and the least squares estimate of σ^2 is

$$\hat{\sigma}^2 = s_e^2 = \frac{\text{SSE}}{n - 2} = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2}$$

The estimate $s_e = \sqrt{\text{SSE}/(n - 2)}$ of σ is called the **residual standard deviation**.

The divisor $n - 2$ in s_e is the number of degrees of freedom (df) associated with the estimate (or, equivalently, with the error sum of squares). This is because to obtain s_e, the two parameters β_0 and β_1 must first be estimated, which results in a loss of 2 df (just as μ had to be estimated in one-sample problems, resulting in an estimated variance based on $n - 1$ df). Equivalently, the normal equations impose two constraints; as a result, if $n - 2$ of the residuals are known, then the remaining two are completely determined (so only $n - 2$ are freely determined; see Exercise 24).

Replacing each y_i in the formula for s_e by the rv Y_i gives the estimator S_e. It can be shown that S_e^2 is an unbiased estimator for σ^2, although the estimator S_e is biased for σ. (The mle of σ^2 based on the normal model has divisor n rather than $n - 2$, so it is biased.)

The interpretation of s_e here is similar to that of σ given earlier. Roughly speaking, s_e is the size of a "typical" or "representative" deviation from the least squares line.

Example 12.6 Japan's high population density has resulted in a multitude of resource usage problems. One especially serious difficulty concerns waste removal. The article "Innovative Sludge Handling Through Pelletization Thickening" (*Water Res*. 1999: 3245–3252) reported the development of a new compression machine for processing sewage sludge. An important part of the investigation involved relating the moisture content of compressed pellets (y, in %) to the machine's filtration rate (x, in kg-DS/m/h). The following data was read from a graph in the paper:

x	125.3	98.2	201.4	147.3	145.9	124.7	112.2	120.2	161.2	178.9
y	77.9	76.8	81.5	79.8	78.2	78.3	77.5	77.0	80.1	80.2
x	159.5	145.8	75.1	151.4	144.2	125.0	198.8	132.5	159.6	110.7
y	79.9	79.0	76.7	78.2	79.5	78.1	81.5	77.0	79.0	78.6

Relevant summary quantities are $\bar{x} = 140.895$, $\bar{y} = 78.74$, $S_{xx} = 18{,}921.8295$, and $S_{xy} = 776.434$, from which

$$\hat{\beta}_1 = \frac{776.434}{18{,}921.8295} = .04103377 \approx .041$$

$$\hat{\beta}_0 = 78.74 - (.04103377)(140.895) = 72.958547 \approx 72.96$$

The equation of the least squares line is $y = 72.96 + .041x$. For numerical accuracy, the fitted values are calculated from $\hat{y}_i = 72.958547 + .04103377x_i$:

$$\hat{y}_1 = 72.958547 + .04103377(125.3) \approx 78.100, \quad e_1 = y_1 - \hat{y}_1 \approx -.200, \text{ etc.}$$

A positive residual corresponds to a point in the scatterplot that lies above the graph of the least squares line, whereas a negative residual results from a point lying below the line. All predicted values (fits) and residuals appear in the accompanying table.

Obs. (i)	Filtrate (x_i)	Moist. Con. (y_i)	Fit ($\hat{y}_i$)	Residual (e_i)
1	125.3	77.9	78.100	−0.200
2	98.2	76.8	76.988	−0.188
3	201.4	81.5	81.223	0.277
4	147.3	79.8	79.003	0.797
5	145.9	78.2	78.945	−0.745
6	124.7	78.3	78.075	0.225
7	112.2	77.5	77.563	−0.063
8	120.2	77.0	77.891	−0.891
9	161.2	80.1	79.573	0.527
10	178.9	80.2	80.299	−0.099
11	159.5	79.9	79.503	0.397
12	145.8	79.0	78.941	0.059
13	75.1	76.7	76.040	0.660
14	151.4	78.2	79.171	−0.971
15	144.2	79.5	78.876	−0.624
16	125.0	78.1	78.088	0.012
17	198.8	81.5	81.116	0.384
18	132.5	77.0	78.396	−1.396
19	159.6	79.0	79.508	−0.508
20	110.7	78.6	77.501	1.099

It can be verified that, rounding error notwithstanding, the residuals (the last column) sum to 0. The corresponding residual sum of squares is

$$\text{SSE} = (-.200)^2 + (-.188)^2 + \cdots + (1.099)^2 = 7.968$$

The estimate of σ^2 is then $\hat{\sigma}^2 = s_e^2 = 7.968/(20 - 2) = .4427$, and the residual standard deviation is $\hat{\sigma} = s_e = \sqrt{.4427} = .665$. Roughly speaking, .665 is the typical difference between the actual moisture concentration of a specimen and its predicted moisture concentration based on the LSRL. ∎

Computation of SSE from the defining formula involves much tedious arithmetic, because both the predicted values and residuals must first be calculated. Use of the following formula does not require these quantities (again see Exercise 24), though $\sum y_i$ and $\sum y_i^2$ are needed.

$$\text{SSE} = S_{yy} - S_{xy}^2/S_{xx}$$

The Coefficient of Determination

Figure 12.10 shows three different scatterplots of bivariate data. In all three plots, the heights of the different points vary substantially, indicating that there is much variability in observed y values. The points in the first plot all fall exactly on a straight line. In this case, all (100%) of the sample variation in y can be attributed to the fact that x and y are linearly related in combination with variation in x. The points in Figure 12.10b do not fall exactly on a line, but compared to overall y variability, the deviations from the least squares line are small. It is reasonable to conclude in this case that much of the observed y variation can be attributed to the approximate linear relationship between the variables postulated by the simple linear regression model. When the scatterplot looks like that of Figure 12.10c, there is substantial variation about the least squares line relative to overall y variation, so the simple linear regression model fails to explain much of the variation in y by relating y to x.
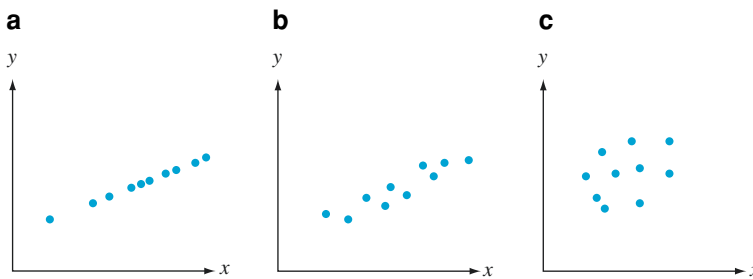
Figure 12.10 Explaining y variation: (**a**) all variation explained; (**b**) most variation explained; (**c**) little variation explained

The error sum of squares SSE can be interpreted as a measure of how much variation in y is left unexplained by the model—that is, how much *cannot* be attributed to a linear relationship. In Figure 12.10a, SSE = 0 and there is no unexplained variation, whereas unexplained variation is small for the data of Figure 12.10b and much larger in Figure 12.10c. A quantitative measure of the *total* amount of variation in the observed y values is given by the **total sum of squares**

$$\text{SST} = \sum (y_i - \bar{y})^2 = S_{yy}$$

Figure 12.11 illustrates the difference between these two sums of squares. The (x, y) points in the two scatterplots are identical. While SSE measures the deviation of the y values from the LSRL (a "model" that uses x as a predictor), SST measures the deviation of the y values from the horizontal line $y = \bar{y}$ (essentially ignoring the presence of x). Since the least squares line is by definition the line having the smallest sum of squared vertical deviations, SSE can't be any larger than SST, and usually it is much smaller.
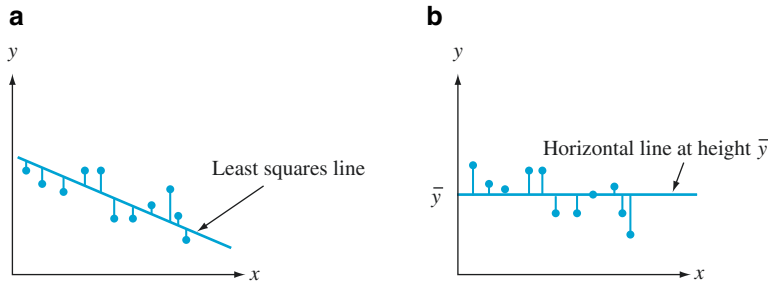
Figure 12.11   Sums of squares illustrated: (**a**) SSE = sum of squared deviations about the least squares line; (**b**) SST = sum of squared deviations about the horizontal line $y = \bar{y}$

Dividing SSE by SST gives the *proportion* of total variation that is not explained by the approximate linear relationship. Subtracting this ratio from 1 results in the proportion of total variation that *is* explained by the relationship.

DEFINITION   The **coefficient of determination**, denoted by R^2, is given by

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

R^2 is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model (i.e., attributed to an approximate linear relationship between y and x).

The closer R^2 is to 1, the more successful the simple linear regression model is in explaining y variation. Multiplying R^2 by 100 gives the percentage of total variation explained by the relationship; software often reports R^2 this way.

Said differently, R^2 is the proportion by which the error sum of squares is *reduced* by the regression line compared to the horizontal line. For example, if SST = 20 and SSE = 2, then $R^2 = 1 - (2/20) = .9$, so the regression reduces the error sum of squares by 90%.

Although it is common to have R^2 values of .9 or more in engineering and the physical sciences, R^2 is likely to be much smaller in social sciences such as psychology and sociology, where values far less than .5 are common but still considered important.

Example 12.7 (Example 12.5 continued) The scatterplot of the truss height-sale price data in Figure 12.8 indicates a fairly high R^2 value. Previous computations showed that SST $= S_{yy} =$ 924.436, $S_{xy} = 901.944$, and $S_{xx} = 913.684$. Using the computational shortcut,

$$\text{SSE} = 924.436 - (901.944)^2/913.684 = 34.081$$

The coefficient of determination is then

$$R^2 = 1 - \frac{34.081}{924.436} = 1 - .037 = .963$$

That is, 96.3% of the observed variation in warehouse price is attributable to (can be explained by) the approximate linear relationship between price and truss height, a fairly impressive result. The R^2 can also be interpreted by saying that the error sum of squares using the regression line is 96.3% less than the error sum of squares using a horizontal line (i.e., ignoring truss height).

Figure 12.12 shows partial Minitab output for the warehouse data; the package will also provide the predicted values and residuals upon request, as well as other information. The formats used by other packages differ slightly from that of Minitab, but the information content is very similar. Quantities in Figure 12.12 such as the standard deviations, t ratios, and the details of the ANOVA table are discussed in Section 12.3.

```
The regression equation is
Sales Price = 23.8 + 0.987 Truss Height

Predictor          Coef      SE Coef       T        P
Constant        23.772←β̂₀     1.113    21.35   0.000
Truss Height   0.98715←β̂₁  0.04684    21.07   0.000

S = 1.41590←sₑ     R-Sq = 96.3%←100R²      R-Sq(adj) = 96.1%

Analysis of Variance

Source           DF        SS           MS        F        P
Regression        1     890.36        890.36   444.11   0.000
Residual Error   17      34.08←SSE     2.00
Total            18     924.44←SST
```

Figure 12.12 Minitab output for the regression of Example 12.7   ∎

For regression there is an analysis of variance identity like the fundamental identity (11.1) in Chapter 11. Add and subtract $\hat{y}_i$ in the total sum of squares:

$$\text{SST} = \sum (y_i - \bar{y})^2 = \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Notice that the middle (cross-product) term is missing on the right; see Exercise 24 for the justification. Of the two sums on the right, the first is SSE $= \sum (y_i - \hat{y}_i)^2$ and the second is something new, the **regression sum of squares**, SSR $= \sum (\hat{y}_i - \bar{y})^2$. The analysis of variance identity for regression is

$$SST = SSE + SSR \qquad (12.5)$$

The coefficient of determination can now be written in a slightly different way:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SST - SSE}{SST} = \frac{SSR}{SST}$$

The ANOVA table in Figure 12.12 shows that SSR = 890.36, from which R^2 = 890.36/924.44 = .936 as before. Hence we interpret the regression sum of squares SSR as the amount of total variation explained by the model, so that R^2 is the ratio of explained variation to total variation.

Exercises: Section 12.2 (13–30)

13. Exercise 4 gave data on x = BOD mass loading and y = BOD mass removal. Values of relevant summary quantities are

$$n = 14 \qquad \sum x_i = 517 \qquad \sum y_i = 346$$
$$S_{xy} = 13{,}048 \qquad S_{xx} = 20{,}003 \qquad S_{yy} = 8903$$

   a. Obtain the equation of the least squares line.
   b. Predict the value of BOD mass removal for a single observation made when BOD mass loading is 35, and calculate the value of the corresponding residual.
   c. Calculate SSE and then a point estimate of σ.
   d. What proportion of observed variation in removal can be explained by the approximate linear relationship between the two variables?
   e. The last two x values, 103 and 142, are much larger than the others. How are the equation of the least squares line and the value of R^2 affected by deletion of the two corresponding observations from the sample? Adjust the given values of the summary quantities, and use the fact that the new value of SSE is 311.79.

14. The accompanying data on x = current density (mA/cm^2) and y = rate of deposition (mm/min) appeared in the article "Plating of 60/40 Tin/Lead Solder for Head

Termination Metallurgy" (*Plating and Surface Finishing*, Jan. 1997: 38–40). Do you agree with the claim by the article's author that "a linear relationship was obtained from the tin–lead rate of deposition as a function of current density"? Explain your reasoning.

x	20	40	60	80
y	.24	1.20	1.71	2.22

15. The efficiency ratio for a steel specimen immersed in a phosphating tank is the weight of the phosphate coating divided by the metal loss (both in mg/ft^2). The article "Statistical Process Control of a Phosphate Coating Line" (*Wire J. Internat.*, May 1997: 78–81) gave the accompanying data on tank temperature (x) and efficiency ratio (y).

Temp.	170	172	173	174	174	175	176
Ratio	.84	1.31	1.42	1.03	1.07	1.08	1.04
Temp.	177	180	180	180	180	180	181
Ratio	1.80	1.45	1.60	1.61	2.13	2.15	.84
Temp.	181	182	182	182	182	184	184
Ratio	1.43	.90	1.81	1.94	2.68	1.49	2.52
Temp.	185	186	188				
Ratio	3.00	1.87	3.08				

   a. Determine the equation of the estimated regression line.

b. Calculate a point estimate for true average efficiency ratio when tank temperature is 182.

c. Calculate the values of the residuals from the least squares line for the four observations for which temperature is 182. Why do they not all have the same sign?

d. What proportion of the observed variation in efficiency ratio can be attributed to the simple linear regression relationship between the two variables?

16. The scientist Francis Galton, an early developer of regression methodology, used "midparent height," the average of the father's and mother's heights, in order to predict children's heights. Here are the heights of 11 female students along with their midparent heights in inches:

Midparent	66.0	65.5	71.5	68.0	70.0	65.5
Daughter	64.0	63.0	69.0	69.0	69.0	65.0

Midparent	67.0	70.5	69.5	64.5	67.5
Daughter	63.0	68.5	69.0	64.0	67.0

a. Construct a scatterplot of daughter's height against the midparent height and comment on the strength of the relationship.

b. Is the daughter's height completely and uniquely determined by the midparent height? Explain.

c. Use the accompanying Minitab output to obtain the equation of the least squares line for predicting daughter height from midparent height, and then predict the height of a daughter whose midparent height is 70 in. Would you feel comfortable using the least squares line to predict daughter height when midparent height is 74 in.? Explain.

Predictor	Coef	SE Coef	T	P
Constant	1.65	13.36	0.12	0.904
midparent	0.9555	0.1971	4.85	0.001

S = 1.45061 R-Sq = 72.3% R-Sq(adj) = 69.2%

```
          Analysis of Variance

Source       DF    SS      MS      F      P
Regression   1   49.471  49.471  23.51  0.001
Residual     9   18.938   2.104
Error
Total        10  68.409
```

d. What are the values of SSE, SST, and the coefficient of determination? How well does the midparent height account for the variation in daughter height?

17. The article "Characterization of Highway Runoff in Austin, Texas, Area" (*J. Environ. Engr*. 1998: 131–137) gave a scatterplot, along with the least squares line, of x = rainfall volume (m^3) and y = runoff volume (m^3) for a particular location. The accompanying values were read from the plot.

x	5	12	14	17	23	30	40	47
y	4	10	13	15	15	25	27	46

x	55	67	72	81	96	112	127
y	38	46	53	70	82	99	100

a. Does a scatterplot of the data support the use of the simple linear regression model?

b. Calculate point estimates of the slope and intercept of the population regression line.

c. Calculate a point estimate of the true average runoff volume when rainfall volume is 50.

d. Calculate a point estimate of the standard deviation σ.

e. What proportion of the observed variation in runoff volume can be attributed to the simple linear regression relationship between runoff and rainfall?

18. A regression of y = calcium content (g/L) on x = dissolved material (mg/cm^2) was reported in the article "Use of Fly Ash or Silica Fume to Increase the Resistance of Concrete to Feed Acids" (*Mag. Concrete Res*. 1997: 337–344). The equation of the estimated regression line was y = 3.678 + .144x, with R^2 = .860, based on n = 23.

a. Interpret the estimated slope .144 and the coefficient of determination .860.

b. Calculate a point estimate of the true average calcium content when the amount of dissolved material is 50 mg/cm^2.

c. The value of total sum of squares was SST = 320.398. Calculate an estimate of the error standard deviation σ in the simple linear regression model.

19. The cetane number is a critical property in specifying the ignition quality of a fuel used in a diesel engine. Determination of this number for a biodiesel fuel is expensive and time-consuming. The article "Relating the Cetane Number of Biodiesel Fuels to Their Fatty Acid Composition: A Critical Study" (*J. Automobile Engr.* 2009: 565–583) included the following data on x = iodine value (g) and y = cetane number for a sample of 14 biofuels. The iodine value is the amount of iodine necessary to saturate a sample of 100 g of oil. The article's authors fit the simple linear regression model to this data, so let's follow their lead.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0

x	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	58.7	61.6	64.0	61.4	54.6	58.8	58.0

a. Obtain the equation of the least squares line, and then calculate a point prediction of the cetane number that would result from a single observation with an iodine value of 100.

b. Calculate and interpret the coefficient of determination.

c. Calculate and interpret a point estimate of the model standard deviation σ.

20. A number of studies have shown lichens (certain plants composed of an alga and a fungus) to be excellent bioindicators of air pollution. The article "The Epiphytic Lichen *Hypogymnia physodes* as a

Biomonitor of Atmospheric Nitrogen and Sulphur Deposition in Norway" (*Environ. Monit. Assess.* 1993: 27–47) gives the following data (read from a graph) on x = NO_3^- wet deposition (gN/m^2) and y = lichen N (% dry weight):

x	.05	.10	.11	.12	.31	.37	.42
y	.48	.55	.48	.50	.58	.52	1.02

x	.58	.68	.68	.73	.85	.92
y	.86	.86	1.00	.88	1.04	1.70

The author used simple linear regression to analyze the data. Use the accompanying Minitab output to answer the following questions:

a. What are the least squares estimates of β_0 and β_1?

b. Predict lichen N for an NO_3^- deposition value of .5.

c. What is the estimate of σ?

d. What is the value of total variation, and how much of it can be explained by the model relationship?

```
The regression equation is
lichen N = 0.365 + 0.967 No3depo

Predictor   Coef          Stdev      t ratio   P
Constant    0.36510       0.09904    3.69      0.004
No3depo     0.9668        0.1829     5.29      0.000
S = 0.1932  R-sq = 71.7%  R-sq (adj) = 69.2%

Analysis of Variance

Source      DF   SS       MS       F       P
Regression   1   1.0427   0.4106   27.94   0.000
Error       11   0.4106   0.0373
Total       11   1.4533
```

21. Visual and musculoskeletal problems associated with the use of visual display terminals (VDTs) have become rather common in recent years. Some researchers have focused on vertical gaze direction as a source of eye strain and irritation. This direction is known to be closely related to ocular surface area (OSA), so a method of measuring OSA is needed. The accompanying representative data on y = OSA (cm^2) and x = width of the palprebral fissure (i.e., the horizontal width of the eye

opening, in cm) is from the article "Analysis of Ocular Surface Area for Comfortable VDT Workstation Layout" (*Ergonomics* 1996: 877–884).

x	.40	.42	.48	.51	.57	.60	.70	.75
y	1.02	1.21	.88	.98	1.52	1.83	1.50	1.80

x	.75	.78	.84	.95	.99	1.03	1.12
y	1.74	1.63	2.00	2.80	2.48	2.47	3.05

x	1.15	1.20	1.25	1.25	1.28	1.30	1.34	1.37
y	3.18	3.76	3.68	3.82	3.21	4.27	3.12	3.99

x	1.40	1.43	1.46	1.49	1.55	1.58	1.60
y	3.75	4.10	4.18	3.77	4.34	4.21	4.92

   a. Construct a scatterplot of this data. Describe what you find.
   b. Calculate the equation of the LSRL.
   c. Interpret the slope of the LSRL.
   d. What OSA would you predict for a subject whose palprebal fissure width is 1.25 cm?
   e. What would be the estimate of expected OSA for people with palprebal fissure width of 1.25 cm?

22. For many years, rubber powder has been used in asphalt cement to improve performance. The article "Experimental Study of Recycled Rubber-Filled High-Strength Concrete" (*Mag. Concrete Res.* 2009: 549–556) included a regression of y = axial strength (MPa) on x = cube strength (MPa) based on the following sample data:

x	112.3	97.0	92.7	86.0	102.0
y	75.0	71.0	57.7	48.7	74.3

x	99.2	95.8	103.5	89.0	86.7
y	73.3	68.0	59.3	57.8	48.5

   a. Verify that a scatterplot supports the assumption that the two variables are related via the simple linear regression model.
   b. Obtain the equation of the least squares line, and interpret its slope.
   c. Calculate and interpret the coefficient of determination.

   d. Calculate and interpret an estimate of the error standard deviation σ in the simple linear regression model.
   e. The largest x value in the sample considerably exceeds the other x values. What is the effect on the equation of the least squares line of deleting the corresponding observation?

23. Show that under the assumptions of the simple linear regression model, the mles of β_0 and β_1 are identical to the least squares estimates. [*Hints*: (1) The pdf of Y_i is normal with mean $\mu_i = \beta_0 + \beta_1 x_i$ and variance σ^2; the likelihood function is the product of the n pdfs. (2) You don't need to differentiate the likelihood function; instead, find the correspondence between that function and the least squares expression $g(b_0, b_1)$.]

24. a. Show that the residuals $e_1, \ldots, e_n$ satisfy both $\sum e_i = 0$ and $\sum (x_i - \bar{x})e_i = 0$. [*Hint*: Use the last expression for e_i in (12.4), along with the fact that for any numbers $a_1, \ldots, a_n, \sum (a_i - \bar{a}) = 0$.]
   b. Show that $\hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$.
   c. Use (a) and (b) to derive the analysis of variance identity for regression, Equation (12.5), by showing that the cross-product term is 0.
   d. Use (b) and Equation (12.5) to verify the computational formula for SSE.

25. A regression analysis is carried out with y = temperature, expressed in °C. How do the resulting values of $\hat{\beta}_0$ and $\hat{\beta}_1$ relate to those obtained if y is re-expressed in °F? Justify your assertion. [*Hint*: (new y_i) = $1.8y_i + 32$.]

26. Show that b_1 and b_0 of Expressions (12.2) and (12.3) satisfy the normal equations.

27. Show that the "point of averages" $(\bar{x}, \bar{y})$ lies on the estimated regression line.

28. Suppose an investigator has data on the amount of shelf space x devoted to display of a particular product and sales revenue y for that product. The investigator may wish to fit a model for which the true

regression line passes through (0, 0). The appropriate model is $Y = \beta_1 x + \varepsilon$. Assume that $(x_1, y_1), \ldots, (x_n, y_n)$ are observed pairs generated from this model, and derive the least squares estimator of β_1. [*Hint*: Write the sum of squared deviations as a function of b_1, a trial value, and use calculus to find the minimizing value of b_1.]

29. a. Consider the data in Exercise 20. Suppose that instead of the least squares line that passes through the points $(x_1, y_1), \ldots, (x_n, y_n)$, we wish the least squares line that passes through $(x_1 - \bar{x}, y_1), \ldots, (x_n - \bar{x}, y_n)$. Construct a scatterplot of the (x_i, y_i) points and then of the $(x_i - \bar{x}, y_i)$ points. Use the plots to explain intuitively how the two least squares lines are related.

    b. Suppose that instead of the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $(i = 1, \ldots, n)$, we wish to fit a model of the form $Y_i = \beta_0^* + \beta_1^*(x_i - \bar{x}) + \varepsilon_i$ $(i = 1, \ldots, n)$. What are the least squares estimators of β_0^* and β_1^*, and how do they relate to $\hat{\beta}_0$ and $\hat{\beta}_1$?

30. Consider the following three data sets, in which the variables of interest are x = commuting distance and y = commuting time. Based on a scatterplot and the values of s_e and R^2, in which situation would simple linear regression be most (least) effective, and why?

	1		2		3	
	x	y	x	y	x	y
	15	42	5	16	5	8
	16	35	10	32	10	16
	17	45	15	44	15	22
	18	42	20	45	20	23
	19	49	25	63	25	31
	20	46	50	115	50	60
S_{xx}	17.50		1270.8333		1270.8333	
S_{xy}	29.50		2722.5		1431.6667	
$\hat{\beta}_1$	1.685714		2.142295		1.126557	
$\hat{\beta}_0$	13.666672		7.868852		3.196729	
SST	114.83		5897.5		1627.33	
SSE	65.10		65.10		14.48	

12.3 Inferences About the Regression Coefficient β_1

In virtually all of our inferential work thus far, the notion of sampling variability has been pervasive. In particular, properties of sampling distributions of various statistics ($\overline{X}$, $\hat{P}$, and so on) have been the basis for developing confidence interval formulas and hypothesis-testing methods. The key idea is that the value of virtually any quantity calculated from sample data (i.e., any statistic) is going to vary from one sample to another.

Example 12.8 Reconsider the data on x = truss height and y = sale price per square foot from $n = 19$ warehouses in Example 12.4 of the previous section. Suppose the simple linear regression model applies here, with parameter values $\beta_1 = 1$, $\beta_0 = 25$, and $\sigma = 1.4$ (consistent with the estimates computed previously). To understand the sampling variability of the statistics $\hat{\beta}_0$ and $\hat{\beta}_1$, we performed the following simulation 250 times in R:

- Generate random errors $\varepsilon_1, \ldots, \varepsilon_{19}$ from a normal distribution with mean 0 and standard deviation $\sigma = 1.4$.
- Using the 19 x_i's from the original data set, generate response values according to the model equation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = 25 + 1x_i + \varepsilon_i \qquad i = 1, 2, \ldots, 19$$

- Perform least squares regression on the simulated (x_i, y_i) pairs to obtain the estimated slope and intercept.

Figure 12.13 shows histograms of the $\hat{\beta}_0$ and $\hat{\beta}_1$ values resulting from this simulation. There is clearly variation in values of the estimated slope and estimated intercept. The equation of the LSRL thus also varies from one sample to the next. Note, though, that the estimates are centered close to the true values, an indication of unbiasedness.
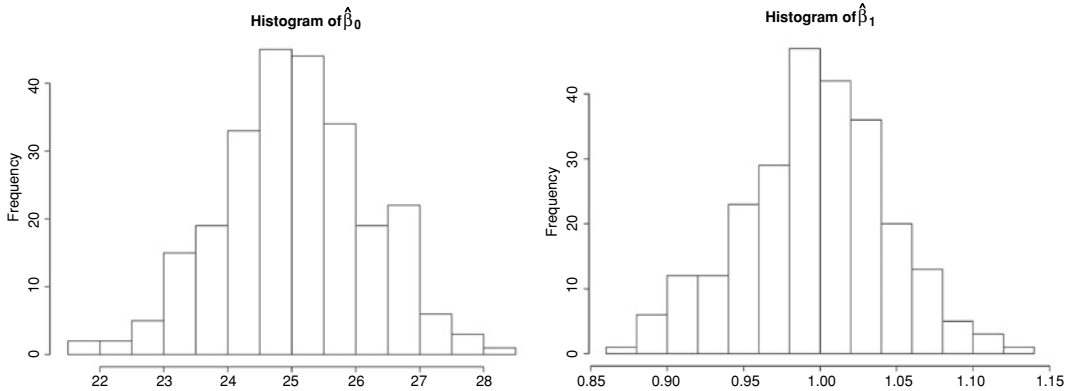
Figure 12.13 Histograms approximating the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ ∎

The slope β_1 of the population regression line is the true change in the mean of the response variable Y associated with a one-unit increase in the explanatory variable x. The slope of the least squares line, $\hat{\beta}_1$, gives a point estimate of β_1. In the same way that a confidence interval for μ and procedures for testing hypotheses about μ were based on properties of the sampling distribution of $\overline{X}$, inferences about β_1 are based on the sampling distribution of $\hat{\beta}_1$.

The values of the x_i's are assumed to be chosen before the study is performed, so only the Y_i's are random. The estimators for β_0 and β_1 are obtained by replacing y_i with Y_i in (12.2) and (12.3):

$$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(Y_i - \overline{Y})}{\sum (x_i - \overline{x})^2} = \frac{\sum (x_i - \overline{x})(Y_i - \overline{Y})}{S_{xx}}, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

Similarly, the estimator for σ results from replacing each y_i in the formula for s_e by the rv Y_i:

$$\hat{\sigma} = S_e = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum (Y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2}{n-2}}$$

The denominator of $\hat{\beta}_1$, $S_{xx} = \sum (x_i - \overline{x})^2$, depends only on the x_i's and not on the Y_i's, so it is a constant. Then, because $\sum \left[(x_i - \overline{x})\overline{Y} \right] = \overline{Y} \sum (x_i - \overline{x}) = \overline{Y} \cdot 0 = 0$, the slope estimator can be re-written as

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})Y_i}{S_{xx}} = \sum c_i Y_i \quad \text{where } c_i = (x_i - \bar{x})/S_{xx}$$

That is, the estimator $\hat{\beta}_1$ is a linear function of the independent rvs Y_1, Y_2, ..., Y_n, each of which is normally distributed. Invoking properties of a linear function of random variables discussed in Section 5.3 leads to the following results (Exercise 40).

PROPERTIES OF THE ESTIMATED SLOPE

1. The mean value of $\hat{\beta}_1$ is $E(\hat{\beta}_1) = \mu_{\hat{\beta}_1} = \beta_1$, so $\hat{\beta}_1$ is an unbiased estimator of β_1 (i.e., the distribution of $\hat{\beta}_1$ is always centered at the true value of β_1).

2. The variance and standard deviation of $\hat{\beta}_1$ are

$$V(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{S_{xx}} \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

Replacing σ by its estimate s_e gives an estimate for $\sigma_{\hat{\beta}_1}$:

$$\hat{\sigma}_{\hat{\beta}_1} = s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{S_{xx}}} = \frac{s_e}{s_x \sqrt{n-1}}$$

3. The estimator $\hat{\beta}_1$ has a normal distribution, because it is a linear function of independent normal rvs.

Properties 1 and 3 manifest themselves in the $\hat{\beta}_1$ histogram of Figure 12.13. According to Property 2, the standard deviation of $\hat{\beta}_1$ equals the standard deviation σ of the random error term—or, equivalently, of any Y_i—divided by $s_x \sqrt{n-1}$. Because the sample standard deviation s_x is a measure of how spread out the x_i's are about $\bar{x}$, we conclude that making observations at x_i values that are quite spread out results in a more precise estimator of the slope parameter (smaller variance of $\hat{\beta}_1$), whereas values of x_i all close to each other imply a highly variable estimator. Of course, if the x_i's are spread out too far, a linear model may not be appropriate throughout the range of observation. Finally, the presence of n in the denominator of $s_{\hat{\beta}_1}$ implies that the estimated slope varies less for larger samples than for smaller ones. We have seen this feature previously in other statistics such as $\bar{X}$ and $\hat{P}$: as sample size n increases, the distribution of the statistic "collapses onto" the true value of the corresponding parameter.

Many inferential procedures discussed previously were based on standardizing an estimator by first subtracting its mean value and then dividing by its estimated standard deviation. In particular, test procedures and a CI for the mean μ of a normal population utilized the fact that the standardized variable $(\bar{X} - \mu)/(S/\sqrt{n})$ has a t distribution with $n - 1$ df. A similar result here provides the key to further inferences concerning β_1.

THEOREM The assumptions of the simple linear regression model imply that the standardized variable

$$T = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{S_e/\sqrt{S_{xx}}}$$

has a t distribution with $n - 2$ df.

The T ratio can be written as

$$T = \frac{\hat{\beta}_1 - \beta_1}{S_e/\sqrt{S_{xx}}} = \frac{\dfrac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\dfrac{(n-2)S_e^2/\sigma^2}{(n-2)}}}$$

The theorem is a consequence of the following facts: (1) $(\hat{\beta}_1 - \beta_1)/(\sigma/\sqrt{S_{xx}}) \sim N(0,1)$, (2) $(n-2)S_e^2/\sigma^2 \sim \chi_{n-2}^2$, and (3) $\hat{\beta}_1$ is independent of S_e. That is, T is a standard normal rv divided by the square root of an independent chi-squared rv over its df, and so has the specified t distribution.

A Confidence Interval for β_1

As in the derivation of previous CIs, we begin with a probability statement:

$$P\left(-t_{\alpha/2,n-2} < \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} < t_{\alpha/2,n-2}\right) = 1 - \alpha$$

Manipulation of the inequalities inside the parentheses to isolate β_1 and substitution of estimates in place of the estimators gives the following CI formula.

A **100(1 − α)% CI for the slope β_1** of the true regression line has endpoints

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1} = \hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot \frac{s_e}{\sqrt{S_{xx}}} = \hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot \frac{s_e}{s_x\sqrt{n-1}}$$

This interval has the same general form as did many of our previous intervals. It is centered at the point estimate of the parameter, and the amount it extends out to each side of the estimate depends on the desired confidence level (through the t critical value) and on the amount of variability in the estimator $\hat{\beta}_1$ (through $s_{\hat{\beta}_1}$, which will tend to be small when there is little variability in the distribution of $\hat{\beta}_1$ and large otherwise).

Example 12.9 The scatterplot in Figure 12.14 shows the size (x, in square feet) and monthly rent (y, in dollars) for a random sample of $n = 77$ two-bedroom apartments in Omaha, NE (courtesy of www.zillow.com/omaha-ne/rentals). The plot suggests, not surprisingly, that rent generally increases with apartment size, and that for any fixed apartment size there is variability in monthly rents.
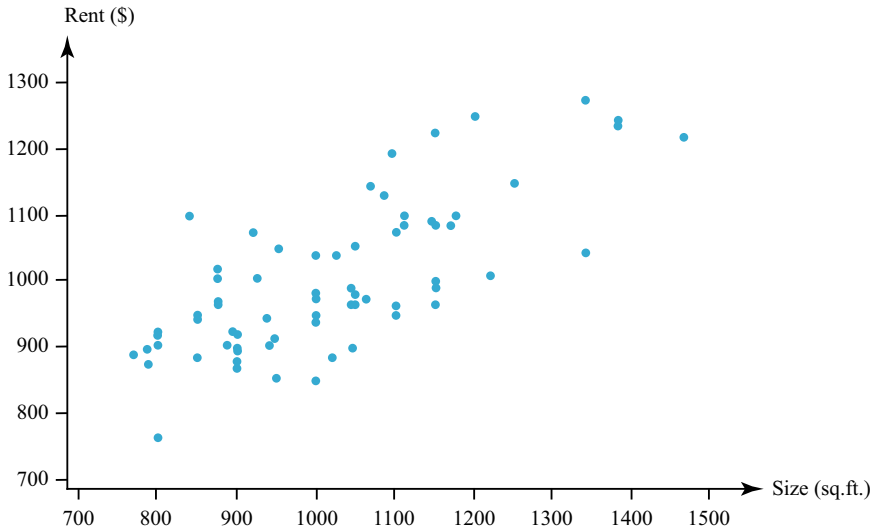
Figure 12.14 Scatterplot of the data from Example 12.8

Summary quantities include

$$\bar{x} = 1023.5 \qquad s_x = 161.9 \qquad S_{xx} = 1{,}991{,}569$$
$$\bar{y} = 1006.6 \qquad s_y = 113.3 \qquad S_{yy} = 975{,}840$$
$$S_{xy} = 1{,}042{,}971$$

from which $\hat{\beta}_1 = .5237$, $\hat{\beta}_0 = 470.6$, SST $= S_{yy} = 975{,}840$, SSE $= 429{,}642$, and $R^2 = .5597$. Roughly 56% of the observed variation in monthly rent can be attributed to the simple linear regression model relationship between rent and apartment size. The remaining 44% of rent variation is due to other apartment features, such as neighborhood, nicer appliances, or dedicated parking. Error df is $n - 2 = 77 - 2 = 75$, giving $s_e^2 = 429{,}642/75 = 5728.56$ and $s_e = 75.69$.

The estimated standard deviation of $\hat{\beta}_1$ is

$$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{S_{xx}}} = \frac{75.69}{\sqrt{1{,}991{,}569}} = .0536$$

The t critical value for a confidence level of 95% is $t_{.025,75} = 1.992$, so a 95% CI for β_1 is

$$.5237 \pm 1.992(.0536) = (.4169, .6305)$$

With a high degree of confidence, we estimate that a one square foot increase in an apartment's size is associated with an increase between $.4169 and $.6305 in the expected monthly rent. This applies to the population of all two-bedroom apartments in Omaha. Multiplying by 100 gives $41.69 to $63.05 as the increase in expected rent associated with a 100 ft^2 size increase.

Looking at the R output of Figure 12.15, we find the value of $s_{\hat{\beta}_1}$ under coefficients as the second number in the standard error column, while the value of s_e is displayed as residual standard error. There is also an estimated standard error for the statistic $\hat{\beta}_0$. For all of the statistics, compare the values in the R output with the values calculated above.

```
Call:
lm(formula = Rent ~ Size, data = Omaha)

Residuals:
     Min        1Q   Median        3Q      Max
-144.314   -52.306    1.658    40.635  189.477

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 470.62001   55.56499   8.470 1.52e-12 ***
Size          0.52369    0.05363   9.765 5.30e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75.69 on 75 degrees of freedom
Multiple R-squared:  0.5597,        Adjusted R-squared:  0.5539
F-statistic: 95.35 on 1 and 75 DF,  p-value: 5.302e-15
```

Figure 12.15 R output for the data of Example 12.9   ∎

Hypothesis-Testing Procedures

As before, the null hypothesis in a test about β_1 will be an equality statement. The null value of β_1 claimed true by the null hypothesis will be denoted by β_{10} (read "beta one naught," *not* "beta ten"). The test statistic results from replacing β_1 in the standardized variable T by the null value β_{10}—that is, from standardizing $\hat{\beta}_1$ under the assumption that H_0 is true. The test statistic thus has a t distribution with $n - 2$ df when H_0 is true, so the type I error probability is controlled at the desired level α by using an appropriate t critical value.

Null hypothesis: $H_0: \beta_1 = \beta_{10}$

Test statistic value: $t = \dfrac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
$H_a: \beta_1 > \beta_{10}$	$t \geq t_{\alpha, n-2}$
$H_a: \beta_1 < \beta_{10}$	$t \leq -t_{\alpha, n-2}$
$H_a: \beta_1 \neq \beta_{10}$	either $t \geq t_{\alpha/2, n-2}$ or $t \leq -t_{\alpha/2, n-2}$

A P-value based on $n - 2$ df can be calculated just as was done previously for t tests in Chapters 9 and 10.

The most commonly encountered pair of hypotheses about β_1 is $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, in which case the test statistic value is the t ratio $\hat{\beta}_1/s_{\hat{\beta}_1}$. When this null hypothesis is true, $E(Y|x) = \beta_0 + 0x = \beta_0$, independent of x, so knowledge of x gives no information about the value of the response variable. A test of these two hypotheses is often referred to as the **model utility test** in simple linear regression.

Unless n is quite small, H_0 will be rejected and the utility of the model confirmed precisely when R^2 is reasonably large. The simple linear regression model should not be used for further inferences, such as estimates of mean value or predictions of future values (the topics of Section 12.4), unless the model utility test results in rejection of H_0 for a suitably small α.

Example 12.10 How is the perceived risk of an investment related to its expected return? Intuitively it would seem as though riskier investments would be associated with higher expected returns. The article "Affect in a Behavioral Asset-Pricing Model" (*Fin. Anal. J.*, March/April 2008: 20–29) reported on an experiment in which each member of a group of investors rated the risk of a company's stock on a 10-point scale ranging from low to high, and members of a different group rated the future return of the stock on the same scale. This was done for a total of 210 companies, and for each one both a risk score x and an expected return score y resulted from averaging responses from the individual raters. The following data is from a subset of ten of these companies (listed for convenience in increasing order of risk):

x	4.3	4.6	5.2	5.3	5.5	5.7	6.1	6.3	6.8	7.5
y	7.7	5.2	7.9	5.8	7.2	7.0	5.3	6.8	6.6	4.7

The scatterplot of the data for these ten companies in Figure 12.16 shows a weak ($R^2 \approx .18$) but also surprising negative relationship between the two variables. Let's carry out the model utility test at significance level $\alpha = .05$ (the scatterplot does not bode well for the model, but stay tuned for the rest of the story).
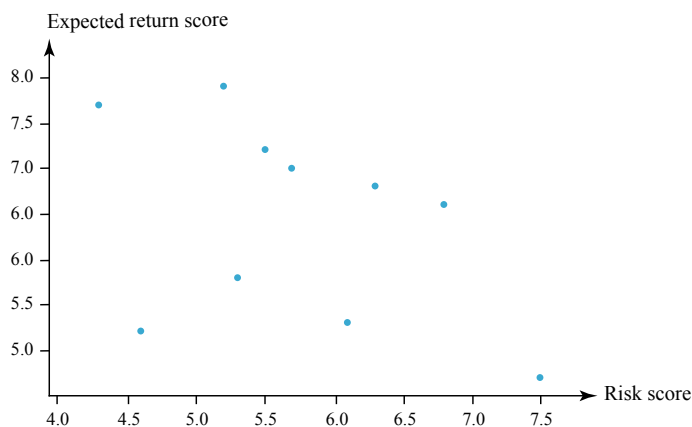
Figure 12.16 Scatterplot for the data in Example 12.10

The parameter of interest is β_1 = true change in expected return score associated with a one-point increase in risk score. The null hypothesis H_0: $\beta_1 = 0$ will be rejected in favor of H_a: $\beta_1 \neq 0$ if the observed test statistic t satisfies either $t \geq t_{\alpha/2, n-2} = t_{.025,8} = 2.306$ or $t \leq -2.306$. Partial Excel output (software not favorably regarded in the statistical community) for this example appears in Figure 12.17. In the output, $\hat{\beta}_1 = -.4913$ and $s_{\hat{\beta}_1} = .3614$, so the test statistic is

$$t = \frac{-.4913 - 0}{.3614} \approx -1.36 \quad \text{(also on output)}$$

	df	SS	MS	F	Significance	
Regression	1	2.0714	2.0714	1.8485	0.2110	
Residual	8	8.9646	1.1206			
Total	9	11.0360				

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	9.2353	2.0975	4.4029	0.0023	4.3983	14.0722
Risk	-0.4913	0.3614	-1.3596	0.2110	-1.3246	0.3420

Figure 12.17 Partial Excel output for Example 12.10

Since -1.36 does not fall in the rejection region, H_0 is not rejected at the .05 level. Equivalently, the two-sided P-value is double the area under the t_8 curve to the left of -1.36, which Excel reports as roughly .211; since .211 > .05, again H_0 is not rejected.

Excel also provides a 95% confidence interval of $(-1.3246, 0.3420)$ for β_1. This is consistent with the results of the hypothesis test: since the value 0 lies in the interval, we have no reason to reject the claim H_0: $\beta_1 = 0$.

Is there truly no relationship, or have we committed a type II error here? With just $n = 10$ observations, it is quite possible we failed to detect a relationship because hypothesis tests do not have much power when the sample size is small. In fact, the authors of the original study examined 210 companies on these same two variables, resulting in an estimated slope of $\hat{\beta}_1 = -0.4$ (similar to our sample) but with an estimated standard error of roughly .0556. The resulting test statistic value is $t = -7.2$ at 208 df, which is highly statistically significant. The authors concluded, based on their larger sample, that risk *is* a useful predictor of future return—although, contrary to intuition, the association between the two appears to be negative. Even though the relationship was statistically significant, note that—even in the full sample—risk only accounted for $R^2 = .185 = 18.5\%$ of the variation in future returns. As in the case of previous test procedures, a large-sample size can result in rejection of H_0 even though the data suggests that the departure from H_0 is of little practical significance. ∎

Regression and ANOVA

The splitting of the total sum of squares SST $= S_{yy} = \sum (y_i - \overline{y})^2$ into a part SSE which measures unexplained variation and a part SSR which measures variation explained by the linear relationship is strongly reminiscent of one-way ANOVA. In fact, H_0: $\beta_1 = 0$ can alternatively be tested against H_a: $\beta_1 \neq 0$ by constructing an ANOVA table (Table 12.1) and rejecting H_0 if $f \geq F_{\alpha,1,n-2}$.

Table 12.1 ANOVA table for simple linear regression

Source of variation	df	Sum of squares	Mean square	F ratio
Regression	1	SSR	MSR = SSR/1	f = MSR/MSE
Error	$n - 2$	SSE	MSE = SSE/$(n - 2)$	
Total	$n - 1$	SST		

The square root of the mean squared error (MSE) is s_e, the residual standard deviation. The F test gives exactly the same result as the model utility t test because $t^2 = f$ and $t^2_{\alpha/2,n-2} = F_{\alpha,1,n-2}$. Virtually all computer packages that have regression options include such an ANOVA table in the output. For example, Figure 12.17 shows ANOVA output for the data of Example 12.10. The ANOVA table at the top of the output has $f = 1.8485$ with a P-value of .211 for the model utility test. The table of parameter estimates gives $t = -1.3596$, again with $P = .211$, and $t^2 = (-1.3596)^2 = 1.8485 = f$. Note that this F test is only for $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$; if the alternative hypothesis is one-sided (> or <) or if the null value β_{10} is not 0, then the t test must be used.

Exercises: Section 12.3 (31–42)

31. Reconsider the situation described in Example 12.1, in which $x =$ firing velocity of a 7.62-mm round and $y =$ body armor penetration area. Suppose the simple linear regression model is valid for x between 650 and 800 m/s, and that $\beta_1 = .25$ and $\sigma = 10$ mm. Consider an experiment in which $n = 7$, and the x values at which observations are made are $x_1 = 650$, $x_2 = 675$, $x_3 = 700$, $x_4 = 725$, $x_5 = 750$, $x_6 = 775$, and $x_7 = 800$.

   a. Calculate $\sigma_{\hat{\beta}_1}$, the standard deviation of $\hat{\beta}_1$.

   b. What is the probability that the estimated slope based on such observations will be between .15 and .35?

   c. Suppose it is also possible to make a single observation at each of the $n = 11$ values 675, 685, 695, 705, ..., 775. If a major objective is to estimate β_1 as precisely as possible, would the experiment with $n = 11$ be preferable to the one with $n = 7$?

32. Exercise 16 of Section 12.2 included Minitab output for a regression of daughter's height on the midparent height.

   a. Use the output to calculate a confidence interval with a confidence level of 95% for the slope β_1 of the population regression line, and interpret the resulting interval.

   b. Suppose it had previously been believed that when midparent height increased by 1 in., the associated true average change in the daughter's height would be at least 1 in. Does the sample data contradict this belief? State and test the relevant hypotheses.

33. Exercise 17 of Section 12.2 gave data on $x =$ rainfall volume and $y =$ runoff volume (both in m^3). Use the accompanying Minitab output to decide whether there is a useful linear relationship between rainfall and runoff, and then calculate a confidence interval for the true average change in runoff volume associated with a 1-m^3 increase in rainfall volume.

```
The regression equation is
runoff = - 1.13 + 0.827 rainfall

Predictor   Coef      Stdev      t ratio   P
Constant    -1.128    2.368      -0.48     0.642
Rainfall    0.82697   0.03652    22.64     0.000
s = 5.240   R-sq = 97.5%      R-sq(adj) = 97.3%
```

34. The invasive diatom species *D. Geminata* has the potential to inflict substantial ecological and economic damage in rivers. The article "Substrate Characteristics Affect Colonization by the Bloom-Forming Diatom *Didymosphenia Geminata*" (*Aquat. Ecol.* 2010: 33–40) described an investigation of colonization behavior. One aspect of particular interest was whether $y =$ colony density was related to $x =$ rock surface area. The article contained a scatterplot and summary of a regression analysis. Here is representative data:

x	50	71	55	50	33	58	79	
y	152	1929	48	22	2	5	35	
x	26	69	44	37	70	20	45	49
y	7	269	38	171	13	43	185	25

a. Fit the simple linear regression model to this data, and then calculate and interpret the coefficient of determination.

b. Carry out a test of hypotheses to determine whether there is a useful linear relationship between density and rock area.

c. The second observation has a very extreme y value (in the full data set consisting of 72 observations, there were two of these). This observation may have had a substantial impact on the fit of the model and subsequent conclusions. Eliminate it and redo parts (a) and (b). What do you conclude?

35. How does lateral acceleration—side forces experienced in turns that are largely under driver control—affect nausea as perceived by bus passengers? The article "Motion Sickness in Public Road Transport: The Effect of Driver, Route, and Vehicle" (*Ergonomics* 1999: 1646–1664) reported data on x = motion sickness dose (calculated in accordance with a British standard for evaluating similar motion at sea) and y = reported nausea (%). Relevant summary quantities are

$$n = 17, \ \sum x_i = 222.1, \ \sum y_i = 193.0,$$
$$S_{xx} = 155.02, \ S_{yy} = 783.88, \ S_{xy} = 238.11$$

Values of dose in the sample ranged from 6.0 to 17.6.

a. Assuming that the simple linear regression model is valid for relating these two variables (this is supported by the raw data), calculate and interpret an estimate of the slope parameter that conveys information about the precision and reliability of estimation.

b. Does it appear that there is a useful linear relationship between these two variables? Answer the question by employing the *P*-value approach.

c. Would it be sensible to use the simple linear regression model as a basis for predicting % nausea when dose = 5.0? Explain your reasoning.

36. Mist (airborne droplets or aerosols) is generated when metal-removing fluids are used in machining operations to cool and lubricate the tool and workpiece. Mist generation is a concern to OSHA, which has substantially lowered the workplace standard. The article "Variables Affecting Mist Generation from Metal Removal Fluids" (*Lubricat. Engr.* 2002: 10–17) gave the accompanying data on x = fluid flow velocity for a 5% soluble oil (cm/s) and y = the extent of mist droplets having diameters smaller than 10 μm (mg/m^3):

x	89	177	189	354	362	442	965
y	.40	.60	.48	.66	.61	.69	.99

a. The investigators performed a simple linear regression analysis to relate the two variables. Does a scatterplot of the data support this strategy?

b. What proportion of observed variation in mist can be attributed to the simple linear regression relationship between velocity and mist?

c. The investigators were particularly interested in the impact on mist of increasing velocity from 100 to 1000 (a factor of 10 corresponding to the difference between the smallest and largest x values in the sample). When x increases in this way, is there substantial evidence that the true average increase in y is less than .6?

d. Estimate the true average change in mist associated with a 1 cm/s increase in velocity, and do so in a way that conveys information about precision and reliability.

37. Refer to the data on $x =$ iodine value and $y =$ cetane number given in Exercise 19.

   a. Does the simple linear regression model specify a useful relationship between the two variables? Use the appropriate test procedure to obtain information about the P-value and then reach a conclusion at significance level .01.

   b. Compute a 95% CI for the expected change in cetane number associated with a 10 g increase in iodine value.

38. Carry out the model utility test using the ANOVA approach for the filtration rate–moisture content data of Example 12.6. Verify that it gives a result equivalent to that of the t test.

39. Use the rules of expected value to show that $\hat{\beta}_0$ is an unbiased estimator for β_0 (making use of the fact that $\hat{\beta}_1$ is unbiased for β_1).

40. a. Verify that $E(\hat{\beta}_1) = \beta_1$ by using the rules of expected value from Chapter 5.

   b. Use the rules of variance from Chapter 5 to verify the expression for $V(\hat{\beta}_1)$ given in this section.

41. Verify that if each x_i is multiplied by a positive constant c and each y_i is multiplied by another positive constant d, the t statistic for testing $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$ is unchanged in value. [*Note:* The value of $\hat{\beta}_1$

will change, which shows that the magnitude of $\hat{\beta}_1$ is not by itself indicative of model utility.]

42. The power for the t test for $H_0: \beta_1 = \beta_{10}$ can be computed in the same manner as it was computed for the t tests of Chapter 9, using the noncentral t distribution. If the alternative value of β_1 is denoted by β_1', the required noncentrality parameter is

$$\delta = \frac{\beta_1' - \beta_{10}}{\sigma / \sqrt{S_{xx}}}$$

and power is calculated based on $n - 2$ df. An article in the *Journal of Public Health Engineering* reports the results of a regression analysis based on $n = 15$ observations in which $x =$ filter application temperature (°C) and $y = \%$ efficiency of BOD removal. (BOD stands for biochemical oxygen demand, and it is a measure of organic matter in sewage.) Calculated quantities include $S_{xx} = 324.4$, $s_e = 3.725$, and $\hat{\beta}_1 = 1.7035$. Consider testing at significance level .01 the hypothesis $H_0: \beta_1 = 1$, which states that the expected increase in % BOD removal is 1 when filter application temperature increases by 1 °C, against the alternative $H_a: \beta_1 > 1$. Determine power when $\beta_1' = 2$, $\sigma = 4$.

12.4   Inferences for the (Mean) Response

Throughout this section we will let $\hat{Y}$ denote the statistic

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

with observed value $\hat{y}$, where x^* denotes a specified value of the explanatory variable x. Once the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have been calculated, $\hat{y}$ can be regarded either as a point estimate of $\mu_{Y|x^*}$ (the expected or true average value of Y when $x = x^*$) or as a prediction of the Y value that will result from a single new observation made when $x = x^*$. The point estimate or prediction by itself gives no information concerning how precisely $\mu_{Y|x^*}$ has been estimated or Y has been predicted. This can be remedied by developing a CI for $\mu_{Y|x^*}$ and a prediction interval (PI) for a single future Y value.

Before we obtain sample data, both $\hat{\beta}_0$ and $\hat{\beta}_1$ are subject to sampling variability—that is, they are both statistics whose values will vary from sample to sample. This variability was shown in Figure 12.13 at the beginning of Section 12.3. Suppose, for example, that $\beta_0 = 50$ and $\beta_1 = 2$. Then a first sample of (x, y) pairs might give $\hat{\beta}_0 = 52.35$ and $\hat{\beta}_1 = 1.895$, a second sample might result in $\hat{\beta}_0 = 46.52$ and $\hat{\beta}_1 = 2.056$, and so on. It follows that $\hat{Y}$ itself varies in value from sample to sample. If the intercept and slope of the population line are the aforementioned values 50 and 2, respectively, and $x^* = 10$, then this statistic is trying to estimate the value $\mu_{Y|10} = 50 + 2(10) = 70$. The estimate from a first sample might be $\hat{y} = 52.35 + 1.895(10) = 71.30$, from a second sample might be $\hat{y} = 46.52 + 2.056(10) = 67.08$, etc. In the same way that a confidence interval for β_1 was based on properties of the sampling distribution of $\hat{\beta}_1$, a confidence interval for a mean y value in regression is based on properties of the sampling distribution of the statistic $\hat{Y}$.

Substitution of the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ into $\hat{Y}$, followed by some algebraic manipulation, leads to the representation of $\hat{Y}$ as a linear function of the Y_i's:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^* = \cdots = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{S_{xx}} \right] Y_i = \sum_{i=1}^{n} d_i Y_i$$

$$\text{where } d_i = (1/n) + (x^* - \bar{x})(x_i - \bar{x})/S_{xx}$$

The coefficients $d_1, d_2, \ldots, d_n$ in this linear function involve the x_i's and x^*, all of which are fixed. Application of the rules of Section 5.3 to this linear function gives the following properties. (Exercise 52 requests a derivation of Property 2.)

SAMPLING DISTRIBUTION OF $\hat{Y}$

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$, where x^* is some fixed value of x. Then

1. The mean value of $\hat{Y}$ is

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \beta_0 + \beta_1 x^* = \mu_{Y|x^*}$$

Thus $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is an unbiased estimator for $\beta_0 + \beta_1 x^*$ (i.e., for $\mu_{Y|x^*}$).

2. The variance of $\hat{Y}$ is

$$V(\hat{Y}) = \sigma_{\hat{Y}}^2 = \sigma^2 \cdot \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$$

and the standard deviation $\sigma_{\hat{Y}}$ is the square root of this expression. The estimated standard deviation of $\hat{Y}$, denoted by $s_{\hat{Y}}$ or $s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}$, results from replacing σ by its estimate s_e:

$$s_{\hat{Y}} = s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

3. $\hat{Y}$ has a normal distribution, because it is a linear function of the Y_i's which are normally distributed and independent.

The variance of $\hat{Y}$ is smallest when $x^* = \bar{x}$ and increases as x^* moves away from $\bar{x}$ in either direction. Thus the estimator of $\mu_{Y|x^*}$ is more precise when x^* is near the center of the x_i's than when it is far from the x values where observations have been made. This implies that both the CI and PI are narrower for an x^* near $\bar{x}$ than for an x^* far from $\bar{x}$. Most statistical computer packages provide both $\hat{Y}$ and $s_{\hat{Y}}$ for any specified x^* upon request.

Inferences Concerning the Mean Response

Just as inferential procedures for β_1 were based on the t variable obtained by standardizing, a t variable obtained by standardizing $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ leads to a CI and test procedures here.

THEOREM The variable

$$T = \frac{\hat{Y} - \mu_{Y|x^*}}{S_{\hat{Y}}} = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} \tag{12.6}$$

has a t distribution with $n - 2$ df.

As was the case for β_1 in the previous section, a probability statement involving this standardized variable can be manipulated to yield a confidence interval for $\mu_{Y|x^*}$.

A **$100(1 - \alpha)\%$ CI for $\mu_{Y|x^*}$**, the mean/expected value of Y when $x = x^*$, has endpoints

$$\hat{y} \pm t_{\alpha/2, n-2} \cdot s_{\hat{Y}} = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2, n-2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \tag{12.7}$$

This CI is centered at the point estimate for $\mu_{Y|x^*}$ and extends out to each side by an amount that depends on the confidence level and on the extent of variability in the estimator on which the point estimate is based.

Example 12.11 Refer back to the Omaha apartment data of Example 12.9, where the response variable was monthly rent and the predictor was square footage. Let's now calculate a confidence interval, using a 95% confidence level, for the mean rent for all 1200 ft.[2] two-bedroom apartments in Omaha—that is, a confidence interval for $\mu_{Y|1200} = \beta_0 + \beta_1(1200)$. The interval is centered at

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(1200) = 470.6 + .5237(1200) = \$1099.04$$

The estimated standard deviation of the statistic $\hat{Y}$ at $x = x^* = 1200$ is

$$s_{\hat{Y}} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = 75.69 \sqrt{\frac{1}{77} + \frac{(1200 - 1023.5)^2}{1,991,569}} = 12.807$$

The 75 df t critical value for a 95% confidence level is 1.992, from which we determine the desired interval to be

$$1099.04 \pm 1.992(12.807) = (1073.54, 1124.57)$$

At the 95% confidence level, we estimate that the average monthly rent of all 1200 square foot, two-bedroom apartments in Omaha is between \$1073.54 and \$1124.57. Remember that if we re-calculated this interval for sample after sample, in the long run about 95% of the calculated intervals would include $\beta_0 + \beta_1(1200)$. We hope that this true mean value lies in the single interval that we have calculated.

For the population of all two-bedroom apartments in Omaha of size 1050 square feet, similar calculations result in $\hat{y} = 1020.50$, $s_{\hat{Y}} = 8.742$, and 95% CI = (1003.08, 1037.91). Notice that not only is the expected rent lower for 1050 ft.^{2} apartments than for 1200 ft.^{2} apartments (no surprise there), but the estimated standard error is also smaller. That's because $x^* = 1050$ is closer to the sample mean of $\bar{x} = 1023.5$ square feet than is $x^* = 1200$.

Figure 12.18 shows a JMP scatterplot with the LSRL and curves corresponding to the confidence limits for each different x value.
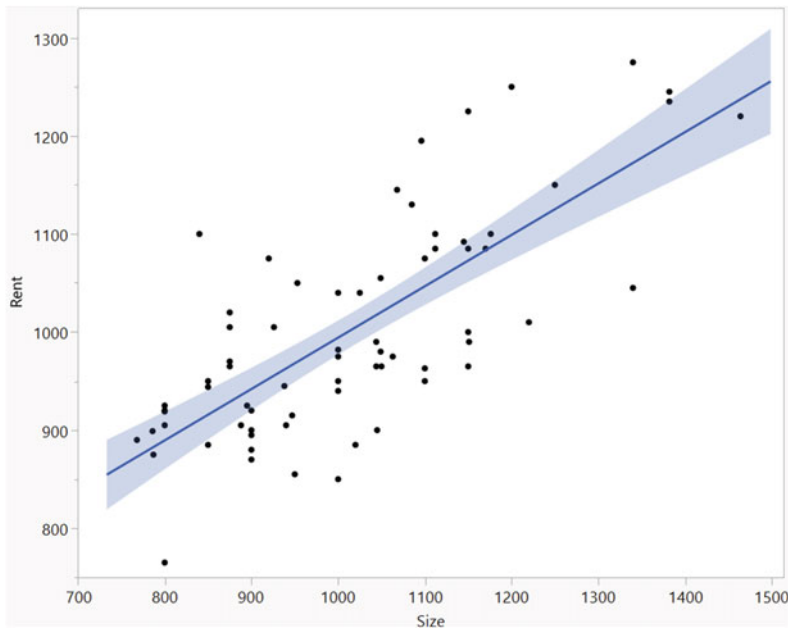
Figure 12.18   JMP scatterplot with confidence limits for the data of Example 12.11   ∎

In some situations, a CI is desired not just for a single x value but for two or more x values, and we must proceed with caution in interpreting the confidence levels of our intervals. For example, in Example 12.11 two 95% CIs were constructed, one for $\mu_{Y|1200}$ and another for $\mu_{Y|1050}$. The *joint* or *simultaneous* confidence level—the long-run proportion of time under repeated sampling that *both* CIs would contain their respective parameters—is *less* than 95%. While it is difficult to determine the exact simultaneous confidence level, Bonferroni's inequality (Chapter 8, Exercise 91) established that if two $100(1 - \alpha)$% CI's are computed, then the joint confidence level of the resulting pair of intervals is at least $100(1 - 2\alpha)$%. Thus, in Example 12.11 we are at least 90% confident (because $\alpha = .05$ and $1 - 2\alpha = 1 - .10 = .90$) that the two statements $1073.54 < \mu_{Y|1200} < 1124.57$ and $1003.08 < \mu_{Y|1050} < 1037.91$ are both true.

More generally, a set of m intervals each with confidence level $100(1 - \alpha)\%$ is guaranteed to have simultaneous confidence of at least $100(1 - m\alpha)\%$. This relationship can be reversed to achieve a desired joint confidence level: replacing α with α/m, if individual $100(1 - \alpha/m)\%$ CIs are constructed for each of m parameters, the resulting simultaneous confidence level is at least $100(1 - \alpha)\%$. For example, if we desired intervals for both $\mu_{Y|1200}$ and $\mu_{Y|1050}$ with (at least) 95% joint confidence, then each individual CI should be calculated using the t critical value corresponding to confidence coefficient $1 - \alpha/m = 1 - .05/2 = .975$.

Tests of hypotheses about $\mu_{Y|x^*}$ are based on the test statistic T obtained by replacing $\mu_{Y|x^*}$ with the null value μ_0 in the numerator of (12.6). For example, the assertion H_0: $\mu_{Y|1200} = \$1100$ in Example 12.11 says that the mean rent for all 1200 ft.^{2} apartments in the population is \$1100 per month. The test statistic value is then $t = (\hat{y} - 1100)/s_{\hat{Y}}$, and the test is upper-, lower-, or two-tailed according to the inequality in H_a.

A Prediction Interval for a Future Value of Y

Analogous to the CI (12.7) for $\mu_{Y|x^*}$, one frequently wishes to obtain an interval of plausible values for the value of Y associated with a *single* future observation when the explanatory variable has value x^*. In Example 12.11, a CI was computed for the true mean rent of all apartments of a certain size, but an individual renter will be more interested in knowing a realistic range of rent values for a single such apartment.

A CI estimates a parameter, or population characteristic, whose value is fixed but unknown to us. In contrast, a future value of Y is not a parameter but instead a random variable; for this reason we refer to an interval of plausible values for a future Y as a **prediction interval (PI)** rather than a confidence interval. (Section 8.2 presented a method for constructing a one-sample t prediction interval for a single future value of a variable.)

When estimating $\mu_{Y|x^*}$, the *estimation error*, $\hat{Y} - \mu_{Y|x^*}$, is the difference between a random variable and a fixed but unknown quantity. In contrast, the *prediction error* is $\hat{Y} - Y = \hat{Y} - (\beta_0 + \beta_1 x^* + \varepsilon)$, a difference between two random variables. With the additional random ε term, there is more uncertainty in prediction than in estimation. As a consequence, a PI will be wider than a CI for the same x^* value. Because the future value Y is independent of the observed Y_i's that determine $\hat{Y}$,

$$
\begin{aligned}
V(\hat{Y} - Y) &= \text{variance of prediction error} \\
&= V(\hat{Y}) + V(Y) \quad \text{independence} \\
&= V(\hat{Y}) + V(\varepsilon) \quad \text{because } \beta_0 + \beta_1 x^* \text{ is a constant} \\
&= \sigma^2 \cdot \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right] + \sigma^2 \\
&= \sigma^2 \cdot \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]
\end{aligned}
$$

Furthermore, because $E(Y) = \beta_0 + \beta_1 x^*$ and $E(\hat{Y}) = \beta_0 + \beta_1 x^*$, the expected value of the prediction error is $E(\hat{Y} - Y) = 0$. It can then be shown that the standardized variable

$$
T = \frac{Y - \hat{Y}}{S_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}
$$

has a t distribution with $n - 2$ df. Substituting this expression for T into the probability statement $P(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1 - \alpha$ and manipulating to isolate Y between the two inequalities yields the following interval.

A $\mathbf{100(1 - \alpha)\%}$ **PI for a future** Y **observation to be made when** $x = x^*$ **has endpoints**

$$\hat{y} \pm t_{\alpha/2,n-2} \cdot \sqrt{s_e^2 + s_{\hat{Y}}^2} = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2,n-2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} \qquad (12.8)$$

The interpretation of the prediction level $100(1 - \alpha)\%$ is identical to that of previous confidence levels—if (12.8) is used repeatedly, in the long run the resulting intervals will actually contain the observed future y values $100(1 - \alpha)\%$ of the time. Notice that the 1 underneath the square root symbol makes the PI (12.8) wider than the CI (12.7), although the intervals are both centered at $\hat{y}$. Also, as $n \to \infty$ the width of the CI approaches 0, whereas the width of the PI approaches $2z_{\alpha/2}\sigma$ (because even with perfect knowledge of β_0 and β_1, there will still be uncertainty in *prediction*).

Example 12.12 (Example 12.11 continued) Let's calculate a 95% prediction interval for the monthly rent of a single 1200 square foot, two-bedroom apartment in Omaha. Relevant quantities from Example 12.11 are

$$\hat{y} = 1099.04 \quad s_{\hat{Y}} = 12.807 \quad s_e = 75.69 \quad t_{.025,75} = 1.992$$

The prediction interval is then

$$1099.04 \pm 1.992\sqrt{75.69^2 + 10.29^2} = 1099.04 \pm 1.992(76.386) = (946.13, 1251.97)$$

Plausible values for the monthly rent of a 1200 ft.^{2} apartment are, at the 95% prediction level, between \$946.13 and \$1251.97. The 95% confidence interval for the mean rent of all such apartments was (1073.54, 1124.57). The prediction interval is much wider than this because of the extra 75.69^2 under the square root. Since apartments of the same size will vary in rent—the estimated sd of rents for apartments of any fixed size is $s_e = \$75.69$—there is necessarily much greater uncertainty in the cost of a single apartment than in the average cost of all such apartments.   ∎

The Bonferroni technique can be employed as in the case of confidence intervals. If a PI with prediction level $100(1 - \alpha/m)\%$ is calculated at each of m different x^* values, the simultaneous or joint prediction level for all m intervals will be at least $100(1 - \alpha)\%$.

Exercises: Section 12.4 (43–52)

43. Global warming is a major issue, and CO_2 emissions are an important part of the discussion. The article "Effects of Atmospheric CO_2 Enrichment on Biomass Accumulation and Distribution in Eldarica Pine Trees" (*J. Exp. Bot.* 1994: 345–349) describes the results of growing pine trees with increasing levels of CO_2 in the air. Here are the observations with x = atmospheric concentration of CO_2 (parts per million) and y = mass in kilograms after 11 months of the experiment.

x	408	408	554	554	680	680	812	812
y	1.1	1.3	1.6	2.5	3.0	4.3	4.2	4.7

Software calculates $s_e = .534$; $\hat{y} = 2.723$ and $s_{\hat{Y}} = .190$ when $x = 600$; and $\hat{y} = 3.992$ and $s_{\hat{Y}} = .256$ when $x = 750$.

a. Explain why $s_{\hat{Y}}$ is larger when $x = 750$ than when $x = 600$.
b. Calculate a confidence interval with a confidence level of 95% for the true average mass of all trees grown with a CO_2 concentration of 600 parts per million.
c. Calculate a prediction interval with a prediction level of 95% for the mass of a tree grown with a CO_2 concentration of 600 parts per million.
d. If a 95% CI is calculated for the true average mass when CO_2 concentration is 750, what will be the simultaneous confidence level for both this interval and the interval calculated in part (b)?

44. Reconsider the filtration rate–moisture content data introduced in Example 12.6.
a. Compute a 90% CI for $\beta_0 + 125\beta_1$, true average moisture content when the filtration rate is 125.
b. Predict the value of moisture content for a single experimental run in which the filtration rate is 125 using a 90% prediction level. How does this interval compare to the interval of part (a)? Why is this the case?
c. How would the intervals of parts (a) and (b) compare to a CI and PI when filtration rate is 115? Answer without actually calculating these new intervals.
d. Interpret both $H_0: \beta_0 + 125\beta_1 = 80$ and $H_a: \beta_0 + 125\beta_1 < 80$, and then carry out a test at significance level .01.

45. Astringency is the quality in a wine that makes the wine drinker's mouth feel slightly rough, dry, and puckery. The paper "Analysis of Tannins in Red Wine Using Multiple Methods: Correlation with Perceived Astringency" (*Amer. J. Enol. Vitic.* 2006: 481–485) reported on an investigation to assess the relationship between perceived astringency and tannin concentration using various analytic methods. Here is data provided by the authors on $x =$ tannin concentration by protein precipitation and $y =$ perceived astringency as determined by a panel of tasters.

x	0.718	0.808	0.924	1.000	0.667	0.529	0.514	0.559
y	0.428	0.480	0.493	0.978	0.318	0.298	−0.224	0.198

x	0.766	0.470	0.726	0.762	0.666	0.562	0.378	0.779
y	0.326	−0.336	0.765	0.190	0.066	−0.221	−0.898	0.836

x	0.674	0.858	0.406	0.927	0.311	0.319	0.518	0.687
y	0.126	0.305	−0.577	0.779	−0.707	−0.610	−0.648	−0.145

x	0.907	0.638	0.234	0.781	0.326	0.433	0.319	0.238
y	1.007	−0.090	−1.132	0.538	−1.098	−0.581	−0.862	−0.551

Relevant summary quantities are as follows:

$$\sum x_i = 19.404, \quad \sum y_i = -.549, \quad S_{xx} = 1.48193150,$$
$$S_{yy} = 11.82637622, \quad S_{xy} = 3.83071088$$

a. Fit the simple linear regression model to this data. Then determine the proportion of observed variation in astringency that can be attributed to the model relationship between astringency and tannin concentration.
b. Calculate and interpret a confidence interval for the slope of the true regression line.
c. Estimate true average astringency when tannin concentration is .6, and do so in a way that conveys information about reliability and precision.
d. Predict astringency for a single wine sample whose tannin concentration is .6, and do so in a way that conveys information about reliability and precision.
e. Is there compelling evidence for concluding that true average astringency is

positive when tannin concentration is .7? State and test the appropriate hypotheses.

46. The simple linear regression model provides a very good fit to the data on rainfall and runoff volume given in Exercise 17 of Section 12.2. The equation of the least squares line is $\hat{y} = -1.128 + .82697x$, $R^2 = .975$, and $s_e = 5.24$.

    a. Use the fact that $s_{\hat{Y}} = 1.44$ when rainfall volume is 40 m^3 to predict runoff in a way that conveys information about reliability and precision. Does the resulting interval suggest that precise information about the value of runoff for this future observation is available? Explain your reasoning.

    b. Calculate a PI for runoff when rainfall is 50 using the same prediction level as in part (a). What can be said about the simultaneous prediction level for the two intervals you have calculated?

47. A simple linear regression is performed on y = salary (\$1000s) and x = years of experience for actuaries. You are told that a 95% CI for the mean salary of actuaries with five years of experience, based on a sample of $n = 10$ observations, is (92.1, 117.7). Calculate a CI with confidence level 99% for the mean salary of actuaries with five years of experience.

48. Refer to Exercise 19 in which x = iodine value in grams and y = cetane number for a sample of 14 biofuels.

    a. Software gives $s_{\hat{Y}} = .802$ when $x = 80$ and $s_{\hat{Y}} = 1.074$ when $x = 120$. Explain why one is much larger than the other.

    b. Calculate a 95% CI for expected cetane number when the iodine value is 80 g.

    c. Calculate a 95% PI for the cetane number of a single biofuel with iodine value 120 g.

49. The article "Optimization of HVAC Control to Improve Comfort and Energy Performance in a School" (*Energy Engr.* 2008: 6–22) gives an analysis of the electrical and gas costs for a high school in Austin, Texas after a new heating and air conditioning (HVAC) system was installed. The accompanying data on x = average outside air temperature (°F) and y = electricity consumption (kWh) for a sample of $n = 20$ months was read from a graph in the article.

x	48	53	56	58	58
y	8200	7600	8000	10000	10400

x	59	59	60	68	69
y	10200	11000	9500	9800	8500

x	69	70	73	75	79
y	11100	11800	12000	12200	11100

x	80	80	84	87	88
y	11400	13600	10000	14000	12500

Summary quantities include $\bar{x} = 68.65$, $S_{xx} = 2692.55$, $\bar{y} = 10645$, $S_{yy} = 60,089,500$, $S_{xy} = 303,515$, $\hat{\beta}_0 = 2906$, $\hat{\beta}_1 = 112.7$.

a. Does the simple linear regression model specify a useful relationship between outside temperature and electricity consumption?

b. Estimate the true change in expected energy consumption associated with a 1 °F increase in outside temperature using a 95% confidence interval, and interpret the interval.

c. Calculate a 95% CI for $\mu_{Y|70}$, the true average monthly energy consumption when temperature = 70 °F.

d. Calculate a 95% PI for a single future observation on energy consumption to be made when temperature = 70 °F.

e. Would the 95% CI and PI when temperature = 85 °F be wider or narrower than the corresponding intervals of parts (c) and (d)? Answer without actually computing the intervals.

f. Would you recommend calculating a 95% PI for an outside temperature of 95 °F? Explain.

g. Calculate simultaneous CI's for true average monthly energy consumption when outside temperature is 60, 70, and 80 °F, respectively. Your simultaneous confidence level should be at least 97%.

50. Consider the following four intervals based on the data of the previous exercise:
   - A 95% CI for energy consumption when temp = 60
   - A 95% PI for energy consumption when temp = 60
   - A 95% CI for energy consumption when temp = 72
   - A 95% PI for energy consumption when temp = 72

   Without computing any of these intervals, what can be said about their widths relative to each other?

51. Many parts of the USA are experiencing increased erosion along waterways due to increased flow from global climate change. The report "Evaluation and Assessment of Environmentally Sensitive Stream Bank Protection Measures" (*Transp. Resour.* *Board* 2016) gives the following data on x = shear stress (lb/ft^2) and y = erosion depth (ft) for six experimental test trays at Colorado State University, built to re-create real stream conditions.

x	0.75	1.50	1.70	1.61	2.43	3.24
y	.01	.06	.10	.03	.13	.24

a. Construct a scatterplot. Does the simple linear regression model appear to be plausible?

b. Carry out a test of model utility.

c. Estimate true average erosion depth when shear stress is 1.75 lb/ft^2 by giving an interval of plausible values.

d. Estimate erosion depth along a single stream where water flow creates a shear stress of 1.75 lb/ft^2 by giving an interval of plausible values.

52. Verify that $V(\hat{\beta}_0 + \hat{\beta}_1 x)$ is indeed given by the expression in the text. [*Hint*: $V(\sum d_i Y_i) = \sum d_i^2 \cdot V(Y_i)$.]

12.5   Correlation

In many situations, the objective in studying the joint behavior of two variables is simply to see *whether* they are related, rather than to use one to predict the value of the other. In this section, we first develop the *sample correlation coefficient r* as a measure of how strongly related two variables x and y are in a sample, and then we relate r to the correlation coefficient ρ defined in Chapter 5.

The Sample Correlation Coefficient *r*

Given n pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$, it is natural to speak of x and y having a *positive relationship* if large x's are paired with large y's and small x's with small y's. Similarly, if large x's are paired with small y's and vice versa, then a *negative relationship* between the variables is implied. Consider standardizing each x value in the sample, i.e., replacing each x_i with $(x_i - \bar{x})/s_x$. Now do the same thing with the y_i's to obtain the standardized y values $(y_i - \bar{y})/s_y$. Our proposed measure of the direction and strength of the relationship between the x's and y's involves the sum of the products of these standardized values.

DEFINITION   The **sample correlation coefficient** for the n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{S_{xy}}{(n-1)s_x s_y} = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}} \tag{12.9}$$

The denominator of Expression (12.9) is clearly positive, so the sign of r (+ or –) is determined by the numerator S_{xy}. If the relationship between x and y is strongly positive, an x_i above the mean $\bar{x}$ will tend to be paired with a y_i above the mean $\bar{y}$, so that $(x_i - \bar{x})(y_i - \bar{y}) > 0$, and this same product will also be positive whenever both x_i and y_i are below their respective means (a negative times a negative equals a positive). Thus a positive relationship implies that S_{xy} will be positive. An analogous argument shows that when the relationship is negative, S_{xy} will be negative, since most of the products $(x_i - \bar{x})(y_i - \bar{y})$ will be negative. This is illustrated in Figure 12.19.
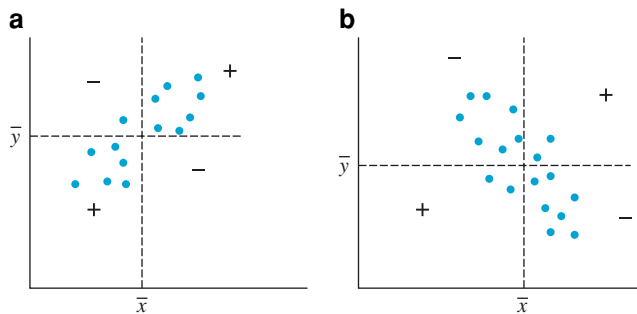
Figure 12.19 (a) Scatterplot with r and S_{xy} positive; (b) scatterplot with r and S_{xy} negative [+ means $(x_i - \bar{x})(y_i - \bar{y}) > 0$, and − means $(x_i - \bar{x})(y_i - \bar{y}) < 0$]

The most important properties of r are as listed below.

PROPERTIES OF r

1. The value of r does not depend on which of the two variables is labeled x and which is labeled y.
2. The value of r is independent of the units in which x and y are measured. In particular, r itself is unitless.
3. The square of the sample correlation coefficient gives the value of the coefficient of determination that would result from fitting the simple linear regression model—in symbols, $r^2 = R^2$.
4. $-1 \leq r \leq 1$.
5. $r = \pm 1$ if and only if all (x_i, y_i) pairs lie on a straight line.

Proof Property 1 should be evident. Property 2 is a direct result of standardizing the two variables; Exercise 64 asks for a formal verification. To prove Property 3, recall that R^2 can be expressed as the

ratio SSR/SST, where $\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$ and $\text{SST} = S_{yy} = \sum (y_i - \bar{y})^2$. It is easily shown (see Exercise 24(b)) that $\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$, and therefore

$$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 = \left(\frac{S_{xy}}{S_{xx}}\right)^2 \cdot S_{xx} = \frac{S_{xy}^2}{S_{xx}} \Rightarrow$$

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \left(\frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}\right)^2 = r^2$$

Because $r^2 = R^2 = \text{SSR/SST} = (\text{SST} - \text{SSE})/\text{SST}$, and the numerator cannot be bigger than the denominator, r must be between -1 and 1. Furthermore, because the ratio can be 1 if and only if $\text{SSE} = 0$, we conclude that $r^2 = 1$ (i.e., $r = \pm 1$) if and only if all the points fall on a straight line. ■

Property 1 stands in marked contrast to what happens in regression analysis, where virtually all quantities of interest (the estimated slope, estimated y-intercept, s_e, etc.) depend on which of the two variables is treated as the response variable. However, Property 3 shows that the proportion of variation in the response variable explained by fitting the simple linear regression model does not depend on which variable plays this role.

Property 2 is equivalent to saying that r is unchanged if each x_i is replaced by cx_i and if each y_i is replaced by dy_i (where c and d are positive, giving a change in the scale of measurement), as well as if each x_i is replaced by $x_i - a$ and y_i by $y_i - b$ (which changes the location of zero on the measurement axis). This implies, for example, that r is the same whether temperature is measured in °F or °C.

Property 4 tells us that the maximum value of r, corresponding to the largest possible degree of positive relationship, is $r = 1$, whereas the most negative relationship is identified with $r = -1$. According to Property 5, the largest positive and largest negative correlations are achieved only when all points lie along a straight line. Any other configuration of points, even if the configuration suggests a deterministic relationship between variables, will yield an r value less than 1 in absolute magnitude. Thus, r *measures the degree of linear relationship* among variables. A value of r near 0 is not necessarily evidence of a lack of a strong relationship, but only the absence of a linear relation, so that such a value of r must be interpreted with caution. Figure 12.20 illustrates several configurations of points associated with different values of r.
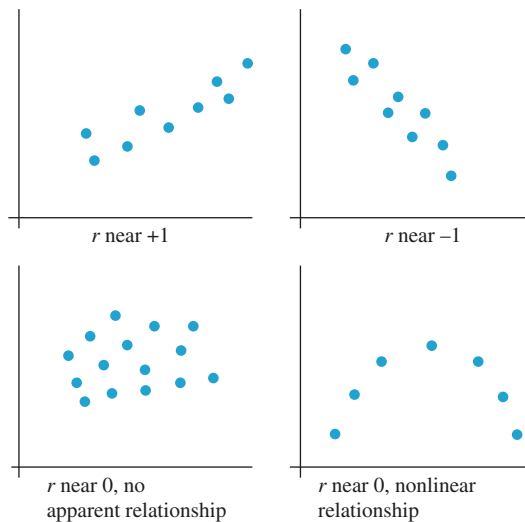
Figure 12.20 Data plots for different values of r

Example 12.13 The article "A Cross-National Relationship Between Sugar Consumption and Major Depression?" (*Depression and Anxiety* 2002: 118–120) reported the following data on x = daily sugar consumption (calories per capita) and y = annual rate of major depression (cases per 100 people) for a sample of six countries.

Country	Sugar consumption	Depression rate
USA	300	3.0
Canada	390	5.2
France	350	4.4
Germany	375	5.0
New Zealand	480	5.7
South Korea	150	2.3

With $n = 6$, $\bar{x} = 340.8$, $s_x = 110.6$, $\bar{y} = 4.267$, and $s_y = 1.338$,

$$r = \frac{1}{6-1}\left[\left(\frac{300 - 340.8}{110.6}\right)\left(\frac{3.0 - 4.267}{1.338}\right) + \cdots + \left(\frac{150 - 340.8}{110.6}\right)\left(\frac{2.3 - 4.267}{1.338}\right)\right] = .944$$

Equivalently, $S_{xx} = 61{,}120.83$, $S_{yy} = 8.953$, $S_{xy} = 698.667$, and $r = S_{xy}/\sqrt{S_{xx} \cdot S_{yy}} = .944$. Since the correlation coefficient is positive and close to 1, the data indicates a strong, positive relationship between sugar consumption and depression rate, at least for these six countries. A scatterplot of this data (not shown) also supports the notion of a strong, positive association. Note that if sugar consumption was converted into grams per capita (a gram of sugar has about 4 calories, so $x_i' = x_i/4$), the summary values for the x data would change but r would remain .944.

Does this study show that increased sugar consumption causes depression? Would forcing people in these countries to eat less sugar reduce the depression rate? Not necessarily: the high r value establishes a strong association between the two variables, but (as discussed in earlier chapters) *association does not imply causation*. Other factors not explored by the investigators may explain why nations with greater sugar consumption report higher depression rates. It should also be noted that aggregating data—here, looking at data on the national rather than individual level—tends to inflate the correlation coefficient by averaging out individual variation that would weaken the apparent relationship between the two variables.   ∎

Correlation and the Regression Effect

The correlation coefficient can be used to obtain an alternative expression for the equation of the least squares regression line:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} + r \cdot \frac{s_y}{s_x}(x - \bar{x})$$

(Exercise 66 requests a derivation of this result.) This expression for the regression line can be interpreted as follows. Suppose $r > 0$. For an x that lies one standard deviation (s_x units) above the mean $\bar{x}$ of the x_i's, the predicted y value is $\bar{y} + r \cdot s_y$, r standard deviations above the mean on the

y scale. If r is negative, the LSRL predicts that the y value when x is one sd above average will be r sd's below average. Critically, since the magnitude of r is typically strictly less than 1, our model predicts that, on a standardized scale, the response variable will be closer to its mean than the explanatory variable is to its mean.

The term *regression analysis* was first used by Francis Galton in the late nineteenth century in connection with his work on the relationship between father's height x and son's height y. After collecting a number of pairs (x_i, y_i), Galton used the principle of least squares to obtain the equation of the LSRL with the objective of using it to predict son's height from father's height. In using the derived line, Galton found that if a father was above average in height, his son was also expected to be above average in height, *but not by as much as the father*. Similarly, the son of a shorter-than-average father was expected to be shorter than average, but not by as much as the father. Thus the predicted height of a son was "pulled back in" toward the mean; because *regression* can be defined as moving backward, Galton adopted the terminology *regression line*. This phenomenon of being pulled back in toward the mean has been observed in many other situations (e.g., a player's batting averages from year to year in baseball) and is called the **regression effect** or **regression to the mean**. See also Section 5.5 for a discussion of this topic in the context of the bivariate normal distribution.

Because of the regression effect, care must be exercised in experiments that involve selecting individuals based on below-average scores. For example, if students are selected because of below-average performance on a test and they are then given special instruction, the regression effect predicts improvement even if the instruction is useless. A similar warning applies in studies of underperforming businesses or hospital patients.

The Population Correlation Coefficient ρ and Inferences About Correlation

The correlation coefficient r is a measure of how strongly related x and y are in the observed sample. We can think of the pairs (x_i, y_i) as having been drawn from a bivariate population of pairs, with (X_i, Y_i) having some joint probability distribution $f(x, y)$. In Chapter 5, we defined the correlation coefficient $\rho(X, Y)$ by

$$\rho = \rho(X, Y) = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right] = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

If we think of $f(x, y)$ as describing the distribution of pairs of values within the entire population, ρ becomes a measure of how strongly related x and y are in that population. Properties of ρ analogous to those for r were given in Chapter 5.

The population correlation coefficient ρ is a parameter or population characteristic, just as μ_X, μ_Y, σ_X, and σ_Y are, and we can use the sample correlation coefficient to make various inferences about ρ. In particular, r is a point estimate for ρ, and the corresponding estimator is

$$\hat{\rho} = R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s_X}\right)\left(\frac{X_i - \overline{Y}}{s_Y}\right) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}\sqrt{\sum (Y_i - \overline{Y})^2}}$$

Many of the intervals and test procedures presented in Chapters 8–10 were based on an assumption of population normality. To test hypotheses about ρ, we must make an analogous assumption about the distribution of pairs of (x, y) values in the population. We are now assuming that *both X and Y are random*, with joint distribution given by the bivariate normal pdf introduced in Section 5.5.

If $X = x$, recall from Section 5.5 that the conditional distribution of Y is normal with mean $\mu_{Y|x} = \mu_2 + (\rho\sigma_2/\sigma_1)(x - \mu_1)$ and variance $(1 - \rho^2)\sigma_2^2$. This is exactly the model used in simple

linear regression with $\beta_0 = \mu_2 - \rho\mu_1\sigma_2/\sigma_1, \beta_1 = \rho\sigma_2/\sigma_1,$ and $\sigma^2 = (1 - \rho^2)\sigma_2^2$ independent of x. The implication is that *if the observed pairs (x_i, y_i) are actually drawn from a bivariate normal distribution, then the simple linear regression model is an appropriate way of studying the behavior of Y for fixed x*. If $\rho = 0$, then $\mu_{Y|x} = \mu_2$ independent of x; in fact, when $\rho = 0$ the joint pdf $f(x, y)$ can be factored into a part involving x only and a part involving y only, which implies that X and Y are independent random variables.

Example 12.14 As discussed in Section 5.5, contours of the bivariate normal distribution are elliptical, and this suggests that a scatterplot of observed (x, y) pairs from such a joint distribution should have a roughly elliptical shape. The article "Methods of Estimation of Visceral Fat: Advantages of Ultrasonography" (*Obesity Res.* 2003: 1488–1494) includes the scatterplot in Figure 12.21 for x = visceral fat (cm^2) measured by ultrasound (US) versus y = visceral fat by computerized tomography (CT) for a sample of $n = 100$ obese women. CT is considered the most accurate technique for body fat measurement but is costly, time-consuming, and involves exposure to ionizing radiation; the US method is noninvasive and less expensive.
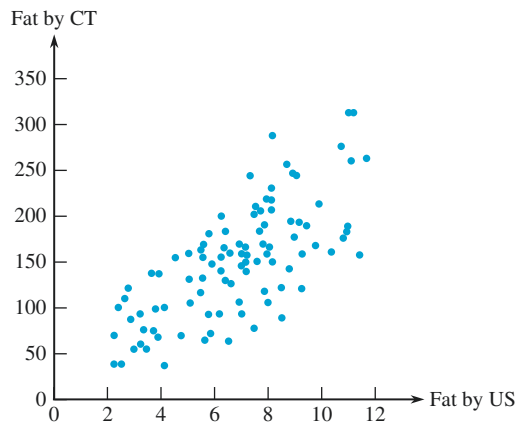
Figure 12.21   Scatterplot for Example 12.14

The pattern in the scatterplot in Figure 12.21 seems consistent with an assumption of bivariate normality. If we let ρ denote the true population correlation coefficient between CT and US measurements, then a point estimate of ρ is $\hat{\rho} = r = .71$, a value given in the article. Of course we would want fat measurements from the two methods to be very highly correlated before regarding one as an adequate substitute for the other. By that standard, $r = .71$ is not all that impressive, but the investigators reported that a test of $H_0: \rho = 0$ (to be introduced shortly) gives P-value $< .001$.   ∎

Assuming that the pairs are drawn from a bivariate normal distribution allows us to test hypotheses about ρ and to construct a CI. There is no completely satisfactory way to check the plausibility of the bivariate normality assumption. A partial check involves constructing two separate normal probability plots, one for the sample x_i's and another for the sample y_i's, since bivariate normality implies that the marginal distributions of both X and Y are normal. If either probability plot deviates substantially from a straight-line pattern, the following inferential procedures should not be used when the sample size n is small. Also, as in Example 12.14, the scatterplot should show a roughly elliptical shape.

TESTING FOR THE ABSENCE OF CORRELATION

When $H_0: \rho = 0$ is true, the test statistic

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

has a t distribution with $n-2$ df (see Exercise 63).

Alternative Hypothesis	Rejection Region for Level α Test
$H_a: \rho > 0$	$t \geq t_{\alpha, n-2}$
$H_a: \rho < 0$	$t \leq -t_{\alpha, n-2}$
$H_a: \rho \neq 0$	either $t \geq t_{\alpha/2, n-2}$ or $t \leq -t_{\alpha/2, n-2}$

A P-value based on $n-2$ df can be calculated as described previously.

Example 12.15 Neurotoxic effects of manganese are well known and are usually caused by high occupational exposure over long periods of time. In the fields of occupational hygiene and environmental hygiene, the relationship between lipid peroxidation, which is responsible for deterioration of foods and damage to live tissue, and occupational exposure had not been previously reported. The article "Lipid Peroxidation in Workers Exposed to Manganese" (*Scand. J. Work Environ. Health* 1996: 381–386) gave data on x = manganese concentration in blood (ppb) and y = concentration (μmol/L) of malondialdehyde, which is a stable product of lipid peroxidation, both for a sample of 22 workers exposed to manganese and for a control sample of 45 individuals. The value of r for the control sample was .29, from which

$$t = \frac{(.29)\sqrt{45-2}}{\sqrt{1-.29^2}} \approx 2.0$$

The corresponding P-value for a two-tailed t test based on 43 df is roughly .052 (the cited article reported only that the P-value > .05). We would not want to reject the assertion that $\rho = 0$ at either significance level .01 or .05. For the sample of exposed workers, $r = .83$ and $t = 6.7$, clear evidence that there is a positive relationship in the entire population of exposed workers from which the sample was selected. Although in general correlation does not necessarily imply causation, it is plausible here that higher levels of manganese cause higher levels of peroxidation. ∎

Because ρ measures the extent to which there is a linear relationship between the two variables in the population, the null hypothesis $H_0: \rho = 0$ states that there is no such population relationship. In Section 12.3, we used the t ratio $\hat{\beta}_1/s_{\hat{\beta}_1}$ to test for a linear relationship between the two variables in the context of regression analysis. It turns out that the two test procedures are completely equivalent because $r\sqrt{n-2}/\sqrt{1-r^2} = \hat{\beta}_1/s_{\hat{\beta}_1}$ (Exercise 63).

Other Inferences Concerning ρ

The procedure for testing $H_0: \rho = \rho_0$ when $\rho_0 \neq 0$ is not equivalent to any procedure from regression analysis. The test statistic is based on a transformation of R called the *Fisher transformation*.

PROPOSITION   When $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a sample from a bivariate normal distribution, the rv

$$V = \frac{1}{2} \ln \left(\frac{1+R}{1-R} \right)$$

has approximately a normal distribution with mean and variance

$$\mu_V = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) \qquad \sigma_V^2 = \frac{1}{n-3}$$

The rationale for the transformation is to obtain a function of R that has a variance independent of ρ; this would not be the case with R itself. The approximation will not be valid if n is quite small.

The test statistic for testing $H_0: \rho = \rho_0$ is

$$Z = \frac{V - \frac{1}{2} \ln[(1+\rho_0)/(1-\rho_0)]}{1/\sqrt{n-3}}$$

Alternative Hypothesis	Rejection Region for Level α Test
$H_a: \rho > \rho_0$	$z \geq z_\alpha$
$H_a: \rho < \rho_0$	$z \leq -z_\alpha$
$H_a: \rho \neq \rho_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

A P-value can be calculated in the same manner as for previous z tests.

Example 12.16 As far back as Leonardo da Vinci, it was known that height and wingspan (measured fingertip to fingertip between outstretched hands) are closely related. For these measurements (in inches) from 16 students in a statistics class notice how close the two values are.

Student	1	2	3	4	5	6	7	8
Height	63.0	63.0	65.0	64.0	68.0	69.0	71.0	68.0
Wingspan	62.0	62.0	64.0	64.5	67.0	69.0	70.0	72.0

Student	9	10	11	12	13	14	15	16
Height	68.0	72.0	73.0	73.5	70.0	70.0	72.0	74.0
Wingspan	70.0	72.0	73.0	75.0	71.0	70.0	76.0	76.5

The scatterplot in Figure 12.22 shows an approximately linear shape, and the point cloud is roughly elliptical. Also, the normal plots for the individual variables are roughly linear, so the bivariate normal distribution can reasonably be assumed.
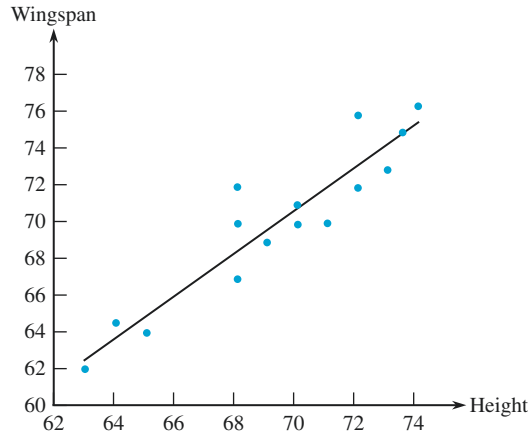
Figure 12.22 Wingspan plotted against height

The correlation is computed to be .9422. Can it be concluded that true correlation between wingspan and height exceeds .8? To carry out a test of H_0: $\rho = .8$ versus H_a: $\rho > .8$, we Fisher transform .9422 and .8:

$$v = \frac{1}{2}\ln\left(\frac{1 + .9422}{1 - .9422}\right) = 1.757 \qquad \mu_V = \frac{1}{2}\ln\left(\frac{1 + .8}{1 - .8}\right) = 1.099$$

The z test statistic is $z = (1.757 - 1.099)/(1/\sqrt{16 - 3}) = 2.37$. Since $2.37 \geq z_{.01} = 2.33$, at level .01 we can reject H_0: $\rho = .8$ in favor of H_a: $\rho > .8$ and conclude that wingspan is highly correlated with height. ∎

To obtain a CI for ρ, we first derive an interval for $\mu_V = \frac{1}{2}\ln[(1 + \rho)/(1 - \rho)]$. Standardizing V, writing a probability statement, and manipulating the resulting inequalities yields

$$v \pm z_{\alpha/2} \cdot \sigma_V = \frac{1}{2}\ln\left(\frac{1 + r}{1 - r}\right) \pm \frac{z_{\alpha/2}}{\sqrt{n - 3}} \tag{12.10}$$

as the endpoints of a $100(1 - \alpha)\%$ interval for μ_V. This interval can then be manipulated to yield a CI for ρ.

A **$100(1 - \alpha)\%$ confidence interval for ρ** is

$$\left(\frac{e^{2c_1} - 1}{e^{2c_1} + 1}, \quad \frac{e^{2c_2} - 1}{e^{2c_2} + 1}\right)$$

where c_1 and c_2 are the left and right endpoints, respectively, in Expression (12.10).

Example 12.17 (Example 12.16 continued) The sample correlation coefficient between wingspan and height was $r = .9422$, giving $v = 1.757$. With $n = 16$, a 95% confidence interval for μ_V is $1.757 \pm 1.96/\sqrt{16 - 3} = (1.213, 2.301) = (c_1, c_2)$.

The 95% interval for ρ is

$$\left(\frac{e^{2(1.213)} - 1}{e^{2(1.213)} + 1}, \frac{e^{2(2.301)} - 1}{e^{2(2.301)} + 1}\right) = (.838, .980)$$

Notice that this interval excludes .8, and that our hypothesis test in Example 12.16 would have rejected H_0: $\rho = .8$ in favor of the alternative H_a: $\rho > .8$ at the .025 level.   ∎

Absent the assumption of bivariate normality, a bootstrap procedure can be used to obtain a CI for ρ or test hypotheses.

In Chapter 5, we cautioned that a large value of the correlation coefficient (near 1 or −1) implies only association and not causation. This applies to both ρ and r. It is easy to find strong but weird correlations in which neither variable is casually related to the other. For example, since Prohibition ended in the 1930s, beer consumption and church attendance have correlated very highly. Of course, the reason is that both variables have increased in accord with population growth.

Exercises: Section 12.5 (53–66)

53. The article "Behavioural Effects of Mobile Telephone Use During Simulated Driving" (*Ergonomics* 1995: 2536–2562) reported that for a sample of 20 experimental subjects, the sample correlation coefficient for x = age and y = time since the subject had acquired a driving license (yr) was .97. Why do you think the value of r is so close to 1? (The article's authors gave an explanation.)

54. The Turbine Oil Oxidation Test (TOST) and the Rotating Bomb Oxidation Test (RBOT) are two different procedures for evaluating the oxidation stability of steam turbine oils. The article "Dependence of Oxidation Stability of Steam Turbine Oil on Base Oil Composition" (*J. Soc. Tribologists Lubricat. Engrs.*, Oct. 1997: 19–24) reported the accompanying observations on x = TOST time (hr) and y = RBOT time (min) for 12 oil specimens.

TOST	4200	3600	3750	3675	4050	2770
RBOT	370	340	375	310	350	200
TOST	4870	4500	3450	2700	3750	3300
RBOT	400	375	285	225	345	285

   a. Calculate and interpret the value of the sample correlation coefficient (as did the article's authors).

b. How would the value of r be affected if we had let x = RBOT time and y = TOST time?

c. How would the value of r be affected if RBOT time was expressed in hours?

d. Construct a scatterplot and normal probability plots and comment.

e. Carry out a test of hypotheses to decide whether RBOT time and TOST time are linearly related.

55. The authors of the paper "Objective Effects of a Six Months' Endurance and Strength Training Program in Outpatients with Congestive Heart Failure" (*Med. Sci. Sports Exerc.* 1999: 1102–1107) presented a correlation analysis to investigate the relationship between maximal lactate level x and muscular endurance y. The accompanying data was read from a plot in the paper.

x	400	750	770	800	850	1025	1200
y	3.80	4.00	4.90	5.20	4.00	3.50	6.30

x	1250	1300	1400	1475	1480	1505	2200
y	6.88	7.55	4.95	7.80	4.45	6.60	8.90

$S_{xx} = 36.9839, \quad S_{yy} = 2{,}628{,}930.357,$
$S_{xy} = 7377.704$

A scatterplot shows a linear pattern.

a. Test to see whether there is a positive correlation between maximal lactate level and muscular endurance in the population from which this data was selected.

b. If a regression analysis was to be carried out to predict endurance from lactate level, what proportion of observed variation in endurance could be attributed to the approximate linear relationship? Answer the analogous question if regression is used to predict lactate level from endurance—and answer both questions without doing any regression calculations.

56. Torsion during hip external rotation and extension may explain why acetabular labral tears occur in professional athletes. The article "Hip Rotational Velocities During the Full Golf Swing" (*J. Sport Sci. Med.* 2009: 296–299) reported on an investigation in which lead hip internal peak rotational velocity (x) and trailing hip peak external rotational velocity (y) were determined for a sample of 15 golfers. Data provided by the article's authors was used to calculate the following summary quantities:

$$S_{xx} = 64{,}732.83, \quad S_{yy} = 130{,}566.96,$$
$$S_{xy} = 44{,}185.87$$

Separate normal probability plots showed very substantial linear patterns.

a. Calculate a point estimate for the population correlation coefficient.

b. If the simple linear regression model was fit to the data, what proportion of variation in external velocity could be attributed to the model relationship? What would happen to this proportion if the roles of x and y were reversed? Explain.

c. Carry out a test at significance level .01 to decide whether there is a linear

relationship between the two velocities in the sampled population; your conclusion should be based on a *P*-value.

d. Would the conclusion of (c) have changed if you had tested appropriate hypotheses to decide whether there is a positive linear association in the population? What if a significance level of .05 rather than .01 had been used?

57. Hydrogen content is conjectured to be an important factor in porosity of aluminum alloy castings. The article "The Reduced Pressure Test as a Measuring Tool in the Evaluation of Porosity/Hydrogen Content in A1–7 Wt Pct Si-10 Vol Pct SiC(p) Metal Matrix Composite" (*Metallurg. Trans.* 1993: 1857–1868) gives the accompanying data on x = content and y = gas porosity for one particular measurement technique.

x	.18	.20	.21	.21	.21	.22	.23
y	.46	.70	.41	.45	.55	.44	.24

x	.23	.24	.24	.25	.28	.30	.37
y	.47	.22	.80	.88	.70	.72	.75

Minitab gives the following output in response to a correlation command:

```
Correlation of Hydrcon and
Porosity = 0.449
```

a. Test at level .05 to see whether the population correlation coefficient differs from 0.

b. If a simple linear regression analysis had been carried out, what percentage of observed variation in porosity could be attributed to the model relationship?

58. The wicking properties of certain fabrics were investigated in the article "Thermal and Water Vapor Transport Properties of Selected Lofty Nonwoven Products" (*Textile Res. J.* 2017: 1413–1424). Use the accompanying data and a .01 significance level to determine whether there is a significant correlation between thickness x (mm) and water vapor resistance

y (m^{2}Pa/W). Is the result of the test surprising in light of the value of r?

x	20	20	30	30	40	40
y	60	56	65	70	96	78

59. The body armor report introduced in Example 12.1 also reported studies in which two different methods were used to measure the same body armor deformations from bullet impact. The goal of these studies was to assess the extent to which two different measurement instruments agree. Eighty-three backface deformations (mm) were measured using a digital caliper (x) and a laser arm (y), resulting in a sample correlation coefficient of $r = .878$.
   a. Compute a 90% CI for the true correlation coefficient ρ.
   b. Test H_0: $\rho = .8$ versus H_a: $\rho > .8$ at level .05.
   c. In a regression analysis of y on x, what proportion of variation in laser arm measurements could be explained by variation in digital caliper measurements?
   d. If you decide to perform a regression analysis with digital caliper measurement as the response variable, what proportion of its variation is explainable by variation in laser arm measurement?

60. It is time-consuming and costly to have trucks stop in order to be weighed on a static scale. The Minnesota Department of Transportation considered using a scale that would weigh trucks while they were moving. Here is data for a sample of trucks that were weighed in motion and also on a static scale (1000s of lbs).

Truck	1	2	3	4	5
In-motion	26.0	29.9	39.5	25.1	31.6
Static	27.9	29.1	38.0	27.0	30.3
Truck	6	7	8	9	10
In-motion	36.2	25.1	31.0	35.6	40.2
Static	34.5	27.8	29.6	33.1	35.5

   a. Determine the sample correlation coefficient r.
   b. Test H_0: $\rho = .85$ versus H_a: $\rho > .85$ at level .05.
   c. How successful do you think the simple linear regression model would be in predicting static weight from in-motion weight? Explain.

61. A sample of $n = 500$ (x, y) pairs was collected and a test of H_0: $\rho = 0$ versus H_a: $\rho \neq 0$ was carried out. The resulting P-value was computed to be .00032.
   a. What conclusion would be appropriate at level of significance .001?
   b. Does this small P-value indicate that there is a very strong relationship between x and y (a value of ρ that differs considerably from 0)? Explain.
   c. Now suppose a sample of $n = 10,000$ (x, y) pairs resulted in $r = .022$. Test H_0: $\rho = 0$ versus H_a: $\rho \neq 0$ at level .05. Is the result statistically significant? Comment on the practical significance of your analysis.

62. Let x be number of hours per week of studying and y be grade point average. Suppose we have one sample of (x, y) pairs for females and another for males. Then we might like to test the hypothesis H_0: $\rho_1 - \rho_2 = 0$ against the alternative that the two population correlation coefficients are different.
   a. Use properties of the transformed variable $V = .5\ln[(1 + R)/(1 - R)]$ to propose an appropriate test statistic and rejection region (let R_1 and R_2 denote the two-sample correlation coefficients).
   b. The paper "Relational Bonds and Customer's Trust and Commitment: A Study on the Moderating Effects of Web Site Usage" (*Serv. Ind. J.* 2003: 103–124) reported that $n_1 = 261$, $r_1 = .59$, $n_2 = 557$, $r_2 = .50$, where the first sample consisted of corporate website users and the second of nonusers; here r is the correlation between an assessment of the

strength of economic bonds and performance. Carry out the test for this data (as did the authors of the cited paper).

63. Verify that the t ratio for testing $H_0: \beta_1 = 0$ in Section 12.3 is identical to t for testing $H_0: \rho = 0$.

64. Verify Property 2 of the correlation coefficient: the value of r is independent of the units in which x and y are measured; that is, if $x_i' = ax_i + c$ and $y_i' = by_i + d$, $a > 0$, $b > 0$, then r for the (x_i', y_i') pairs is the same as r for the (x_i, y_i) pairs.

65. Consider a time series—that is, a sequence of observations $X_1, X_2, \ldots$ on some response variable (e.g., concentration of a pollutant) over time—with observed values $x_1, x_2, \ldots, x_n$ over n time periods. Then the *lag 1 autocorrelation coefficient*, which assess the strength of relationship between series values one time unit apart, is defined as

$$r_1 = \frac{\sum_{i=1}^{n-1} (x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Autocorrelation coefficients r_2, r_3, … for lags 2, 3, … are defined analogously.

a. Calculate the values of r_1, r_2, and r_3 for the temperature data from Chapter 1 Exercise 95.

b. Consider the pairs (x_1, x_2), (x_2, x_3), …, (x_{n-1}, x_n). What is the difference between the formula for the sample correlation coefficient r applied to these pairs and the formula for r_1? What if n, the length of the series, is large? What

about r_2 compared to r for the $n - 2$ pairs (x_1, x_3), (x_2, x_4), …, (x_{n-2}, x_n)?

c. Analogous to the population correlation coefficient ρ, let ρ_i $(i = 1, 2, 3, \ldots)$ denote the theoretical or long-run autocorrelation coefficients at the various lags. If all these ρ's are zero, there is no (linear) relationship between observations in the series at *any* lag. In this case, if n is large, each R_i has approximately a normal distribution with mean 0 and standard deviation $1/\sqrt{n}$ and different R_i's are almost independent. Therefore $H_0: \rho_i = 0$ can be rejected at a significance level of approximately .05 if either $r_i \geq 2/\sqrt{n}$ or $r_i \leq -2/\sqrt{n}$. If $n = 100$ and $r_1 = .16$, $r_2 = -.09$, $r_3 = -.15$, is there evidence of theoretical autocorrelation at any of the first three lags?

d. If you are testing the null hypothesis in (c) for more than one lag, why might you want to increase the cutoff constant 2 in the rejection region? [*Hint*: What about the probability of committing at least one type I error?]

66. Let s_x and s_y denote the sample standard deviations of the observed x's and y's, respectively.

a. Show that $S_{xx} = (n - 1)s_x^2$ and similarly for the y's.

b. Show that an alternative expression for the estimated regression line $\hat{\beta}_0 + \hat{\beta}_1 x$ is

$$y = \bar{y} + r \cdot \frac{s_y}{s_x} (x - \bar{x})$$

12.6   Investigating Model Adequacy: Residual Analysis

In the last several sections we have taken for granted that our data is consistent with the simple linear regression model, which makes certain assumptions about the "true error" term ε. Table 12.2 summarizes these model assumptions.

Since we do not observe the true errors in practice, our assessment of the plausibility of these assumptions is made based on the observed residuals $e_1, \ldots, e_n$. Certain graphs of the residuals, some of which appeared in the context of ANOVA in Chapter 11, will allow us to validate the regression assumptions. Moreover, these graphs may reveal other unusual or noteworthy features of the data.

Table 12.2 Assumptions of the simple linear regression model

Assumption	In terms of Y	In terms of ε
Linearity	$E(Y\|x)$ is a linear function of x.	For any fixed x, $E(\varepsilon) = 0$.
Normality	For any fixed x, the Y distribution is normal.	For any fixed x, the rv ε is normally distributed.
Constant variance	The variance of Y at any fixed x value is independent of x.	$V(\varepsilon) = \sigma^2$, independent of x.
Independence	Y_i's for different observations are independent.	ε_i's for different observations are independent.

Residuals and Standardized Residuals

Suppose the simple linear regression model is correct, and let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the predicted y value of the ith observation. Then the ith residual is $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$. To derive properties of the residuals, let $Y_i - \hat{Y}_i$ represent the ith residual as a random variable (i.e., before observations are actually made). Then

$$E(Y_i - \hat{Y}_i) = E(Y_i) - E(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x_i) = 0$$

It can also be shown (Exercise 74) that

$$V(Y_i - \hat{Y}_i) = \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}} \right] \tag{12.11}$$

Notice that the further x_i is from $\overline{x}$, the *smaller* the variance will be. This is because the least squares line is "pulled toward" observations whose x values are extreme relative to the other x values. Replacing σ by s_e and taking the square root of Equation (12.11) gives the estimated standard deviation of the ith residual.

Let's now standardize each residual by subtracting the mean value (zero) and then dividing by the estimated standard deviation.

DEFINITION   The **standardized residuals** are

$$e_i^* = \frac{e_i - 0}{s_{e_i}} = \frac{y_i - \hat{y}_i}{s_e \sqrt{1 - \dfrac{1}{n} - \dfrac{(x_i - \overline{x})^2}{S_{xx}}}} \qquad i = 1, \ldots, n$$

If, for example, a particular standardized residual is 1.5, then the residual itself is 1.5 standard deviations larger than what would be expected from fitting the correct model. Though the standard deviations of the residuals differ from one another, if n is reasonably large the bracketed term in (12.11) will be approximately 1, so some sources use $e_i^* \approx e_i/s_e$ as the standardized residual. Computation of the e_i^*'s can be tedious, but the most widely used statistical computer packages automatically provide these values and can construct various plots involving them.

Example 12.18 Does stress really accelerate aging? A study described in the article "Accelerated Telomere Shortening in Response to Life Stress" (*Proc. Nat. Acad. Sci.* 2004: 17312–17315) investigated the relationship between x = perceived stress level (on a quantitative scale) and y = telomere length, a biological measure of cell longevity (smaller telomere lengths indicate shorter lifespan at the cellular level). Figure 12.23 shows a scatterplot of (x, y) pairs for 38 subjects; the plot suggests a negative, weak-to-moderate ($r = -.32$) association between stress level and telomere length.
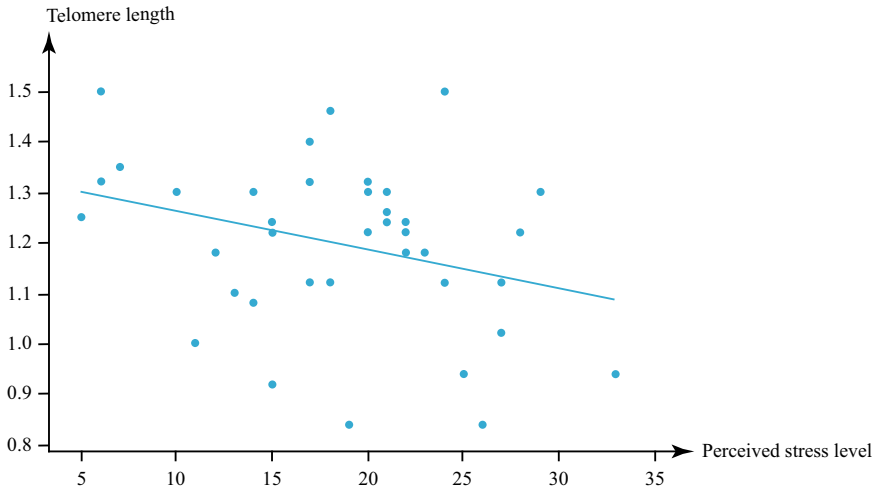
Figure 12.23 Scatterplot of data in Example 12.18

The accompanying table displays the data, residuals, and standardized residuals obtained from software. The estimated standard deviations of the residuals are slightly different, e.g., for the first two observations, $s_{e_1} \approx .156$ while $s_{e_2} \approx .157$.

x_i	y_i	e_i	e_i^*	x_i	y_i	e_i	e_i^*
14	1.30	0.068	0.439	7	1.35	0.065	0.431
17	1.32	0.111	0.710	11	1.00	−0.255	−1.649
14	1.08	−0.152	−0.971	15	1.24	0.016	0.103
27	1.02	−0.112	−0.729	5	1.25	−0.050	−0.341
22	1.24	0.070	0.446	21	1.26	0.082	0.524
12	1.18	−0.067	−0.431	24	1.50	0.345	2.218
22	1.18	0.010	0.062	21	1.24	0.062	0.396
24	1.12	−0.035	−0.224	6	1.50	0.207	1.387
25	0.94	−0.207	−1.337	20	1.30	0.114	0.729
18	1.46	0.259	1.650	22	1.22	0.050	0.318
28	1.22	0.096	0.627	26	0.84	−0.300	−1.941
21	1.30	0.122	0.779	10	1.30	0.038	0.246
19	0.84	−0.353	−2.250	18	1.12	−0.081	−0.515
23	1.18	0.017	0.112	17	1.12	−0.089	−0.564
15	1.22	−0.004	−0.025	20	1.22	0.034	0.220
15	0.92	−0.304	−1.943	13	1.10	−0.139	−0.895
27	1.12	−0.012	−0.078	33	0.94	−0.146	−0.994
17	1.40	0.191	1.220	20	1.32	0.134	0.857
6	1.32	0.027	0.182	29	1.30	0.183	1.209

∎

Diagnostic Plots for Checking Assumptions

The basic plots that many statisticians recommend for an assessment of model validity are the following:

1. e_i^* (or e_i) on the vertical axis and x_i on the horizontal axis—i.e., a plot of the (x_i, e_i^*) or (x_i, e_i) pairs
2. e_i^* (or e_i) on the vertical axis and $\hat{y}_i$ on the horizontal axis—i.e., a plot of the $(\hat{y}_i, e_i^*)$ or $(\hat{y}_i, e_i)$ pairs
3. A normal probability plot of the e_i^*'s (or e_i's)

Plots 1 and 2 are called **residual plots** (against the explanatory variable and the fitted values, respectively). These two plots generally look quite similar, since $\hat{y}$ is simply a linear function of x; the advantage of plot 2 is that we may also use it for assumption diagnostics in multiple regression, as we'll see in Section 12.7. Diagnostic plots 2 and 3 were both utilized in Chapter 11 for validating ANOVA assumptions.

A residual plot can be used to validate two assumptions for simple linear regression: linearity and constant variance. Ideally, residuals should be randomly distributed about a horizontal line passing through 0. Figure 12.24 shows two prototype scatterplots and the corresponding residual plots. Figure 12.24a shows a scatterplot for which a straight line, at least initially, may appear a good fit. However, the associated residual plot exhibits strong curvature, suggesting the relationship between x and (the mean of) y is not actually linear. Figure 12.24b shows nonconstant variance: as x increases, so does the spread of the residuals about the mean line. (It is certainly possible to have a residual plot indicating both nonlinearity and nonconstant variance.)
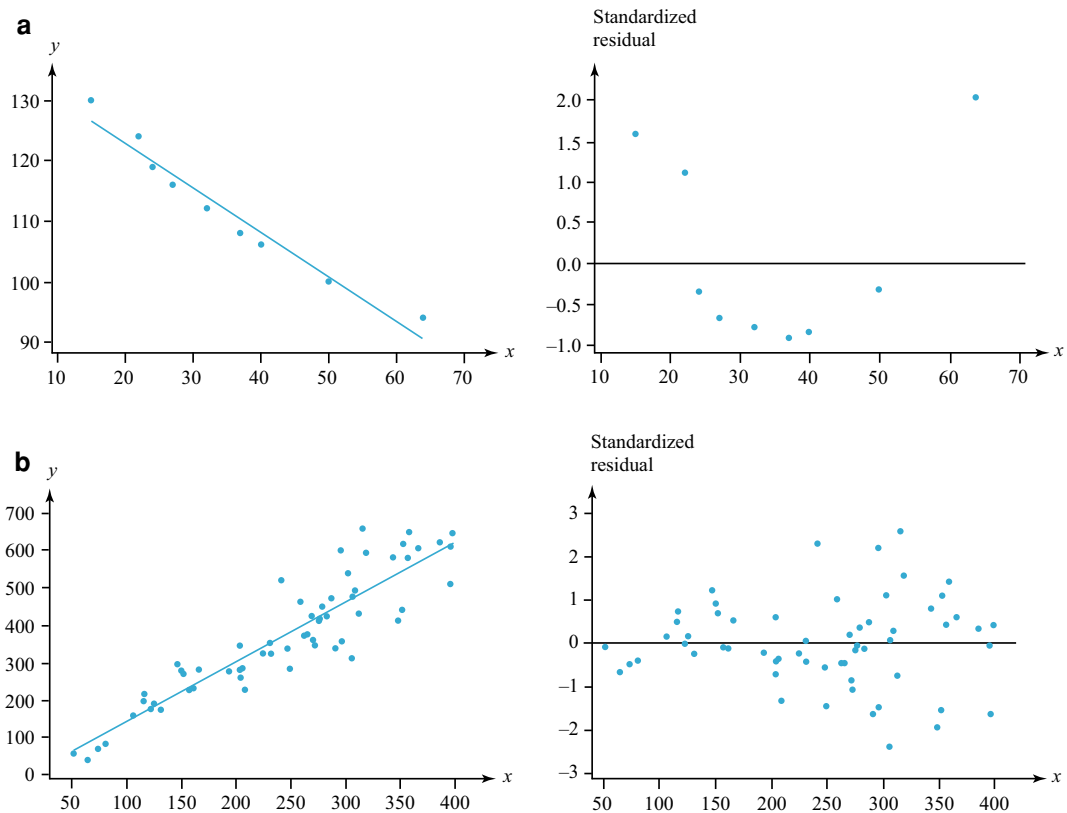
Figure 12.24   Scatterplots and residual plots: (**a**) violation of the linearity assumption; (**b**) violation of the constant variance assumption

The assumption of normally distributed errors is, naturally, checked by a normal probability plot of the residuals or standardized residuals. As before, approximate normality of the residuals becomes less important as n increases for most of our inferential procedures in regression. For example, the t test in Section 12.3 is still valid for large n even if the residuals are clearly nonnormal. The exception to this is a prediction interval (PI) for a future y value presented in Section 12.4.

Example 12.19 (Example 12.18 continued) Figure 12.25 presents a residual-versus-fit plot (e_i^* vs. $\hat{y}_i$) and a normal probability plot of the e_i^*'s for the stress–telomere data. The lack of a pattern in Figure 12.25a, e.g., lack of curvature, validates the linearity assumption, while the relatively equal vertical spread throughout the graph indicates the constant variance assumption is reasonable here. The points in Figure 12.25b plot are quite straight, suggesting that the standardized residuals—and, by extension, the true errors ε_i—might reasonably come from a normally distributed population.
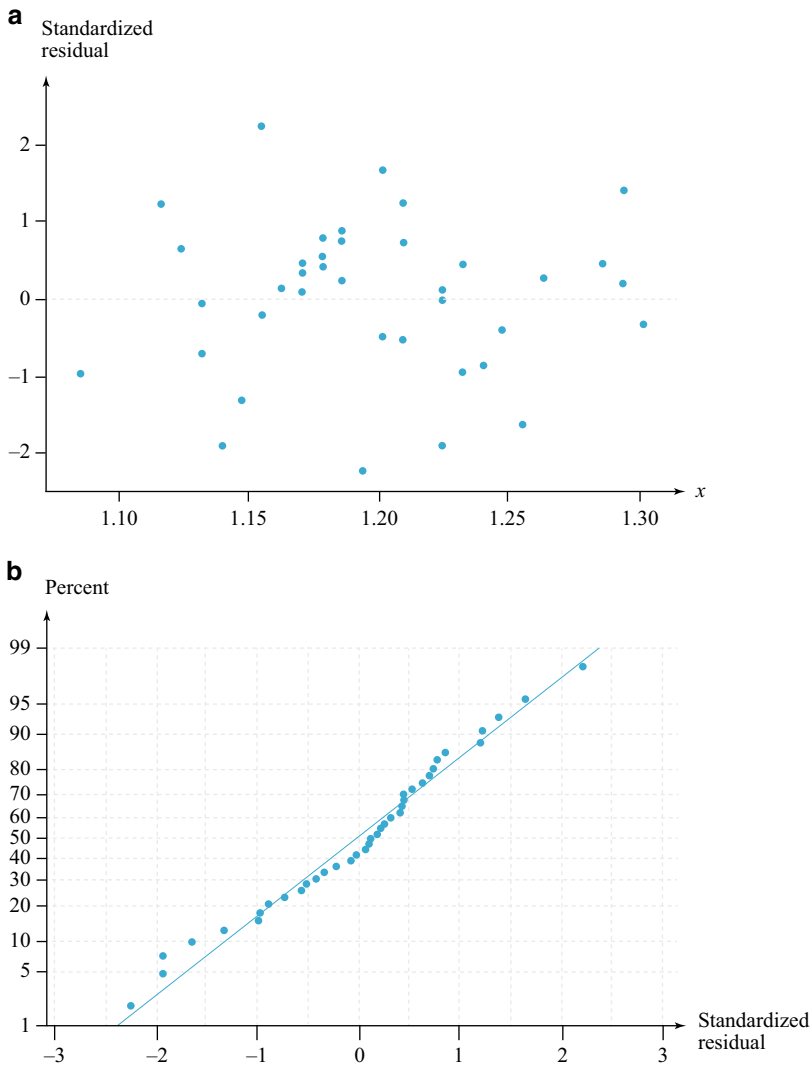
Figure 12.25 Plots for the data from Example 12.19

What about the final assumption, independence? Provided that the observations were obtained through random sampling (or, in the case of an experiment, treatments were randomly assigned to subjects), it is reasonable to treat the response values as independent. In Example 12.18, the 38 subjects were volunteers, but there is no reason to think their stress levels or telomere lengths are related. (The lack of random sampling does, however, call into question the extent to which the study's results can be generalized to a larger population of individuals.)   ■

Other Diagnostic Tools

Besides violations of the inference requirements for simple linear regression, bivariate data will sometimes present other difficulties:

1. The selected model fits the data well except for a few discrepant or outlying data values, which may have greatly influenced the choice of the best-fit function.
2. When the observations (x_i, y_i) appear in time order (i.e., the subscript i is actually a time index), the errors exhibit dependence over time.
3. One or more relevant explanatory variables have been omitted from the model.

Figure 12.26 presents plots corresponding to these three scenarios. Some unusual observations can be detected by a residual plot, particularly those with large standardized residuals (see Figure 12.26a). However, detection of all types of unusual observations can be difficult, especially in a multiple regression setting. A more complete analysis of unusual observations for both simple and multiple regression is presented in Section 12.9.
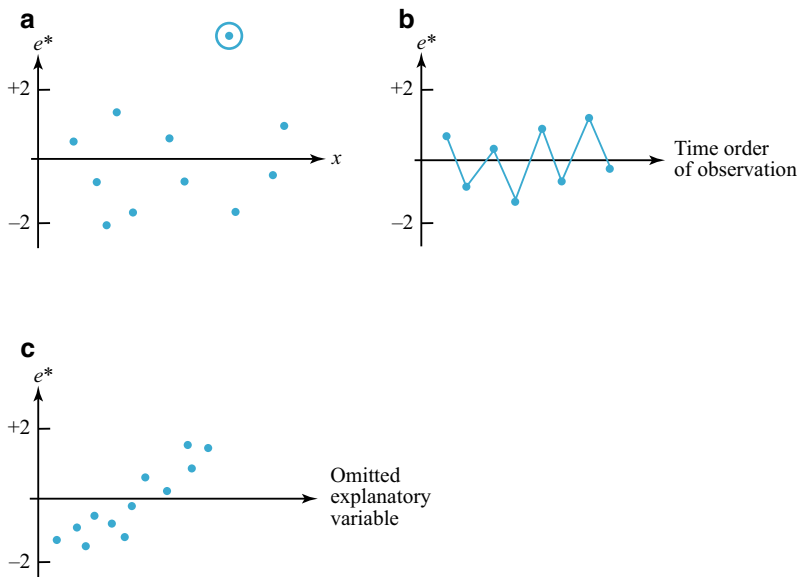
Figure 12.26 Plots that indicate abnormality in data: (**a**) a discrepant observation; (**b**) dependence in errors; (**c**) an omitted explanatory variable

Figure 12.26b shows a plot of the standardized residuals against time order; this is only appropriate when the data is collected sequentially (in successive time periods) rather than randomly. The line segments connecting successive points emphasize the sequential nature of the observations.

Observe that the residuals have a perfect alternating pattern: the first residual is above the mean line, the next one is below, the next above, and so on. This is an example of *autocorrelation:* a time-dependent pattern in the residuals. The methods of this chapter should be applied with great caution to modeling sequential observations—known as *time series* data—since the independence assumption is generally not met for this type of data.

Figure 12.26c shows a plot of the e_i^*'s against an explanatory variable *other than x*. The presence of a pattern suggests that this other explanatory variable should be added to the model, resulting in a multiple regression model. In Example 12.18, we might find that the residuals from the regression of y = telomere length on x_1 = stress level are linearly related to the values of x_2 = subject's age. If so, it makes sense to use both x_1 and x_2 to predict y (the topic of Section 12.7).

Remedies for Assumption Violations
We now briefly indicate what remedies are available for some of the difficulties encountered in this section. Several of these are discussed in greater detail in subsequent sections of the book. For a more comprehensive discussion, one or more of the bibliographic references on regression analysis should be consulted.

If the relationship between x and y appears nonlinear (e.g., as indicated in the residual plot of Figure 12.24a), then a model other than $Y = \beta_0 + \beta_1 x + \varepsilon$ may be fit. This can be achieved by transformation of the x and/or y variable, or by inclusion of higher-order polynomial terms (see Section 12.8).

Transformations of the y variable can also be used to remedy nonconstant variance. For instance, if the spread of the residuals grows with x as in the residual plot of Figure 12.24b, the transformation $y' = \ln(y)$ is often applied. Another popular approach to addressing nonconstant variance is the method of **weighted least squares**. The basic idea of weighted least squares is to find coefficients b_0 and b_1 to minimize the expression

$$g_w(b_0, b_1) = \sum w_i [y_i - (b_0 + b_1 x_i)]^2$$

where the w_i's are "weights" determined by the variance structure of the errors. For example, if the standard deviation of Y is proportional to x (for $x > 0$)—that is, $V(Y) = kx^2$—then it can be shown that the weights $w_i = 1/x_i^2$ yield minimum-variance estimators of β_0 and β_1. The books by Kutner et al. and by Chatterjee and Hadi explain weighted least squares in detail (see the bibliography). Weighted least squares are used quite frequently by econometricians (economists who use statistical methods) to estimate parameters.

Generally speaking, violations of the normality assumption cannot be fixed, though such a problem may naturally be resolved while addressing linearity and constant variance issues. Again, if the sample size is reasonably large then normality is not as important, except for prediction intervals. If a small data set has the feature that *only* the normality assumption is violated, consult your friendly neighborhood statistician for information on computer-intensive methods (e.g., bootstrapping).

When plots or other evidence suggest that the data set contains outliers or points having large influence on the resulting fit, one possible approach is to omit these outlying points and re-compute the estimated regression equation. This would certainly be correct if it was found that the outliers resulted from errors in recording data values or experimental errors. If no assignable cause can be found for the outliers, it is still desirable to report the estimated equation both with and without outliers. Another approach is to retain possible outliers but to use an estimation principle that puts relatively less weight on outlying values than does the principle of least squares. One such principle is minimize absolute deviations (MAD), which selects b_0 and b_1 to minimize $\sum |y_i - (b_0 + b_1 x_i)|$. Unlike the least squares estimates, there are no nice formulas for the MAD estimates; their values

must be found by using an iterative computational procedure. Such procedures are also used when it is suspected that the ε_i's have a distribution that is not normal but instead has "heavy tails" (making it much more likely than for the normal distribution that discrepant values will enter the sample); *robust regression* procedures are those that produce reliable estimates for a wide variety of underlying error distributions. Least squares estimators are not robust, in the same way that the sample mean $\overline{X}$ is not a robust estimator for μ.

Exercises: Section 12.6 (67–76)

67. The x values and standardized residuals for the temperature-energy use data of Exercise 49 (Section 12.4) are displayed in the accompanying table. Construct a standardized residual plot and comment on its appearance.

x	48	53	56	58	58
e^*	−0.110	−1.153	−1.077	0.486	0.836
x	59	59	60	68	69
e^*	0.560	1.258	−0.148	−0.660	−1.869
x	69	70	73	75	79
e^*	0.356	0.858	0.743	0.724	−0.622
x	80	80	84	87	88
e^*	−0.460	1.471	−2.133	1.181	−0.302

68. Suppose the variables x = commuting distance and y = commuting time are related according to the simple linear regression model with $\sigma = 10$.

    a. If $n = 5$ observations are made at the x values $x_1 = 5$, $x_2 = 10$, $x_3 = 15$, $x_4 = 20$, and $x_5 = 25$, calculate the (true) standard deviations of the five corresponding residuals.

    b. Repeat part (a) for $x_1 = 5$, $x_2 = 10$, $x_3 = 15$, $x_4 = 20$, and $x_5 = 50$.

    c. What do the results of parts (a) and (b) imply about the deviation of the estimated line from the observation made at the largest sampled x value?

69. Nickel-based alloys are especially difficult to machine due to characteristics including high hardness and low thermal conductivity. The article "Multi-response Optimization Using ANOVA and Desirability Function Analysis: A Case Study in End Milling of Inconel Alloy" (*ARPN J. Engr.*

Appl. Sci. 2014: 457–463) reports the following data on x = cutting velocity (m/min) and y = material removal rate (mm^2/min) from one experiment.

x	25	25	25	50	50
y	258.48	268.80	270.18	338.58	343.86

x	50	75	75	75
y	354.24	414.36	424.80	451.80

    a. The LSRL for this data is $y = 182.7 + 3.29x$. Calculate and plot the residuals against x and then comment on the appropriateness of the simple linear regression model.

    b. Use $s_e = 11.759$ to calculate the standardized residuals from a simple linear regression. Construct a standardized residual plot and comment. Also construct a normal probability plot and comment.

70. As the air temperature drops, river water becomes supercooled and ice crystals form. Such ice can significantly affect the hydraulics of a river. The article "Laboratory Study of Anchor Ice Growth" (*J. Cold Regions Engr.* 2001: 60–66) described an experiment in which ice thickness (mm) was studied as a function of elapsed time (hr) under specified conditions. The following data was read from a graph in the article: $n = 33$; $x = .17, .33, .50, .67, \ldots, 5.50$; $y = .50, 1.25, 1.50, 2.75, 3.50, 4.75, 5.75, 5.60, 7.00, 8.00, 8.25, 9.50, 10.50, 11.00, 10.75, 12.50, 12.25, 13.25, 15.50, 15.00, 15.25, 16.25, 17.25, 18.00, 18.25, 18.15, 20.25, 19.50, 20.00, 20.50, 20.60, 20.50, 19.80$.

a. The R^2 value resulting from a least squares fit is .977. Given the high R^2, does it seem appropriate to assume an approximate linear relationship?

b. The residuals, listed in the same order as the x values, are

−1.03	−0.92	−1.35	−0.78	−0.68	−0.11	0.21
−0.59	0.13	0.45	0.06	0.62	0.94	0.80
−0.14	0.93	0.04	0.36	1.92	0.78	0.35
0.67	1.02	1.09	0.66	−0.09	1.33	−0.10
−0.24	−0.43	−1.01	−1.75	−3.14		

Plot the residuals against x, and reconsider the question in (a). What does the plot suggest?

71. The accompanying data on x = true density (kg/mm^3) and y = moisture content (% d. b.) was read from a plot in the article "Physical Properties of Cumin Seed" (*J. Agric. Engr. Res*. 1996: 93–98).

x	7.0	9.3	13.2	16.3	19.1	22.0
y	1046	1065	1094	1117	1130	1135

The equation of the least squares line is $y = 1008.14 + 6.19268x$ (this differs very slightly from the equation given in the article); $s_e = 7.265$ and $R^2 = .968$.

a. Carry out a test of model utility and comment.

b. Compute the values of the residuals and plot the residuals against x. Does the plot suggest that a linear regression function is inappropriate?

c. Compute the values of the standardized residuals and plot them against x. Are there any unusually large (positive or negative) standardized residuals? Does this plot give the same message as the plot of part (b) regarding the appropriateness of a linear regression function?

72. Continuous recording of heart rate can be used to obtain information about the level of exercise intensity or physical strain during sports participation, work, or other daily activities. The article "The Relationship Between Heart Rate and Oxygen Uptake During Non-Steady State Exercise" (*Ergonomics* 2000: 1578–1592) reported on a study to investigate using heart rate response (x, as a percentage of the maximum rate) to predict oxygen uptake (y, as a percentage of maximum uptake) during exercise. The accompanying data was read from a graph in the paper.

HR	43.5	44.0	44.0	44.5	44.0	45.0	48.0	49.0
VO$_2$	22.0	21.0	22.0	21.5	25.5	24.5	30.0	28.0

HR	49.5	51.0	54.5	57.5	57.7	61.0	63.0	72.0
VO$_2$	32.0	29.0	38.5	30.5	57.0	40.0	58.0	72.0

Use a statistical software package to perform a simple linear regression analysis. Considering the list of potential difficulties in this section, see which of them apply to this data set.

73. Example 12.6 presented the residuals from a simple linear regression of moisture content y on filtration rate x.

a. Plot the residuals against x. Does the resulting plot suggest that a straight-line regression function is a reasonable choice of model? Explain your reasoning.

b. Using $s_e = .665$, compute the values of the standardized residuals. Is $e_i^* \approx e_i/s_e$ for $i = 1, \ldots, n$, or are the e_i^*'s not close to being proportional to the e_i's?

c. Plot the standardized residuals against x. Does the plot differ significantly in general appearance from the plot of part (a)?

74. Express the ith residual $Y_i - \hat{Y}_i$ (where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) in the form $\sum c_j Y_j$, a linear function of the Y_j's. Then use rules of variance to verify that $V(Y_i - \hat{Y}_i)$ is given by Expression (12.11).

75. Consider the following classic four (x, y) data sets; the first three have the same x values, so these values are listed only once (Frank Anscombe, "Graphs in Statistical Analysis," *Amer. Statist*. 1973: 17–21):

1–3	1	2	3	4	4
x	y	y	y	x	y
10.0	8.04	9.14	7.46	8.0	6.58
8.0	6.95	8.14	6.77	8.0	5.76
13.0	7.58	8.74	12.74	8.0	7.71
9.0	8.81	8.77	7.11	8.0	8.84
11.0	8.33	9.26	7.81	8.0	8.47
14.0	9.96	8.10	8.84	8.0	7.04
6.0	7.24	6.13	6.08	8.0	5.25
4.0	4.26	3.10	5.39	19.0	12.50
12.0	10.84	9.13	8.15	8.0	5.56
7.0	4.82	7.26	6.42	8.0	7.91
5.0	5.68	4.74	5.73	8.0	6.89

For each of these four data sets, the values of $\bar{x}$, $\bar{y}$, S_{xx}, S_{yy}, and S_{xy} are virtually identical, so all quantities computed from these five will be essentially identical for the four sets—the equation of the least squares line ($y = 3 + .5x$), SSE, s_e, R^2, t intervals, t statistics, and so on. The summaries provide no way of distinguishing among the four data sets. Based on a scatterplot and a residual plot for each set, comment on the appropriateness or inappropriateness of fitting a straight-line model; include in your comments any specific suggestions for how a "straight-line analysis" might be modified or qualified.

76. If there is at least one x value at which more than one observation has been made, the **lack of fit test** is a formal procedure for testing

H_0: $\mu_{Y|x} = \beta_0 + \beta_1 x$ for some values β_0, β_1 (the true regression function is linear)

   versus

H_a: H_0 is not true (the true regression function is not linear)

Suppose observations are made at c levels $x_1, x_2, \ldots, x_c$. Let $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ denote the n_i observations when $x = x_i$ ($i = 1, \ldots, c$). With $n = \sum n_i$, SSE has $n - 2$ df. We break SSE into two pieces, SSPE (pure error) and SSLF (lack of fit), as follows:

$$SSPE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$$

$$SSLF = SSE - SSPE$$

The n_i observations at x_i contribute $n_i - 1$ df to SSPE, so the number of degrees of freedom for SSPE is $\sum_i (n_i - 1) = n - c$, and the degrees of freedom for SSLF is $n - 2 - (n - c) = c - 2$. Let MSPE = SSPE/$(n - c)$, MSLF = SSLF/$(c - 2)$. Then it can be shown that whereas $E(MSPE) = \sigma^2$ whether or not H_0 is true, $E(MSLF) = \sigma^2$ if H_0 is true and $E(MSLF) > \sigma^2$ if H_0 is false.

Test statistic: $F = MSLF/MSPE$

Rejection region: $f \geq F_{\alpha, c-2, n-c}$

The following data comes from the article "Single Cell Isolation Process with Laser Induced Forward Transfer" (*J. Bio. Engr.* 2017), with x = laser pulse energy (µJ), w = titanium thickness (nm) and y = number of viable cells resulting from a new cell-isolation technique.

x	5	5	5	5	5	5	5	5	5
w	80	40	120	80	40	120	80	40	120
y	24	21	16	25	22	14	24	22	14

x	9	9	9	9	9	9	9	9	9
w	80	40	120	80	40	120	80	40	120
y	18	35	24	16	37	25	19	37	24

x	13	13	13	13	13	13	13	13	13
w	80	40	120	80	40	120	80	40	120
y	31	47	42	29	46	38	33	38	40

a. Construct a scatterplot of y vs x. Does it appear that x and the mean of y are linearly related?

b. Carry out the lack of fit test on the (x, y) data at significance level .05.

c. Repeat parts (a) and (b) for y vs w.

12.7 Multiple Regression Analysis

In *multiple regression*, the objective is to build a probabilistic model that relates a response variable y to more than one explanatory or predictor variable. Let k represent the number of predictor variables ($k \geq 2$) and denote these predictors by $x_1, x_2, \ldots, x_k$. For example, in attempting to predict the selling price of a house, we might have $k = 4$ with x_1 = size (ft^2), x_2 = age (years), x_3 = number of bedrooms, and x_4 = number of bathrooms.

THE MULTIPLE LINEAR REGRESSION MODEL	There are parameters $\beta_0, \beta_1, \ldots, \beta_k$ and σ such that for any fixed values of the explanatory variables $x_1, \ldots, x_k$, the response variable Y is related to the x_j's through the model equation $$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \qquad (12.12)$$ where the random variable ε is assumed to follow a $N(0, \sigma)$ distribution. It is also assumed that the ε_i's (and thus the Y_i's) associated with different observations are independent of one another.

As before, ε is the random error term (or random deviation) in the model, and the assumptions for statistical inference may be stated in terms of ε. Equation (12.12) says that the **true** (or **population**) **regression function**, $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, gives the expected value of Y as a function of $x_1, \ldots, x_k$. The β_j's are the **true** (or **population**) **regression coefficients**.

Interpret the regression coefficients carefully! Performing multiple regression on k explanatory variables is not the same thing as creating k separate, simple linear regression models. For example, β_1 (the coefficient on the predictor x_1) cannot be interpreted in multiple regression without reference to the other predictor variables in the model. Here's a correct interpretation:

$$\beta_1 = \text{the change in the average value of } y \text{ associated with a one-unit increase in } x_1,$$
adjusting for the effects of the other explanatory variables

As an example, for the four predictors of home price mentioned above, β_1 is interpreted as the change in expected selling price when size increases by 1 ft^2, adjusting for the effects of age, number of bedrooms, and number of bathrooms. The other coefficients are interpreted similarly.

Some statisticians refer to β_1 as describing the effect of x_1 on y "after removing the effects of" the other predictors or "in the presence of" the other predictors; either of these interpretations is acceptable. You may hear β_1 defined as the change in the mean of y associated with a one-unit increase in x_1 "while holding the other variables fixed." This is correct only if it is possible to increase the value of one predictor while the values of all others remain constant.

Estimating Parameters

The data in simple linear regression consists of n pairs $(x_1, y_1), \ldots, (x_n, y_n)$. Suppose that a multiple regression model contains two explanatory variables, x_1 and x_2. Then each observation will consist of three numbers (a triple): a value of x_1, a value of x_2, and a value of y. More generally, with k predictor variables, each observation will consist of $k + 1$ numbers (a "$k + 1$ tuple"). The values of the predictors in the individual observations are denoted using double-subscripting:

$$x_{ij} = \text{the value of the } j\text{th predictor } x_j \text{ in the } i\text{th observation}$$
$$(i = 1, \ldots, n; j = 1, \ldots, k)$$

Thus the first subscript is the observation number and the second subscript is the predictor number. For example, x_{83} is the value of the third predictor in the eighth observation (to avoid confusion, a comma can be inserted between the two subscripts, e.g. $x_{12,3}$). The first observation in our data set is then $(x_{11}, x_{12}, \ldots, x_{1k}, y_1)$, the second is $(x_{21}, x_{22}, \ldots, x_{2k}, y_2)$, and so on.

Consider candidates $b_0, b_1, \ldots, b_k$ for estimates of the β_j's and the corresponding candidate regression function $b_0 + b_1 x_1 + \cdots + b_k x_k$. Substituting the predictor values for any individual observation into this candidate function gives a prediction for the y value that would be observed, and subtracting this prediction from the actual observed y value gives the prediction error. As in Section 12.2, the **principle of least squares** says we should square these prediction errors, sum, and then take as the least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ the values of the b_j's that minimize the sum of squared prediction errors. To carry out this procedure, define the criterion function (sum of squared prediction errors) by

$$g(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} \left[y_i - (b_0 + b_1 x_{i1} + \ldots + b_k x_{ik}) \right]^2,$$

then take the partial derivative of $g(\cdot)$ with respect to each b_j ($j = 0, 1, \ldots, k$) and equate these $k + 1$ partial derivatives to 0. The result is a system of $k + 1$ equations, the **normal equations**, in the $k + 1$ unknowns (the b_j's):

$$nb_0 + \left(\sum x_{i1}\right)b_1 + \left(\sum x_{i2}\right)b_2 + \cdots + \left(\sum x_{ik}\right)b_k = \sum y_i$$
$$\left(\sum x_{i1}\right)b_0 + \left(\sum x_{i1}^2\right)b_1 + \left(\sum x_{i1}x_{i2}\right)b_2 + \cdots + \left(\sum x_{i1}x_{ik}\right)b_k = \sum x_{i1}y_i$$
$$\vdots$$
$$\left(\sum x_{ik}\right)b_0 + \left(\sum x_{i1}x_{ik}\right)b_1 + \cdots + \left(\sum x_{i,k-1}x_{ik}\right)b_{k-1} + \left(\sum x_{ik}^2\right)b_k = \sum x_{ik}y_i$$

$$(12.13)$$

Notice that the normal equations are linear in the unknowns (because the criterion function is quadratic). We will assume that the system has a unique solution, the least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. The result is an **estimated regression function**

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

Section 12.9 uses matrix algebra to deal with the system of equations and develop inferential procedures for multiple regression. For the moment, though, we shall take advantage of the fact that all of the commonly used statistical software packages are programmed to solve the equations and provide the results needed for inference.

Sometimes interest in the individual regression coefficients is the main reason for a regression analysis. The article "Autoregressive Modeling of Baseball Performance and Salary Data" (*Proc. of the Statistical Graphics Section*, *Amer. Stat. Assoc.* 1988, 132–137) describes a multiple regression of runs scored as a function of singles, doubles, triples, home runs, and walks (combined with hit-by-pitcher). The estimated regression equation is

$$\text{runs} = -2.49 + .47 \text{ singles} + .76 \text{ doubles} + 1.14 \text{ triples} + 1.54 \text{ home runs} + .39 \text{ walks}$$

This is very similar to the popular slugging percentage statistic, which gives weight 1 to singles, 2 to doubles, 3 to triples, and 4 to home runs. However, the slugging percentage gives no weight to walks, whereas the estimated regression function puts weight .39 on walks, more than 80% of the weight it assigns to singles. The importance of walks is well known among statisticians who follow baseball, and it is interesting that there are now some statistically savvy people in major league baseball management who are emphasizing walks in choosing players.

Example 12.20 Fuel efficiency of an automobile is determined to a large extent by various intrinsic characteristics of the vehicle. Consider the following multivariate data set consisting of $n = 38$ observations on x_1 = weight (1000s of pounds), x_2 = engine displacement (i.e., engine size, in^3), x_3 = number of cylinders, and y = fuel efficiency, measured in gallons per 100 miles:

x_1	x_2	x_3	y	x_1	x_2	x_3	y
4.360	350	8	5.92	3.830	318	8	5.49
4.054	351	8	6.45	2.585	140	4	3.77
3.605	267	8	5.21	2.910	171	6	4.57
3.940	360	8	5.41	1.975	86	4	2.93
2.155	98	4	3.33	1.915	98	4	2.85
2.560	134	4	3.64	2.670	121	4	3.65
2.300	119	4	3.68	1.990	89	4	3.17
2.230	105	4	3.24	2.135	98	4	3.39
2.830	131	5	4.93	2.670	151	4	3.52
3.140	163	6	5.88	2.595	173	6	3.47
2.795	121	4	4.63	2.700	173	6	3.73
3.410	163	6	6.17	2.556	151	4	2.99
3.380	231	6	4.85	2.200	105	4	2.92
3.070	200	6	4.81	2.020	85	4	3.14
3.620	225	6	5.38	2.130	91	4	2.68
3.410	258	6	5.52	2.190	97	4	3.28
3.840	305	8	5.88	2.815	146	6	4.55
3.725	302	8	5.68	2.600	121	4	4.65
3.955	351	8	6.06	1.925	89	4	3.13

We've chosen to use this representation of fuel efficiency (similar to European measurement), rather than the traditional American "miles per gallon" version, because the former is linearly related to our predictors while the latter is not. One consequence is that *lower y* values are better, in the sense that they indicate vehicles with better fuel efficiency.

Our goal is to predict fuel efficiency (y) from the predictor variables x_1, x_2, and x_3 (so $k = 3$). Figure 12.27 shows R output from a request to fit a linear function to the fuel efficiency data. The least squares coefficients appear in the estimate column of the coefficients block:

$$\hat{\beta}_0 = -1.64351 \quad \hat{\beta}_1 = 2.33584 \quad \hat{\beta}_2 = -0.01065 \quad \hat{\beta}_3 = 0.21774$$

Thus the estimated regression equation is

$$y = -1.64351 + 2.33584x_1 - 0.01065x_2 + 0.21774x_3$$

Consider an automobile that weighs 3000 lbs, has a displacement of 175 in^3, and is equipped with a six-cylinder engine. A prediction for the resulting fuel efficiency is obtained by substituting the values of the predictors into the fitted equation:

```
Call:
lm(formula = Fuel.Eff ~ Weight + Disp + Cyl, data = df)

Residuals:
     Min       1Q    Median       3Q      Max
-0.63515 -0.30801   0.00029  0.23637  0.63957

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.64351    0.48512   -3.388 0.001795 **
Weight       2.33584    0.28810    8.108 1.87e-09 ***
Disp        -0.01065    0.00269   -3.959 0.000364 ***
Cyl          0.21774    0.11566    1.883 0.068342 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3749 on 34 degrees of freedom
Multiple R-squared:  0.9034,  Adjusted R-squared:  0.8948
F-statistic: 105.9 on 3 and 34 DF,  p-value: < 2.2e-16
```

Figure 12.27 Multiple regression output for Example 12.20

$$\hat{y} = -1.64351 + 2.33584(3) - 0.01065(175) + 0.21774(6) = 4.8067$$

Such an automobile is predicted to use 4.8 gallons over 100 miles. This is also a point estimate of the mean fuel efficiency of all vehicles with these specifications ($x_1 = 3$, $x_2 = 175$, $x_3 = 6$).

The intercept $\hat{\beta}_0 = -1.64351$ has no contextual meaning here, since you can't really have a vehicle with no weight and no engine. The coefficient $\hat{\beta}_1 = 2.33584$ on x_1 indicates that, after adjusting for both engine displacement (x_2) and number of cylinders (x_3), a 1000-lb increase in weight is associated with an estimated increase of about 2.34 gallons per 100 miles, on average. Equivalently, a 100-lb weight increase is predicted to increase average fuel consumption by 0.234 gallons per 100 miles, accounting for engine size (i.e., displacement and number of cylinders).

Consider fitting the simple linear model to just y and x_3. The resulting LSRL is

$$y = 1.057 + 0.6067\,x_3,$$

suggesting that a one-cylinder increase is associated with about a 0.6 gallon increase in average fuel consumption per 100 miles. This is our estimate of the relationship between x_3 and y *ignoring the effects of any other explanatory variables*; notice that it differs substantially from the previous estimate of 0.21774, which adjusted for both the vehicle's weight and its engine displacement. (Since these cars have 4, 6, or 8 cylinders, it's perhaps more appropriate to double these coefficients as an indication of the effect of a two-cylinder increase.) ■

Descriptive Measures of Fit
The predicted (or fitted) value $\hat{y}_i$ results from substituting the values of the predictors from the ith observation into the estimated equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$$

The corresponding residual is $e_i = y_i - \hat{y}_i$. As in simple linear regression, the closer the residuals are to zero, the better the job our estimated regression function is doing in predicting the y values in our sample. For the fuel efficiency data, the values of the three predictors in the first observation are $x_{11} = 4.360$, $x_{12} = 350$, and $x_{13} = 8$, so

$$\hat{y}_1 = -1.64351 + 2.33584(4.360) - 0.01065(350) + 0.21774(8) = 6.555$$
$$e_1 = y_1 - \hat{y}_1 = 5.92 - 6.555 = -0.635$$

Residuals are sometimes important not just for judging the quality of a regression. Several enterprising students developed a multiple regression model using age, size in square feet, etc. to predict the price of four-unit apartment buildings. They found that one building had a large negative residual, meaning that the price was much lower than predicted. As it turned out, the reason was that the owner had "cash-flow" problems and needed to sell quickly.

In simple linear regression, after fitting a straight line to bivariate data and obtaining the residuals, we calculated sums of squares and used them to obtain two assessments of how well the line summarized the relationship: the residual standard deviation and the coefficient of determination. Let's now follow the same path in multiple regression:

$$\text{SSE} = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \quad \text{SSR} = \sum (\hat{y}_i - \bar{y})^2 \quad \text{SST} = \sum (y_i - \bar{y})^2$$

These are the same expressions introduced in Section 12.2, and again it can be shown that SST = SSE + SSR. The interpretation is that the total variation in the values of the response variable is the sum of explained variation (SSR) and unexplained variation (SSE).

Each sum of squares has an associated number of degrees of freedom (df). In particular,

$$\text{df for SSE} = n - (k+1)$$

This is because the $k + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_k$ must be estimated before SSE can be obtained, entailing a reduction of $k + 1$ df for SSE. Notice that for the case of simple linear regression, $k = 1$ and df for SSE $= n - (1 + 1) = n - 2$ as before.

DEFINITION   The standard deviation about the estimated multiple regression function (or simply the **residual standard deviation**) is

$$s_e = \sqrt{\frac{\text{SSE}}{n - (k+1)}}$$

The **coefficient of (multiple) determination**, denoted by R^2, is given by

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}$$

Roughly speaking, s_e is the size of a typical deviation from the fitted equation. The residual standard deviation is also our point estimate of the model parameter σ, i.e., $\hat{\sigma} = s_e$. R^2 is the proportion of variation in observed y values that can be explained by (i.e., attributed to) the multiple regression model. Software often reports $100R^2$, the percent of explained variation. The closer R^2 is to 1, the

larger the proportion of observed y variation that is being explained. (In fact, the positive square root of R^2, called the *multiple correlation coefficient*, turns out to be the sample correlation coefficient between the observed y_i's and the predicted $\hat{y}_i$'s—another measure of the quality of the fit of the estimated regression function.)

Unfortunately, there is a potential problem with R^2 in multiple regression: its value can be inflated by including predictors in the model that are relatively unimportant or even frivolous. For example, suppose we plan to obtain a sample of 20 recently sold houses in order to relate sale price to various characteristics of a house. Natural predictors include interior size, lot size, age, number of bedrooms, and distance to the nearest school. Suppose we also include in the model the diameter of the doorknob on the door of the master bedroom, the height of the toilet bowl in the master bath, and so on until we have 19 predictors. Then unless we are extremely unlucky in our choice of predictors, the value of R^2 will be 1 (because 20 coefficients are perfectly estimated from 20 observations)! Rather than seeking a model that has the highest possible R^2 value, which can be achieved just by "packing" our model with predictors, what is desired is a relatively simple model based on just a few important predictors whose R^2 value is high.

It is therefore desirable to adjust R^2 to take account of the fact that its value may be quite high just because many predictors were used relative to the amount of data. The **adjusted coefficient of multiple determination** or **adjusted R^2** is defined by

$$R_a^2 = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{\text{SSE}/[n - (k+1)]}{\text{SST}/(n-1)} = 1 - \frac{n-1}{n-(k+1)}\frac{\text{SSE}}{\text{SST}}$$

The ratio multiplying SSE/SST in adjusted R^2 exceeds 1 (the denominator is smaller than the numerator), so adjusted R^2 is smaller than R^2 itself, and in fact will be much smaller when k is large relative to n. A value of R_a^2 much smaller than R^2 is a warning flag that the chosen model has too many predictors relative to the amount of data.

Example 12.21 (Example 12.20 continued) Figure 12.27 shows that $s_e \approx 0.3749$, $R^2 = 90.34\%$, and $R_a^2 = 89.48\%$ for the fuel efficiency data. Fuel efficiency predictions based on the estimated regression equation are typically off by about 0.375 gal/100 mi (positive or negative) from vehicles' actual fuel efficiency. The model explains about 90% of the observed variation in fuel efficiency. ∎

A Model Utility Test

In multiple regression, is there a single indicator that can be used to judge whether a particular model (equivalently, a particular set of predictors $x_1, \ldots, x_k$) will be useful? The value of R^2 certainly communicates a preliminary message, but this value is sometimes deceptive because it can be greatly inflated by using a large number of predictors (large k) relative to the sample size n (this is the rationale behind adjusting R^2).

The model utility test in *simple* linear regression involved the null hypothesis H_0: $\beta_1 = 0$, according to which there is no useful relation between y and the single predictor x. Here we consider the assertion that $\beta_1 = 0$, $\beta_2 = 0$, $\ldots$, $\beta_k = 0$, which says that there is no useful relationship between y and *any* of the k predictors. If at least one of these β_j's is not 0, the corresponding predictor(s) is (are) useful. The test is based on the F statistic derived from the regression ANOVA table (see Sections 10.5 and 11.1 for more about F tests). A prototype multiple regression ANOVA table appears in Table 12.3.

Table 12.3 ANOVA table for multiple regression

Source of variation	df	Sum of squares	Mean square	f
Regression	k	SSR	MSR = SSR/k	MSR/MSE
Error	$n - (k + 1)$	SSE	MSE = SSE/$[n - (k + 1)]$	
Total	$n - 1$	SST		

MODEL UTILITY TEST IN MULTIPLE REGRESSION

Null hypothesis: H_0: $\beta_1 = 0$, $\beta_2 = 0$, ..., $\beta_k = 0$

Alternative hypothesis: H_a: at least one $\beta_j \neq 0$

Test statistic value: $f = \dfrac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} = \dfrac{\text{MSR}}{\text{MSE}}$ (12.14)

When H_0 is true, the test statistic has an F distribution with k numerator df and $n - (k + 1)$ denominator df.

Rejection Region for a Level α Test   **P-value**

$f \geq F_{\alpha, k, n-(k+1)}$ area to the right of f under the $F_{k, n-(k+1)}$ curve

To understand why the test statistic value (12.14) should be compared to this particular F distribution, divide the fundamental ANOVA identity by σ^2:

$$\frac{\text{SST}}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} + \frac{\text{SSR}}{\sigma^2}$$

When H_0 is true, the observations $Y_1, \ldots, Y_n$ all have the same mean $\mu = \beta_0$ and common variance σ^2. It follows from a proposition in Section 6.4 that $\text{SST}/\sigma^2 \sim \chi_{n-1}^2$. It can also be shown that (1) $\text{SSE}/\sigma^2 \sim \chi_{n-(k+1)}^2$ and (2) SSE and SSR are independent. Together, (1) and (2) imply—again, see Section 6.4—that SSR/σ^2 has a chi-squared distribution, with df $= (n - 1) - (n - (k + 1)) = k$. Finally, by definition the F distribution is the ratio of two independent chi-squared rvs divided by their respective dfs. Applying this to SSR/σ^2 and SSE/σ^2 leads to the F ratio

$$\frac{\dfrac{\text{SSR}/\sigma^2}{k}}{\dfrac{\text{SSE}/\sigma^2}{n - (k + 1)}} = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} = \frac{\text{MSR}}{\text{MSE}} \sim F_{k, n-(k+1)}$$

The test statistic is identical in structure to the ANOVA F statistic from Chapter 11: the numerator measures the variation explained by the proposed model, while the denominator measures the unexplained variation. The larger the value of F, the stronger the evidence that a statistically significant relationship exists between y and the predictors. In fact, the model utility test statistic value can be re-written as

$$f = \frac{R^2}{1 - R^2} \cdot \frac{n - (k+1)}{k}$$

so the test statistic is proportional to $R^2/(1 - R^2)$, the ratio of the explained to unexplained variation. If the proportion of explained variation is high relative to unexplained, we would naturally want to reject H_0 and confirm the utility of the model. However, the factor $[n - (k + 1)]/k$ decreases as k increases, and if k is large relative to n, it will reduce f considerably.

Because the model utility test considers *all* of the explanatory variables simultaneously, it is sometimes called a **global F test**.

Example 12.22 What impacts the salary offers made to faculty in management and information science (MIS)? The accompanying table shows part of the data available on a sample of 167 MIS faculty, which includes the following variables (based on publicly available data provided by Prof. Dennis Galletta, University of Pittsburgh):

y = salary offer, in thousands of dollars
x_1 = year the salary offer was made, coded as years after 2000 (so 2009 is coded as 9)
x_2 = previous experience, in years
x_3 = teaching load, converted into the number of three-credit semester courses per year.

Observation	Salary (y)	Year (x_1)	Experience (x_2)	Teaching load (x_3)
1	90.0	3	5	3
2	91.5	3	12	4
3	105.0	4	7	4
4	79.2	6	3	5
5	95.0	9	0.5	6
⋮	⋮	⋮	⋮	⋮

The model utility test hypotheses are

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
H_a: at least one of these three β_j's is not 0

Figure 12.28 shows output from the JMP statistical package. The values of s_e (Root Mean Square Error), R^2, and adjusted R^2 certainly suggest a useful model. The value of the model utility F ratio is

$$f = \frac{\text{SSR}/k}{\text{SSE}/[n - (k+1)]} = \frac{20064.099/3}{21418.105/163} = \frac{6688.03}{131.40} = 50.8985$$

This value also appears in the F ratio column of the ANOVA table in Figure 12.28. Since $f = 50.8985 \geq F_{.01,3,163} \approx 3.9$, H_0 should be rejected at significance level .01. In fact, the ANOVA table in the JMP output shows that P-value < .0001. The null hypothesis should therefore be rejected at any reasonable significance level. We conclude that there is a useful linear relationship between y and *at least one* of the three predictors in the model. Note this does not mean that *all three* predictors are necessarily useful; we will say more about this shortly.

Summary of Fit

RSquare	0.48368
RSquare Adj	0.474177
Root Mean Square Error	11.46296
Mean of Response	95.42565
Observations (or Sum Wgts)	167

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	20064.099	6688.03	50.8985
Error	163	21418.105	131.40	**Prob > F**
C. Total	166	41482.204		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	115.55831	3.969873	29.11	<.0001*
Year	2.1853404	0.496914	4.40	<.0001*
Experience	0.2068286	0.232131	0.89	0.3742
TeachLoad	-6.580262	0.564235	-11.66	<.0001*

Figure 12.28 Multiple regression output from JMP for the data of Example 12.22   ■

Inferences about Individual Regression Coefficients

When the assumptions for the multiple linear regression model are met, we may also construct CIs and perform hypothesis tests for the individual population regression coefficients $\beta_1, \ldots, \beta_k$. Inferences concerning a single coefficient β_j are based on the standardized variable

$$T = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}}$$

which, assuming the model is correct, has a t distribution with $n - (k + 1)$ df. A matrix formula for $s_{\hat{\beta}_j}$ is given in Section 12.9, and the result is part of the output from all standard regression computer packages. A CI for β_j allows us to estimate with confidence the effect of the predictor x_j on the response variable, while adjusting for the effects of the other explanatory variables in the model.

By far the most commonly tested hypotheses about an individual β_j are H_0: $\beta_j = 0$ versus H_a: $\beta_j \neq 0$, in which case the test statistic value implifies to $t = \hat{\beta}_j / s_{\hat{\beta}_j}$. This null hypothesis states that, with the other explanatory variables in the model, the variable x_j does not provide any *additional* useful information about y. This is referred to as a **variable utility test**. It is sometimes the case that a predictor variable will be judged useful for predicting y under the simple linear regression model using x_j alone but not in the multiple regression setting (i.e., in the presence of other predictors). This usually indicates that the other variables do a better job predicting y, and that the additional

information in x_j is effectively "redundant." Occasionally, the opposite will happen: a predictor that isn't very useful by itself proves statistically significant in the presence of some other variables.

Next, let $\boldsymbol{x}^* = (x_1^*, \ldots, x_k^*)$ denote a particular value of $\boldsymbol{x} = (x_1, \ldots, x_k)$. Then the point estimate of $\mu_{Y|\boldsymbol{x}^*}$, the expected value of Y when $\boldsymbol{x} = \boldsymbol{x}^*$, is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_k x_k^*$. The estimated standard deviation $s_{\hat{Y}}$ of the corresponding estimator is a complicated expression involving the sample x_{ij}'s, but a matrix formula is given in Section 12.9. The better statistical software packages will calculate it on request. Inferences about $\mu_{Y|\boldsymbol{x}^*}$ are based on standardizing its estimator to obtain a t variable having $n - (k + 1)$ df.

---

1. A $100(1 - \alpha)\%$ CI for β_j, the coefficient of x_j in the model equation (12.12), is

$$\hat{\beta}_j \pm t_{\alpha/2, n-(k+1)} \cdot s_{\hat{\beta}_j}$$

2. A test for $H_0: \beta_j = \beta_{j0}$ uses the test statistic value $t = (\hat{\beta}_j - \beta_{j0})/s_{\hat{\beta}_j}$ based on $n - (k + 1)$ df. The test is upper-, lower-, or two-tailed according to whether H_a contains the inequality $>$, $<$, or $\neq$.

3. A $100(1 - \alpha)\%$ CI for $\mu_{Y|\boldsymbol{x}^*}$ is

$$\hat{y} \pm t_{\alpha/2, n-(k+1)} \cdot s_{\hat{Y}}$$

4. A $100(1 - \alpha)\%$ PI for a future y value is

$$\hat{y} \pm t_{\alpha/2, n-(k+1)} \cdot \sqrt{s_e^2 + s_{\hat{Y}}^2}$$

Simultaneous intervals for which the joint confidence or prediction level is controlled can be obtained by applying the Bonferroni technique discussed in Section 12.4.

Example 12.23 (Example 12.22 continued) The JMP output in Figure 12.28 includes variable utility tests and 95% confidence intervals for the coefficients. The results of testing $H_0: \beta_2 = 0$ versus $H_a: \beta_2 \neq 0$ (x_2 = years of experience) are

$$\hat{\beta}_2 = .2068, \quad s_{\hat{\beta}_2} = .2321, \quad t = .2068/.2321 = 0.89, \quad P\text{-value} = .3742$$

so H_0 is not rejected here. Adjusting for the year a salary offer was made and the position's teaching load, years of prior experience does not provide additional useful information about MIS faculty members' salary offers. The other two variables *are* useful (P-value $< .0001$ for each). A 95% CI for β_3 is

$$\hat{\beta}_3 \pm t_{.025, 167-(3+1)} s_{\hat{\beta}_3} = -6.58 \pm 1.975(.5642) = (-7.694, -5.466),$$

which agrees with the interval given in Figure 12.28. Thus after adjusting for offer year and prior experience, a one-course increase in teaching load is associated with a *decrease* in expected salary offer between \$5466 and \$7694. (If that seems counterintuitive, bear in mind that elite institutions can

offer both higher salaries and lighter teaching loads, while the opposite is true at a typical state university.)

The predicted salary offer in 2015 ($x_1 = 15$) for a newly minted PhD ($x_2 = 0$, no experience) and a five-course annual teaching load ($x_3 = 5$) is

$$\hat{y} = 115.558 + 2.18534(15) + .2068(0) - 6.580(5) = 115.437$$

(that is, \$115,437). The estimated standard deviation for this predicted value can be obtained from software: $s_{\hat{Y}} = 4.929$. So a 95% confidence interval for the *mean* offer under these settings is

$$\hat{y} \pm t_{.025,163} s_{\hat{Y}} = 115.437 \pm 1.975(4.929) = (105.704, 125.171)$$

which can also be obtained from software. A 95% prediction interval for a single future salary offer under these settings is

$$\hat{y} \pm t_{.025,163} \cdot \sqrt{s_e^2 + s_{\hat{Y}}^2} = 115.437 \pm 1.975\sqrt{11.463^2 + 4.929^2} = (90.798, 140.076)$$

Of course, the PI is much wider than the corresponding CI.

Since x_2 (years of prior experience) was deemed not useful, the model could also be re-fit without that variable, resulting in somewhat different estimated regression coefficients and, consequently, slightly different intervals above. ∎

Assessing Model Adequacy

The model assumptions of linearity, normality, constant variance, and independence are essentially the same for simple and multiple regression. Scatterplots of y versus each of the explanatory variables can give a preliminary sense of whether linearity is plausible, but the residual plots detailed in Section 12.6 are preferred. The standardized residuals in multiple regression result from dividing each residual e_i by its estimated standard deviation; a matrix formula for the standard deviation of e_i is given in Section 12.9. We recommend a normal probability plot of the standardized residuals as a basis for validating the normality assumption. Plots of the standardized residuals versus each predictor and/or versus $\hat{y}$ should show no discernible pattern. If the linearity and/or constant variance conditions appear violated, transformation of the response variable (possibly in tandem with transforming some x_j's) may be required. The book by Kutner et al. discusses transformations as well as other diagnostic plots.

Example 12.24 (Example 12.23 continued) Figure 12.29 shows a normal probability plot of the standardized residuals, as well as a plot of the standardized residuals versus the fitted values $\hat{y}_i$, for the MIS salary data. The probability plot is sufficiently straight that there is no reason to doubt the assumption of normally distributed errors. The residual-vs-fit plot shows no pattern, validating the linearity and constant variance assumptions.

Plots of the standardized residuals against the explanatory variables x_1, x_2, and x_3 (not shown) also exhibit no discernable pattern.
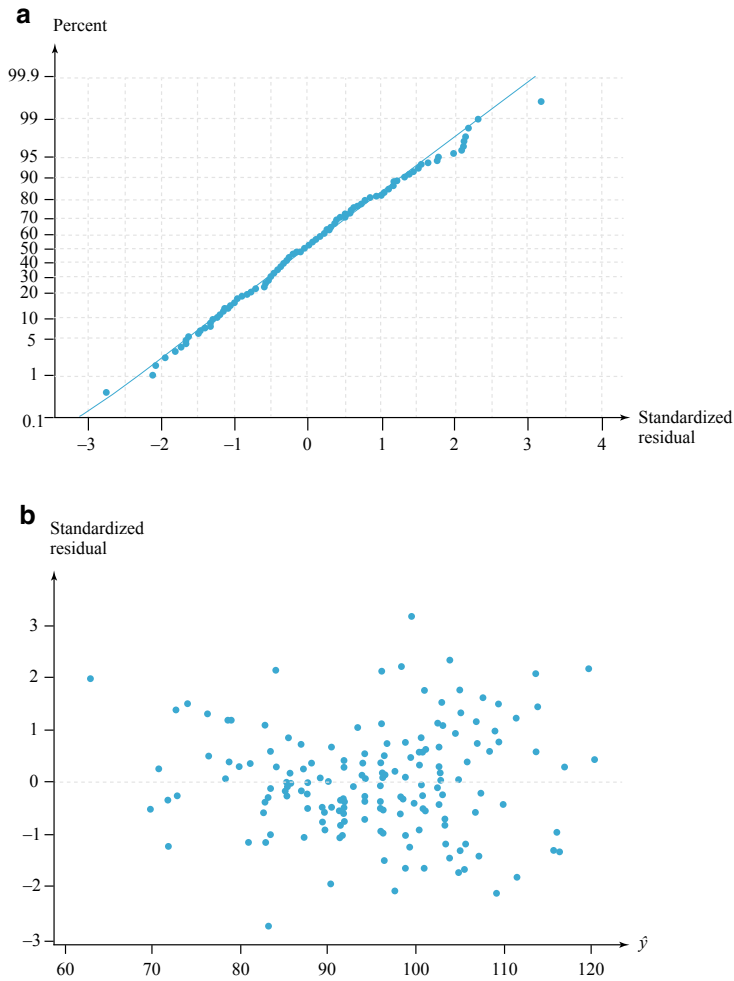
Figure 12.29 Residual plots for the MIS salary data   ■

Exercises: Section 12.7 (77–87)

77. Let y = weekly sales at a fast-food outlet (in dollars), x_1 = number of competing outlets within a 1-mile radius, and x_2 = population within a 1-mile radius (in thousands of people). Suppose that the true regression model is

$$Y = 10000 - 1400x_1 + 2100x_2 + \varepsilon$$

a. Determine expected sales when the number of competing outlets is 2 and there are 8000 people within a 1-mile radius.

b. Determine expected sales for an outlet that has three competing outlets and 5000 people within a 1-mile radius.

c. Interpret β_1 and β_2.

d. Interpret β_0. In what context does this value make sense?

78. Cardiorespiratory fitness is widely recognized as a major component of overall physical well-being. Direct measurement of maximal oxygen uptake (VO_2max) is the single best measure of such fitness, but direct measurement is time-consuming and expensive. It is therefore desirable to have a prediction equation for VO_2max in terms of easily obtained quantities. Consider the variables

$y = VO_2\text{max(L/min)}$ $x_1 = \text{weight(kg)}$
$x_2 = \text{age(yr)}$
$x_3 = \text{time necessary to walk 1 mile(min)}$
$x_4 = \text{heart rate at the end of the walk(beats/min)}$

Here is one possible model, for male students, consistent with the information given in the article "Validation of the Rockport Fitness Walking Test in College Males and Females" (Res. Q. Exerc. Sport 1994: 152–158):

$Y = 5.0 + .01x_1 - .05x_2 - .13x_3 - .01x_4 + \varepsilon$
$\sigma = .4$

a. Interpret β_1 and β_3.
b. What is the expected value of VO_2max when weight is 76 kg, age is 20 year, walk time is 12 min, and heart rate is 140 beats/min?
c. What is the probability that VO_2max will be between 1.00 and 2.60 for a single observation made when the values of the predictors are as stated in part (b)?

79. Athletic footwear is a multibillion-dollar industry, and manufacturers need to understand which features are most important to customers. The article "Overall Preference of Running Shoes Can Be Predicted by Suitable Perception Factors Using a Multiple Regression Model" (*Human Factors* 2017: 432–441) reports a survey of 100 young male runners in Beijing and Singapore. Each participant was asked to assess the Li Ning Hyper Arc (a running shoe) on five features: y = overall preference, x_1 = fit, x_2 = cushioning, x_3 = arch support and x_4 = stability. All measurements were made on a 0–15 visual analog scale, with $0 = dislike\ extremely$ and $15 = like\ extremely$.

a. The estimated regression equation reported in the article is $y = -.66 + .35x_1 + .34x_2 + .09x_3 + .32x_4$. Interpret the coefficient on x_2. [*Note*: The units are simply "points."]
b. Estimate the true mean rating from runners whose ratings on fit, cushioning, arch support, and stability are 9.0, 8.7, 8.9, and 9.2, respectively. (These were the average ratings across all 100 participants.) What would be more informative than this point estimate?
c. The authors report R^2 = .777 for this four-variable model. Perform a model utility test at the .01 significance level. Can we conclude that all four predictors provide useful information?
d. The article also reports variable utility test statistic values for each predictor; in order, they are t = 6.23, 4.92, 1.35, and 5.51. Perform all four variable utility tests at a simultaneous .01 level. Are all four predictors considered useful?

80. Roads in Egypt are notoriously haphazard, often lacking proper engineering consideration. The article "Effect of Speed Hump Characteristics on Pavement Condition" (*J. Traffic Transp. Engr.* 2017: 103–110) reports a study by two Egyptian civil engineering faculty of 52 speed bumps on a major road in upper Egypt. They measured each speed bump's height (x_1), width (x_2),

and distance (x_3) from the preceding speed bump (all in meters). They also evaluated each bump using a pavement condition index (PCI), a 0–100 scale with 100 the best condition.

a. With y = PCI, the estimated regression equation reported in the article is $y = 59.05 - 243.668x_1 + 11.675x_2 + .012x_3$. Interpret the coefficient on x_1. [*Hint:* Does a one-meter increase make sense?]

b. Estimate the pavement condition index of a speed bump 0.13 m tall, 2.5 m wide, and 666.7 m from the preceding speed bump.

c. The authors report $R^2 = .770$ for this three-variable model. Interpret this value, and then carry out the model utility test at the .05 significance level.

81. The accompanying data on sale price (thousands of dollars), size (thousands of sq. ft.), and land-to-building ratio for 10 large industrial properties appeared in the paper "Using Multiple Regression Analysis in Real Estate Appraisal" (*Appraisal J.* 2001: 424–430).

Price	Size	L/B ratio	Price	Size	L/B ratio
10600	2167	2.011	8000	2867	2.279
2625	752	3.543	10000	1698	3.123
10500	2423	3.632	6670	1046	4.771
1850	225	4.653	5825	1109	7.569
20000	3918	1.712	4517	405	17.190

a. Use software to create an estimated multiple regression equation for predicting the sale price of a property from its size and land-to-building ratio.

b. Interpret the estimated regression coefficients in this example.

c. Based on the data, what is the predicted sale price for a 500,000 ft.^{2} industrial property with a land-to-building ratio of 4.0?

82. There has been a shift recently away from using harsh chemicals to dye textiles in favor of natural plant extracts. The article "Ecofriendly Dyeing of Silk with Extract of Yerba Mate" (*Textile Res. J.* 2017: 829–837) describes an experiment to study the effects of dye concentration (mg/L), temperature (°C), and pH on dye adsorption (mg of dye per gram of fabric). [The article also included pictures of the resulting color from each treatment combination; dye adsorption is a proxy for color here.]

Conc.	Temp.	pH	Adsorption
10	70	3.0	250
20	70	3.0	520
10	90	3.0	387
20	90	3.0	593
10	70	4.0	157
20	70	4.0	377
10	90	4.0	225
20	90	4.0	451
15	80	3.5	353
15	80	3.5	382
15	80	3.5	373

a. Obtain the estimated regression equation for this data. Then, interpret the coefficient on temperature.

b. Calculate a point estimate for mean dye adsorption when concentration = 15 mg/L, temperature = 80 °C, and pH = 3.5 (i.e., the settings of the last three experimental runs).

c. The model utility test results in a test statistic value of $f = 44.02$. What can be concluded at the $\alpha = .01$ level?

d. Calculate and interpret a 95% CI for the settings specified in part (b).

e. Calculate and interpret a 95% PI for the settings specified in part (b).

f. Perform variable utility tests on each of the predictors. Can each one be judged useful provided that the other two are included in the model?

83. Carbon nanotubes (CNTs) are used for everything from structural reinforcement to communication antennas. The article "Fast Mechanochemical Synthesis of Carbon Nanotube-Polyanaline Hybrid Materials" (*J. Mater. Res.* 2018: 1486–1495) reported the following data on y = electrical conductivity (S/cm), x_1 = multi-walled CNT weight (mg), $x_2 = CN_x$ weight (mg), and x_3 = water volume (ml) from one experiment.

y	x_1	x_2	x_3
1.337	0.0	20.1	2
4.439	20.2	0.0	2
4.512	40.6	20.5	2
4.153	20.3	40.4	2
1.727	20.1	20.1	7
2.415	0.0	40.4	7
3.008	20.3	20.1	7
3.869	20.3	20.3	7
2.140	40.9	40.4	7
2.025	20.4	20.2	7
3.407	40.4	0.0	7
2.545	20.4	20.1	7
4.426	20.3	0.0	12
2.863	40.4	20.3	12
2.096	0.0	20.4	12
2.006	20.8	40.6	12

a. Obtain and interpret R^2 and s_e for the model with predictors x_1, x_2, and x_3.
b. Test for model utility using $\alpha = .05$.
c. Does the data support the manufacturing goal of (relatively) consistent electrical conductivity across differing values of the experimental factors? Explain.

84. Electric vehicles have greatly increased in popularity recently, but their short battery life (with a few exceptions) continues to be of concern. The article "Design of Robust Battery Capacity Model for Electric Vehicle by Incorporation of Uncertainties" (*Int. J. Energy Res.* 2017: 1436–1451) includes the following data on temperature (°C), discharge rate, and battery capacity (A-h, ampere-hours) for a certain type of lithium-ion battery.

Temp	Disch. rate	Capacity	Temp	Disch. rate	Capacity
0	0.50	0.96001	0	1.25	1.06506
0	1.75	0.85001	20	1.25	1.34459
0	3.00	0.89001	25	1.25	1.45473
25	0.50	1.38001	30	1.25	1.32355
40	1.75	1.05396	40	1.25	1.54713
40	3.00	0.96337	50	1.25	1.47159
20	0.50	1.27100	0	1.50	0.85171
20	1.75	1.24897	20	1.50	1.20890
20	3.00	1.20751	25	1.50	1.29703
25	3.00	1.32001	30	1.50	1.16097
50	3.00	1.39139	40	1.50	1.32047
40	0.50	1.00208	50	1.50	1.36305
50	0.50	1.44001	30	1.25	1.32355
25	1.75	1.30001	40	1.25	1.54713
30	1.75	0.82697	50	1.25	1.47159
50	1.75	1.42713	0	1.50	0.85171
0	1.00	1.08485	20	1.50	1.20890
20	1.00	1.40556	25	1.50	1.29703
25	1.00	1.47986	30	1.50	1.16097
30	1.00	0.91734	40	1.50	1.32047
40	1.00	1.18187	50	1.50	1.36305
50	1.00	1.53058			

a. Determine the estimated regression equation based on this data (y = capacity).
b. Calculate a point estimate for the expected capacity of a lithium-ion battery of this type when operated at 30 °C with discharge rate 1.00 (meaning the battery should drain in 1 h).
c. Perform a model utility test at the .05 significance level.
d. Calculate a 95% CI for expected capacity at the settings in part (b).
e. Calculate a 95% PI for the capacity of a single such battery at the settings in part (b).
f. Perform variable utility tests on both predictors at a simultaneous .05 significance level. Are both temperature and discharge rate useful predictors of capacity?

85. Steel microfibers are an alternative to conventional rebar reinforcement of concrete structures. The article "Measurement of Average Tensile Force for Individual Steel Fiber Using New Direct Tension Test"

(*J. Test. Eval.* 2016: 2403–2413) proposes a new evaluation method for concrete specimens infused with twisted steel micro rebar (TSMR) fibers, which were loaded until a crack occurred. The accompanying data on $y =$ load until cracking (lb), $x_1 =$ diameter at break (in), $x_2 =$ number of TSMR fibers per in^2, and $x_3 =$ concrete compressive strength (psi) appears in the article.

y	x_1	x_2	x_3	y	x_1	x_2	x_3
213	2.778	1.2	7020	688	2.735	3.4	6130
706	2.725	3.3	7020	180	2.705	2.6	6130
440	2.683	3.4	7020	1046	2.755	7.7	6130
984	2.835	8.2	5680	272	2.734	3.2	6880
155	2.779	2.3	5680	1168	2.725	7.4	7270
251	2.712	2.3	5680	821	2.750	6.1	7270
311	2.712	1.6	7960	418	2.740	4.4	7270
989	2.686	6.0	7960	1102	2.708	9.2	7390
326	2.821	4.0	7960	56	2.860	1.6	2950
479	2.810	2.7	7090	91	3.002	1.3	2950
324	2.845	1.9	7090	97	2.749	1.9	2950
404	2.755	3.0	7090				

a. Determine the estimated regression equation for this data.
b. Perform variable utility tests on each of the three explanatory variables. Can each one be judged useful given that the other two are included in the model?
c. Calculate R^2 and adjusted R^2 for this three-predictor model.
d. Perform a multiple regression of y on just x_2 and x_3. Determine both R^2 and adjusted R^2 for this reduced model. How do they compare to the values in part (c)? Explain.

86. An investigation of a die casting process resulted in the accompanying data on $x_1 =$ furnace temperature, $x_2 =$ die close time, and $y =$ temperature difference on the die surface ("A Multiple-Objective Decision-Making Approach for Assessing Simultaneous Improvement in Die Life and Casting Quality in a Die Casting Process," *Qual. Engr.* 1994: 371–383).

x_1	1250	1300	1350	1250	1300
x_2	6	7	6	7	6
y	80	95	101	85	92

x_1	1250	1300	1350	1350
x_2	8	8	7	8
y	87	96	106	108

Minitab output from fitting the multiple regression model with predictors x_1 and x_2 is given here.

```
The regression equation is
tempdiff = −200 + 0.210 furntemp
             + 3.00 clostime

Predictor      Coef    St dev   t ratio      p
Constant    −199.56     11.64    −17.14  0.000
furntemp   0.210000  0.008642     24.30  0.000
clostime     3.0000    0.4321      6.94  0.000
s = 1.058   R-sq = 99.1%      R-sq(adj) = 98.8%

Analysis of variance

Source      DF      SS      MS       F      P
Regression   2  715.50  357.75  319.31  0.000
Error        6    6.72    1.12
Total        8  722.22
```

a. Carry out the model utility test.
b. Calculate and interpret a 95% confidence interval for β_2, the population regression coefficient of x_2.
c. When $x_1 = 1300$ and $x_2 = 7$, the estimated standard deviation of $\hat{Y}$ is $s_{\hat{Y}} = .353$. Calculate a 95% confidence interval for true average temperature difference when furnace temperature is 1300 and die close time is 7.
d. Calculate a 95% prediction interval for the temperature difference resulting from a single experimental run with a furnace temperature of 1300 and a die close time of 7.
e. Use appropriate diagnostic plots to see if there is any reason to question the regression model assumptions.

87. The article "Analysis of the Modeling Methodologies for Predicting the Strength of Air-Jet Spun Yarns" (*Textile Res.*

J. 1997: 39–44) reported on a study carried out to relate yarn tenacity (y, in g/tex) to yarn count (x_1, in tex), percentage polyester (x_2), first nozzle pressure (x_3, in kg/cm^2), and second nozzle pressure (x_4, in kg/cm^2). The estimate of the constant term in the corresponding multiple regression equation was 6.121. The estimated coefficients for the four predictors were $-.082$, .113, .256, and $-.219$, respectively, and the coefficient of multiple determination was .946. Assume that $n = 25$.

a. State and test the appropriate hypotheses to decide whether the fitted model specifies a useful linear relationship between the response variable and at least one of the four model predictors.

b. Calculate the value of adjusted R^2 and comment.

c. Calculate a 99% confidence interval for true mean yarn tenacity when yarn count is 16.5, yarn contains 50% polyester, first nozzle pressure is 3, and second nozzle pressure is 5 if the estimated standard deviation of predicted tenacity under these circumstances is .350.

12.8 Quadratic, Interaction, and Indicator Terms

The fit of a multiple regression model can often be improved by creating new predictors from the original explanatory variables. In this section we discuss the two primary examples: *quadratic terms* and *interaction terms*. We also explain how to incorporate categorical predictor variables into the multiple regression model through the use of *indicator variables*.

Polynomial Regression

Let's return for a moment to the case of bivariate data consisting of n (x, y) pairs. Suppose that a scatterplot shows a parabolic rather than linear shape. Then it is natural to specify a **quadratic regression model**:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

The corresponding population regression function $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ gives the mean or expected value of Y for any particular x.

What does this have to do with multiple regression? Re-write the quadratic model equation as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad \text{where } x_1 = x \text{ and } x_2 = x^2$$

Now this looks exactly like a multiple regression equation with two predictors. Although we interpret this model as a quadratic function of x, the multiple linear regression model (12.12) only requires that the response be a linear function of the β_j's and ε. Nothing precludes one predictor being a mathematical function of another one. So, from a modeling perspective, *quadratic regression is a special case of multiple regression*. Thus any software package capable of carrying out a multiple regression analysis can fit the quadratic regression model. The same is true of cubic regression and even higher-order polynomial models, although in practice very rarely are such higher-order predictors needed.

The coefficient β_1 on the linear predictor x_1 cannot be interpreted as the change in expected Y when x_1 increases by one unit while x_2 is held fixed. This is because it is impossible to increase x without also increasing x^2. A similar comment applies to β_2. More generally, the interpretation of regression coefficients requires extra care when some predictor variables are mathematical functions of others.

Example 12.25 Reconsider the solar cell data of Example 12.2. Figure 12.2 clearly shows a parabolic relationship between x = sheet resistance and y = cell efficiency. To calculate the "least squares parabola" for this data, software fits a multiple regression model with two predictors: $x_1 = x$ and $x_2 = x^2$. The first few rows of data for this scenario are as follows:

y	$x_1 = x$	$x_2 = x^2$
13.91	43.58	1898.78
13.50	50.94	2594.63
13.59	60.03	3603.60
13.86	66.82	4464.91
$\vdots$	$\vdots$	$\vdots$

(In most software packages, it is not necessary to calculate x^2 for each observation; rather, the user can merely instruct the software to fit a quadratic model.) The coefficients that minimize the residual sum of squares are $\hat{\beta}_0 = 4.008$, $\hat{\beta}_1 = .3617$, and $\hat{\beta}_2 = -.003344$, so the estimated regression equation is

$$y = 4.008 + .3617x_1 - .003344x_2 = 4.008 + .3617x - .003344x^2$$

Figure 12.30 shows this parabola superimposed on a scatterplot of the original (x, y) data. Notice that the negative coefficient on x^2 matches the concave-downward contour of the data.
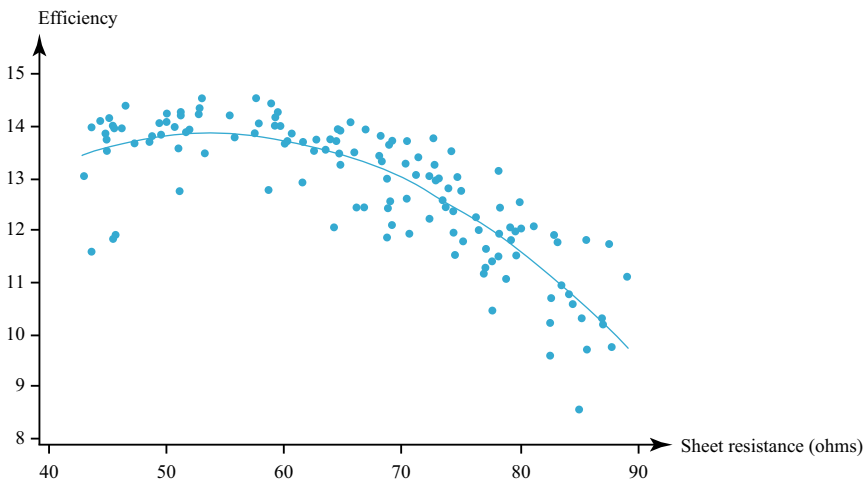

Figure 12.30 Scatterplot for Example 12.25 with a best-fit parabola

The estimated equation can now be used to make estimates and predictions at any particular x value. For example, the predicted efficiency at $x = 60$ ohms is determined by substituting $x_1 = x = 60$ and $x_2 = x^2 = 60^2 = 3600$:

$$y = 4.008 + .3617(60) - .003344(60)^2 = 13.67 \text{ percent}$$

Using software, a 95% CI for the mean efficiency of all 60-ohm solar panels is (13.50, 13.84), while a 95% PI for the efficiency of a single future 60-ohm panel is (12.32, 15.03). As always, the prediction interval is substantially wider than the confidence interval. ∎

Models with Interaction

Suppose that an industrial chemist is interested in the relationship between product yield (y) from a certain reaction and two explanatory variables, x_1 = reaction temperature and x_2 = pressure at which the reaction is carried out. The chemist initially proposes the relationship

$$Y = 1200 + 15x_1 - 35x_2 + \varepsilon$$

for temperature values between 80 and 100 in combination with pressure values ranging from 50 to 70. The population regression function $1200 + 15x_1 - 35x_2$ gives the mean y value for any particular values of the predictors. Consider this mean y value for three different particular temperatures:

$$x_1 = 90: \text{mean } y \text{ value} = 1200 + 15(90) - 35x_2 = 2550 - 35x_2$$
$$x_1 = 95: \text{mean } y \text{ value} = 2625 - 35x_2$$
$$x_1 = 100: \text{mean } y \text{ value} = 2700 - 35x_2$$

Graphs of these three mean y value functions are shown in Figure 12.31a. Each graph is a straight line, and the three lines are parallel, each with a slope of -35. Thus irrespective of the fixed value of temperature, the change in mean yield associated with a one-unit increase in pressure is -35.
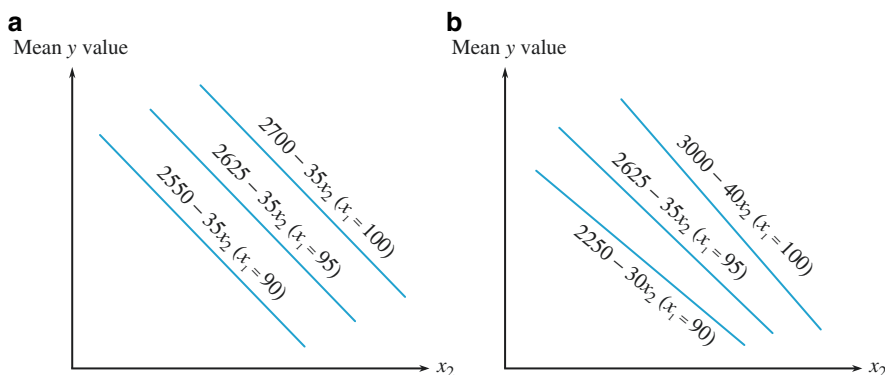

Figure 12.31 Graphs of the mean y value for two different models: (**a**) $1200 + 15x_1 - 35x_2$;
(**b**) $-4500 + 75x_1 + 60x_2 - x_1x_2$

In reality, when pressure increases the decline in average yield should be more rapid for a high temperature than for a low temperature, so the chemist has reason to doubt the appropriateness of the proposed model. Rather than the lines being parallel, the line for a temperature of 100 should be steeper than the line for a temperature of 95, and that line in turn should be steeper than the line for $x_1 = 90$. A model that has this property includes, in addition to x_1 and x_2, a third predictor variable, $x_3 = x_1 \cdot x_2$. One such model is

$$Y = -4500 + 75x_1 + 60x_2 - x_1x_2 + \varepsilon$$

for which the population regression function is $-4500 + 75x_1 + 60x_2 - x_1x_2$. This gives

$$\text{(mean } y \text{ value when temperature is } 100) = -4500 + (75)(100) + 60x_2 - 100x_2$$
$$= 3000 - 40x_2$$
$$\text{(mean value when temperature is } 95) = 2625 - 35x_2$$
$$\text{(mean value when temperature is } 90) = 2250 - 30x_2$$

These are graphed in Figure 12.31b. Now each different value of x_1 yields a line with a different slope, so the change in expected yield associated with a l-unit increase in x_2 depends on the value of x_1. When this is the case, the two predictor variables are said to *interact*.

DEFINITION   If the effect on y of one explanatory variable x_1 depends on the value of a second explanatory variable x_2, then x_1 and x_2 have an **interaction effect** on the (mean) response.

We can model this interaction by including as an additional predictor $x_3 = x_1x_2$, the product of the two explanatory variables, known as an **interaction term**.

The general equation for a multiple regression model based on two explanatory variables x_1 and x_2 and also including an interaction term is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \quad \text{where } x_3 = x_1 x_2$$

When an interaction effect is present, this model will usually give a much better fit to the data than would the no-interaction model. Failure to consider a model with interaction too often leads an investigator to conclude incorrectly that the relationship between y and a set of explanatory variables is not very substantial.

In applied work, quadratic predictors x_1^2 and x_2^2 are often also included, to model a potentially curved relationship. This leads to the **complete second-order model**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$$

This model replaces the straight lines of Figure 12.31 with parabolas (each one is the graph of the population regression function as x_2 varies when x_1 has a particular value).

Example 12.26 The need to devise environmentally friendly remedies for heavy-metal contaminated sites has become a global issue of serious concern. The article "Polyaspartate Extraction of Cadmium Ions from Contaminated Soil" (*J. Hazard. Mater.* 2018: 58–68) describes one possible cleanup method. Researchers varied five experimental factors: x_1 = polyaspartate (PA) concentration (mM), x_2 = PA-to-soil ratio, x_3 = initial cadmium (Cd) concentration in soil, x_4 = pH, and x_5 = extraction time (hours). One of the response variables of interest was y = residual Cd concentration, based on a total of $n = 47$ experimental runs.

Consider fitting a "first-order" model using all five predictors. Results from software include

$$y = -38.7 - .174x_1 - 1.308x_2 + .242x_3 + 10.26x_4 + 2.328x_5$$

$$s_e = 33.124(\text{df} = 41), R^2 = 71.25\%, R_a^2 = 67.74\%$$

Variable utility tests indicate that all predictors except x_1 are useful; it's possible that x_1 is redundant with x_2. At the other extreme, a complete second-order model here involves 20 predictor variables: the original x_j's (five), their squares (another five), and all $\binom{5}{2} = 10$ possible interaction terms: $x_1 x_2$, $x_1 x_3$, ..., $x_4 x_5$. Summary quantities from fitting this enormous model include $s_e = 24.878$ (df = 26), $R^2 = 89.72\%$, and $R_a^2 = 81.80\%$. The reduced standard deviation and greatly increased adjusted R^2 both suggest that at least some of the 15 second-order terms are useful, and so it was wise to incorporate these additional terms.

By considering the relative importance of the terms based on P-values, the researchers reduced their model to "just" 12 terms: all first-order terms, two of the quadratic terms, and five of the ten interactions. Based on the resulting estimated regression equation (shown in the article, but not here) researchers were able to determine the values of $x_1, \ldots, x_5$ that minimize residual Cd concentration. (Optimization is one of the added benefits of quadratic terms. For example, in Figure 12.30, we can see there is an x value for which solar cell efficiency is maximized. A linear model has no such local maxima or minima.)

It's worth noting that while x_1 was not considered useful in the first-order model, several second-order terms involving x_1 *were* significant. When fitting second-order models, it is recommended to fit the complete model first and delete useless terms rather than building up from the simpler first-order model; using the latter approach, important quadratic and interaction effects can be missed.   ∎

One issue that arises with fitting a model with an abundance of terms, as in Example 12.26, is the potential to commit many type I errors when performing variable utility t tests on every predictor. Exercise 94 presents a method called the *partial F test* for determining whether a group of predictors can all be deleted while controlling the overall type I error rate.

Models with Categorical Predictors

Thus far we have explicitly considered the inclusion of only quantitative (numerical) predictor variables in a multiple regression model. Using simple numerical coding, categorical variables such as sex, type of college (private or state), or type of wood (pine, oak, or walnut) can also be incorporated into a model. Let's first focus on the case of a dichotomous variable, one with just two possible categories—alive or dead, US or foreign manufacture, and so on. With any such variable, we associate an **indicator** (or **dummy**) **variable** whose possible values 0 and 1 indicate which category is relevant for any particular observation.

Example 12.27   Is it possible to predict graduation rates from freshman test scores? Based on the median SAT score of entering freshmen at a university, can we predict the percentage of those freshmen who will get a degree there within six years? To investigate, let y = six-year graduation rate, x_2 = median freshman SAT score, and x_1 = a variable defined to indicate private or public status:

$$x_1 = \begin{cases} 1 & \text{if the university is private} \\ 0 & \text{if the university is public} \end{cases}$$

The corresponding multiple regression model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

The mean graduation rate depends on whether the university is public or private:

$$\text{mean graduation rate} = \beta_0 + \beta_2 x_2 \qquad \text{when } x_1 = 0 \text{ (public)}$$
$$\text{mean graduation rate} = \beta_0 + \beta_1 + \beta_2 x_2 \quad \text{when } x_1 = 1 \text{ (private)}$$

Thus there are two parallel lines with vertical separation β_1, as shown in Figure 12.32a. The coefficient β_1 is the *difference* in mean graduation rates between private and public universities, after adjusting for median SAT score. If $\beta_1 > 0$ then, on average, for a given SAT, private universities will have a higher graduation rate.
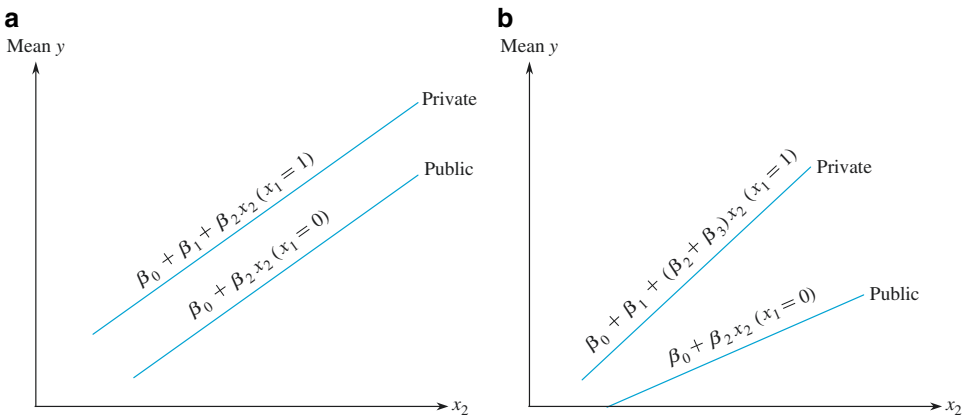

Figure 12.32 Regression functions for models with one indicator variable (x_1) and one quantitative variable (x_2): (**a**) no interaction; (**b**) interaction

A second possibility is a model with an interaction term:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Now the mean graduation rates for the two types of university are

$$\text{mean graduation rate} = \beta_0 + \beta_2 x_2 \qquad\qquad \text{when } x_1 = 0 \text{ (public)}$$
$$\text{mean graduation rate} = \beta_0 + \beta_1 + (\beta_2 + \beta_3) x_2 \quad \text{when } x_1 = 1 \text{ (private)}$$

Here we have two lines, where β_1 is the difference in intercepts and β_3 is the difference in slopes, as shown in Figure 12.32b. Unless $\beta_3 = 0$, the lines will not be parallel and there will be interaction effect, meaning that the separation between public and private graduation rates depends on SAT.

To make inferences, we obtained a random sample of 20 Master's level universities from the 2017 data file available on www.collegeresults.org.

University	Grad rate	Median SAT	Sector
Appalachian State University	73.4	1140	Public
Brenau University	49.4	973	Private
Campbellsville University	34.1	1008	Private
Delta State University	39.6	1028	Public
DeSales University	70.1	1072	Private
Lasell College	54.1	966	Private
Marshall University	49.3	1010	Public
Medaille College	43.9	906	Private
Mount Saint Joseph University	60.7	1011	Private
Mount Saint Mary College	53.8	991	Private
Muskingum University	48.2	1009	Private
Pacific University	64.4	1122	Private
Simpson University	56.7	985	Private
SUNY Oneonta	70.9	1082	Public
Texas A&M University-Texarkana	29.7	1016	Public
Truman State University	74.9	1224	Public
University of Redlands	77.0	1101	Private
University of Southern Indiana	39.6	1005	Public
University of Tennessee-Chattanooga	45.2	1088	Public
Western State Colorado University	41.0	1026	Public

First of all, does the interaction predictor provide useful information over and above what is contained in x_1 and x_2? To answer this question, we should test the hypothesis $H_0: \beta_3 = 0$ versus $H_a: \beta_3 \neq 0$ first. If H_0 is not rejected (meaning interaction is not informative) then we can use the parallel lines model to see if there is a separation (β_1) between lines. Of course, it does not make sense to estimate the difference between lines if the difference depends on x_2, which is the case when there is interaction.

Figure 12.33 shows R output for these two tests. The coefficient for interaction has a P-value of roughly .42, so there is no reason to reject the null hypothesis $H_0: \beta_3 = 0$. Since we fail to reject the "no-interaction" hypothesis, we drop the interaction term and re-run the analysis. The estimated regression equation specified by R is

$$y = -124.56039 + 13.33553x_1 + 0.16474x_2$$

The t ratio values and P-values indicate that both $H_0: \beta_1 = 0$ and $H_0: \beta_2 = 0$ should be rejected at the .05 significance level. The coefficient on x_1 indicates that a private university is estimated to have a graduation rate about 13 percentage points higher than a state university with the same median SAT.

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             -152.37773   49.18039  -3.098 0.006904 **
SectorPrivate             70.06920   69.28401   1.011 0.326908
Median.SAT                 0.19077    0.04592   4.154 0.000746 ***
SectorPrivate:Median.SAT  -0.05457    0.06649  -0.821 0.423857
```

Testing without interaction

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -124.56039   35.29279  -3.529 0.002575 **
SectorPrivate  13.33553    4.64033   2.874 0.010530 *
Median.SAT      0.16474    0.03289   5.009 0.000108 ***
```

Figure 12.33 R output for an interaction model and a "parallel lines" model

At the same time, the coefficient on x_2 shows that, after adjusting for a university's sector (private or public), a 100-point increase in median SAT score is associated with roughly a 16.5 percentage point increase in the school's six-year graduation rate. ∎

You might think that the way to handle a three-category variable is to define a single numerical predictor with coded values such as 0, 1, and 2 corresponding to the three categories. This is incorrect: doing so imposes an ordering on the categories that is not necessarily implied by the context, and it forces the difference in mean response between the 0 and 1 categories to equal the difference for categories 1 and 2 (because $1 - 0 = 2 - 1$ and the model is linear in its predictors). The correct way to incorporate three categories is to define two different indicator variables. Suppose, for example, that y = score on a posttest taken after instruction, x_1 = score on an ability pretest taken before instruction, and that there are three methods of instruction in a mathematics unit: (1) with symbols, (2) without symbols, and (3) a hybrid method. Then let

$$x_2 = \begin{cases} 1 & \text{instruction method 1} \\ 0 & \text{otherwise} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{instruction method 2} \\ 0 & \text{otherwise} \end{cases}$$

For an individual taught with method 1, $x_2 = 1$ and $x_3 = 0$, whereas for an individual taught with method 2, $x_2 = 0$ and $x_3 = 1$. For an individual taught with method 3, $x_2 = x_3 = 0$, and it is not possible that $x_2 = x_3 = 1$ because an individual cannot be taught simultaneously by both methods 1 and 2. The no-interaction model would have only the predictors x_1, x_2, and x_3. The following interaction model allows the change in mean posttest score associated with a one-point increase in pretest to depend on the method of instruction:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon$$

Construction of a picture like Figure 12.32 with a graph for each of the three possible (x_2, x_3) pairs gives three nonparallel lines (unless $\beta_4 = \beta_5 = 0$).

More generally, incorporating a categorical variable with c possible categories into a multiple regression model requires the use of $c - 1$ indicator variables (e.g., five methods of instruction would necessitate using four indicator variables). Thus even one categorical variable can add many predictors to a model.

Indicator variables can be used for categorical variables without any other predictors in the model. For example, consider Example 11.3, which compared the maximum power of five different thermoelectric modules. Using a regression with four indicator variables (to represent the five categories) produces the exact same ANOVA table presented in Example 11.3. In particular, the "treatment sum of squares" SSTr in Chapter 11 and the "regression sum of squares" SSR of this chapter are identical, as are SSE, SST, and hence the results of the F test. In a sense, *analysis of variance is a special case of multiple regression, with the only predictor variables being indicators for various categories.*

Analysis that involves both quantitative and categorical predictors, as in Example 12.27, is sometimes called **analysis of covariance**, and the quantitative variable(s) are called **covariates**. This terminology is typically applied when the effect of the categorical predictor is of primary interest, while the inclusion of the quantitative variables serves to reduce the amount of unexplained variation in the model.

Exercises: Section 12.8 (88–98)

88. The article "Selling Prices/Sq. Ft. of Office Buildings in Downtown Chicago – How Much Is It Worth to Be an Old but Class-A Building?" (*J. Real Estate Res*. 2010: 1–22) considered a regression model to relate $y = \ln(\$/ft^2)$ to 16 predictors, including age, age squared, number of stories, occupancy rate, and indicator variables for whether it is a class-A building, whether a building has a restaurant and whether it has conference rooms. The model was fit to data resulting from 203 sales.

    a. The coefficient of multiple determination was .711. What is the value of the adjusted coefficient of multiple determination? Does it suggest that the relatively high R^2 value was the result of including too many predictors in the model relative to the amount of data available?

    b. Using the R^2 value from (a), carry out a test of hypotheses to see whether there is a useful linear relationship between the response variable and at least one of the predictors.

    c. The estimated coefficient of the indicator variable for whether or not a building was class-A was .364. Interpret this estimated coefficient, first in terms of y and then in terms of $\$/ft^2$.

    d. The t ratio for the estimated coefficient of (c) was 5.49. What does this tell you?

89. Cerium dioxide (also called ceria) is used in many applications, including pollution control and wastewater treatment. The article "Mechanical Properties of Gelcast Cerium Dioxide from 23 to 1500 °C" (*J. Engr. Mater. Technol.* 2017) reports an experiment to determine the relationship between y = elastic modulus (GPa) and x = temperature for ceria specimens under certain conditions. A scatterplot in the article suggests a quadratic relationship.

    a. The article reports the estimated equation $y = -1.92 \times 10^{-5}x^2 - .0191x + 89.0$. Over what temperature range does elastic modulus increase with temperature, and for what temperature range does it decrease?

    b. Predict the elastic modulus of a ceria specimen at 800 °C.

    c. The coefficient of determination is reported as $R^2 = .948$. Use the fact that the data consisted of $n = 28$ observations to perform a model utility test at the .01 level.

    d. Information consistent with the article suggests that at $x = 800$, $s_{\hat{Y}} = 2.9$ GPa. Use this to calculate a 95% CI for $\mu_{Y|800}$.

    e. The residual standard deviation for the quadratic model is roughly $s_e = 2.37$ GPa. Use this to calculate a 95% PI at $x = 800$, and interpret this interval.

90. Many studies have researched how traffic load affects road asphalt, but fewer have examined the effect of extreme cold weather. The article "Effects of Large Freeze-Thaw Cycles on Stiffness and Tensile Strength of Asphalt Concrete" (*J. Cold Regions Engr.* 2016) reports the following data on y = indirect tensile strength (MPa) and x = temperature (°C) for six asphalt specimens in one particular experiment.

x	−35	−20	−10	0	10	22
y	3.01	3.56	3.47	2.72	2.15	1.20

    a. Verify that a scatterplot of the data is consistent with the choice of a quadratic regression model.

    b. Determine the estimated quadratic regression equation.

    c. Calculate a point prediction for the indirect tensile strength of this asphalt type at 0 °C.

d. What proportion of the observed variation in tensile strength can be attributed to the quadratic regression relationship?

e. Obtain a 95% CI for $\mu_{Y|0}$, the true expected tensile strength of this asphalt type at 0 °C.

f. Obtain and interpret a 95% PI at $x = 0$.

91. Ethyl vanillin is used in food, cosmetics, and pharmaceuticals for its vanilla-like scent. The article "Determination and Correlation of Ethyl Vanillin Solubility in Different Binary Solvents at Temperatures from 273.15 to 313.15 K" (*J. Chem. Engr. Data* 2017: 1788–1796) reported an experiment to determine y = ethyl vanillin solubility (mole fraction) as a function of x_1 = initial mole fraction of the chemical propan-2-one in the solvent mixture and x_2 = temperature (°K). The experiment was run at seven x_1 and nine x_2 values. The accompanying table shows the response, y, at each combination. [*Note:* 273.15 °K corresponds to 0 °C.]

$x_1 =$	.4	.5	.6	.7	.8	.9	1.0
$x_2 =$							
273.15	10.6	13.4	16.8	18.5	19.5	19.8	19.9
278.15	12.8	16.2	19.2	21.0	21.4	21.5	21.6
283.15	13.7	18.8	20.9	23.1	23.4	23.5	23.7
288.15	17.3	20.7	23.8	26.4	26.7	26.9	27.1
293.15	20.5	25.2	26.9	29.1	29.6	29.8	30.0
298.15	23.7	27.2	29.8	31.2	32.1	32.4	33.0
303.15	27.4	31.6	33.4	35.0	36.2	36.2	36.9
308.15	34.3	36.9	38.7	40.7	41.0	41.3	41.6
313.15	38.5	42.1	43.4	45.3	45.5	45.7	45.9

a. Create scatterplots of y versus x_1 and y versus x_2. Does it appear the predictors are linearly related to y, or would quadratic terms be appropriate?

b. Would a scatterplot of x_1 versus x_2 indicate whether an interaction term might be suitable? Why or why not?

c. Perform a regression using the complete second-order model. Based on a residual analysis, does it appear that the model assumptions are satisfied?

d. Test various hypotheses to determine which term(s) should be retained in the model.

92. In the construction industry, a "project labor agreement" (PLA) between clients and contractors stipulates that all bidders on a project will use union labor and abide by union rules. The article "Do Project Labor Agreements Raise Construction Costs?" (*CS-BIGS* 2007: 71–79) investigated construction costs for 126 schools in Massachusetts over an eight-year period. Among the variables considered were y = project cost, in dollars per square foot; x_1 = project size, in 1000s of ft^2; $x_2 = 1$ for new construction and 0 for remodel; and $x_3 = 1$ if a PLA was in effect and 0 otherwise.

a. What would it mean in this context to say that x_1 and x_3 have an interaction effect?

b. What would it mean in this context to say that x_2 and x_3 have an interaction effect?

c. No second-order terms are statistically significant here, and the estimated regression equation for the first-order model is $y = 138.69 - .1236x_1 + 17.89x_2 + 18.83x_3$. Interpret the coefficient on x_1. Does the sign make sense?

d. Interpret the coefficient on x_3.

e. The estimated standard error of $\hat{\beta}_3$ is 4.96. Test the hypotheses $H_0: \beta_3 = 0$ vs. $H_a: \beta_3 > 0$ at the .01 significance level. Does the data indicate that PLAs tend to raise construction costs?

93. A regression analysis carried out to relate y = repair time for a water filtration system (hr) to x_1 = elapsed time since the previous service (months) and x_2 = type of repair (1 if electrical and 0 if mechanical) yielded the following model based on $n = 12$ observations: $y = .950 + .400x_1 + 1.250x_2$. In addition, SST = 12.72, SSE = 2.09, and $s_{\hat{\beta}_2} = .312$.

a. Does there appear to be a useful linear relationship between repair time and the two model predictors? Carry out a test of the appropriate hypotheses using a significance level of .05.

b. Given that elapsed time since the last service remains in the model, does type of repair provide useful information about repair time? State and test the appropriate hypotheses using a significance level of .01.

c. Calculate and interpret a 95% CI for β_2.

d. The estimated standard deviation of a prediction for repair time when elapsed time is six months and the repair is electrical is .192. Predict repair time under these circumstances by calculating a 99% prediction interval. Does the interval suggest that the estimated model will give an accurate prediction? Why or why not?

94. Sometimes an investigator wishes to decide whether a group of m predictors ($m > 1$) can simultaneously be eliminated from the model. The null hypothesis says that all β's associated with these m predictors are 0, which is interpreted to mean that as long as the other $k - m$ predictors are retained in the model, the m predictors under consideration collectively provide no useful information about y. The test is carried out by first fitting the "full" model with all k predictors to obtain SSE(full) and then fitting the "reduced" model consisting just of the $k - m$ predictors *not* being considered for deletion to obtain SSE(red). The test statistic is

$$F = \frac{[\text{SSE(red)} - \text{SSE(full)}]/m}{\text{SSE(full)}/[n - (k + 1)]}$$

The test is upper-tailed and based on m numerator df and $n - (k + 1)$ denominator df. This procedure is called the **partial F test**.

Refer back to Example 12.26. The following are the SSEs and numbers of predictors for the first-order, complete second-order, and the researchers' final model.

Model	Predictors	SSE
First-order	5	44,985
Complete second-order	20	16,092
Final	12	17,794

a. Use the partial F test to compare the first-order and complete second-order models. Is there evidence that at least some of the second-order terms should be retained?

b. Use the partial F test to compare the final model to the complete second-order model. Do you agree with the researchers' decision to eliminate the "other" eight predictors?

95. Utilization of sucrose as a carbon source for the production of chemicals is uneconomical. Beet molasses is a readily available and low-priced substitute. The article "Optimization of the Production of β-Carotene from Molasses by *Blakeslea trispora*" (*J. Chem. Tech. Biotech.* 2002: 933–943) carried out a multiple regression analysis to relate the response variable y = amount of β-carotene (g/dm^3) to the three predictors: amount of linoleic acid, amount of kerosene, and amount of antioxidant (all g/dm^3).

Obs.	Linoleic	Kerosene	Antiox.	Betacarotene
1	30.00	30.00	10.00	0.7000
2	30.00	30.00	10.00	0.6300
3	30.00	30.00	18.41	0.0130
4	40.00	40.00	5.00	0.0490
5	30.00	30.00	10.00	0.7000
6	13.18	30.00	10.00	0.1000
7	20.00	40.00	5.00	0.0400
8	20.00	40.00	15.00	0.0065
9	40.00	20.00	5.00	0.2020
10	30.00	30.00	10.00	0.6300
11	30.00	30.00	1.59	0.0400
12	40.00	20.00	15.00	0.1320
13	40.00	40.00	15.00	0.1500
14	30.00	30.00	10.00	0.7000
15	30.00	46.82	10.00	0.3460
16	30.00	30.00	10.00	0.6300
17	30.00	13.18	10.00	0.3970
18	20.00	20.00	5.00	0.2690
19	20.00	20.00	15.00	0.0054
20	46.82	30.00	10.00	0.0640

a. Fitting the complete second-order model in the three predictors resulted in $R^2 = .987$ and adjusted $R^2 = .974$, whereas fitting the first-order model gave $R^2 = .016$. What would you conclude about the two models?

b. For $x_1 = x_2 = 30$, $x_3 = 10$, a statistical software package reported that $\hat{y} = .66573$, $s_{\hat{Y}} = .01785$, and $s_e = .044$ based on the complete second-order model. Predict the amount of β-carotene that would result from a single experimental run with the designated values of the explanatory variables, and do so in a way that conveys information about precision and reliability.

96. Snowpacks contain a wide spectrum of pollutants that may represent environmental hazards. The article "Atmospheric PAH Deposition: Deposition Velocities and Washout Ratios" (*J. Environ. Engr.* 2002: 186–195) focused on the deposition of polyaromatic hydrocarbons. The authors proposed a multiple regression model for relating deposition over a specified time period (y, in $\mu g/m^2$) to two rather complicated predictors x_1 (μg-s/m^3) and x_2 ($\mu g/m^2$) defined in terms of PAH air concentrations for various species, total time, and total amount of precipitation. Here is data on the species fluoranthene and corresponding Minitab output:

x_1	x_2	y
92017	.0026900	278.78
51830	.0030000	124.53
17236	.0000196	22.65
15776	.0000360	28.68
33462	.0004960	32.66
243500	.0038900	604.70
67793	.0011200	27.69
23471	.0006400	14.18
13948	.0004850	20.64
8824	.0003660	20.60
7699	.0002290	16.61
15791	.0014100	15.08
10239	.0004100	18.05
43835	.0000960	99.71
49793	.0000896	58.97
40656	.0026000	172.58
50774	.0009530	44.25

The regression equation is
flth dep = −33.5 + 0.00205 x1 + 29836 x2

Predictor	Coef	SE Coef	T	P
Constant	−33.46	14.90	−2.25	0.041
x1	0.0020548	0.0002945	6.98	0.000
x2	29836	13654	2.19	0.046

S = 44.28   R-Sq = 92.3%   R-Sq(adj) = 91.2%

Analysis of variance

Source	DF	SS	MS	F	P
Regression	2	330,989	165,495	84.39	0.000
Residual error	14	27,454	1961		
Total	16	35,8443			

Formulate questions and perform appropriate analyses. Construct the appropriate residual plots, including plots against the predictors. Based on these plots, justify adding a quadratic predictor, and fit the model with this additional predictor. Does this predictor provide additional useful information over and above what x_1 and x_2 contribute, and does it help the appearance of the diagnostic plots? Also, the data includes a clear outlier. Re-run the regression without the outlier and determine whether quadratic terms are appropriate.

97. The following data set has ratings from ratebeer.com along with values of IBU (international bittering units, a measure of bitterness) and ABV (alcohol by volume) for 25 beers. Notice which beers have the lowest ratings and which are highest.

Beer	IBU	ABV	Rating
Amstel Light	18	3.5	1.93
Anchor Liberty Ale	54	5.9	3.60
Anchor Steam	33	4.9	3.31
Bud Light	7	4.2	1.15
Budweiser	11	5	1.38
Coors	14	5	1.63
DAB Dark	32	5	2.73
Dogfish 60 min IPA	60	6	3.76
Great Divide Titan IPA	65	6.8	3.81
Great Divide Hercules Double IPA	85	9.1	4.05
Guinness Extra Stout	60	5	3.38
Harp Lager	21	4.3	2.85
Heineken	23	5	2.13
Heineken Premium Light	11	3.2	1.62
Michelob Ultra	4	4.2	1.01
Newcastle Brown Ale	18	4.7	3.05
Pilsner Urquell	35	4.4	3.28
Redhook ESB	29	5.77	3.06
Rogue Imperial Stout	88	11.6	3.98
Samuel Adams Boston Lager	31	4.9	3.19
Shiner Light	13	4.03	2.57
Sierra Nevada Pale Ale	37	5.6	3.61
Sierra Nevada Porter	40	5.6	3.60
Terrapin All-Amer. Imperial Pilsner	75	7.5	3.46
Three Floyds Alpha King	66	6	4.04

a. Find the correlations (and the corresponding *P*-values) among Rating, IBU, and ABV.

b. Regress rating on IBU and ABV. Notice that although both predictors have strongly significant correlations with Rating, they do not both have significant regression coefficients. How do you explain this?

c. Plot the residuals from the regression of (b) to check the assumptions. Also plot rating against each of the two predictors. Which of the assumptions is clearly not satisfied?

d. Regress rating on IBU and ABV with the square of IBU as a third predictor. Again check assumptions.

e. How effective is the regression in (d)? Interpret the coefficients with regard to statistical significance and sign. In particular, discuss the relationship to IBU.

f. Summarize your conclusions.

98. The article "Promoting Healthy Choices: Information versus Convenience" (*Amer. Econ. J.: Appl. Econ.* 2010: 164–178) reported on a field experiment at a fast-food sandwich chain to see whether calorie information provided to patrons would affect calorie intake. One aspect of the study involved fitting a multiple regression model with seven predictors to data consisting of 342 observations. Predictors in the model included age and indicator variables for sex, whether or not a daily calorie recommendation was provided, and whether or not calorie information about choices was provided. The reported value of the *F* ratio for testing model utility was 3.64.

a. At significance level .01, does the model appear to specify a useful linear relationship between calorie intake and at least one of the predictors?

b. What can be said about the *P*-value for the model utility *F* test?

c. What proportion of the observed variation in calorie intake can be attributed to the model relationship? Does this seem very impressive? Why is the *P*-value as small as it is?

d. The estimated coefficient for the indicator variable *calorie information provided* was -71.73, with an estimated standard error of 25.29. Interpret the coefficient. After adjusting for the effects of other predictors, does it appear that true average calorie intake depends on whether or not calorie information is provided? Carry out a test of appropriate hypotheses.

12.9 Regression with Matrices

Throughout this chapter we have explored linear models with both one and several predictors. It should perhaps not be surprising that such models can be imbedded in the language of linear algebra, i.e., in matrix form. In this section, we re-write the model equation and least squares estimates in terms of certain matrices and then derive matrix-based formulas for several of the quantities mentioned in earlier sections. (The focus here will be on multiple regression, since simple linear regression is just the special case where $k = 1$.) In fact, all software packages that perform regression analysis rely on these matrix representations for computation.

The Model Equation in Matrix Form

In Section 12.7 we used the following additive model equation to relate a response variable *y* to explanatory variables $x_1, \ldots, x_k$:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma)$ and the ε_i's for different observations are independent of one another. Suppose there are n observations, each consisting of a y value and values of the k predictors (so each observation consists of $k + 1$ numbers). Then the n equations for the various observations can be expressed compactly using matrix notation:

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
&\ \vdots \\
Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n
\end{aligned}
\quad \Rightarrow \quad
\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}
=
\begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}
\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}
+
\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}
$$

$$(12.15)$$

The dimensions of the four matrices in (12.15), from left to right, are $n \times 1$, $n \times (k + 1)$, $(k + 1) \times 1$, and $n \times 1$. If we denote these four matrices by $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$, then the multiple linear regression model is equivalent to

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

We will use $\mathbf{y}$ to denote the $n \times 1$ column vector of observed y values: $\mathbf{y} = [y_1, \ldots, y_n]'$, where $'$ denotes matrix transpose. The vector $\mathbf{y}$ (or $\mathbf{Y}$) is called the **response vector**, while the matrix $\mathbf{X}$ is known as the **design matrix**. The design matrix consists of one row for each observation (n rows total) and one column for each predictor, along with a leading column of 1's to accommodate the constant term.

Parameter Estimation in Matrix Form

We now estimate β_0, β_1, …, β_k using the principle of least squares. Let $\|\mathbf{u}\|$ denote the (Euclidean) length of a column vector $\mathbf{u}$, i.e., $\|\mathbf{u}\|^2 = \sum u_i^2 = \mathbf{u}'\mathbf{u}$. Then our goal is to find $b_0, b_1, \ldots, b_k$ to minimize

$$g(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik})]^2 = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2,$$

where $\mathbf{b}$ is the column vector with entries b_0, b_1, …, b_k. One solution method was outlined in Section 12.7: if we set the partial derivatives of g with respect to $b_0, b_1, \ldots, b_k$ equal to zero, the result is the normal equations in Expression (12.13). In matrix form, (12.13) becomes

$$
\begin{bmatrix}
\sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik} \\
\sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1} x_{i1} & \cdots & \sum_{i=1}^{n} x_{i1} x_{ik} \\
& \vdots & & \\
\sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik} x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik} x_{ik}
\end{bmatrix}
\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}
=
\begin{bmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} x_{i1} y_i \\
\vdots \\
\sum_{i=1}^{n} x_{ik} y_i
\end{bmatrix}
$$

The matrix on the left is $\mathbf{X}'\mathbf{X}$ and the one on the far right is $\mathbf{X}'\mathbf{y}$. The normal equations then become $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$. We will assume throughout this section that $\mathbf{X}'\mathbf{X}$ has an inverse, so the vector of estimated coefficients is given by

$$\hat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{12.16}$$

All statistical software packages use the matrix version of the normal equations to calculate the estimated regression coefficients. (At the end of this section, we present an alternative derivation of $\hat{\boldsymbol{\beta}}$ that relies on linear algebra rather than partial derivatives from calculus.)

Example 12.28 For illustrative purposes, suppose we wish to predict horsepower using engine size (liters) and fuel type (premium or regular) based on a sample of $n = 6$ cars.

Horsepower	Engine size	Fuel type
132	2.0	Regular
167	2.0	Premium
170	2.5	Regular
204	2.5	Premium
230	3.0	Regular
260	3.0	Premium

Define variables y = horsepower, x_1 = engine size, and $x_2 = 1$ for premium fuel and 0 for regular (an indicator variable). Then the response vector and design matrix here are

$$\mathbf{y} = \begin{bmatrix} 132 \\ 167 \\ 170 \\ 204 \\ 230 \\ 260 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 2.0 & 0 \\ 1 & 2.0 & 1 \\ 1 & 2.5 & 0 \\ 1 & 2.5 & 1 \\ 1 & 3.0 & 0 \\ 1 & 3.0 & 1 \end{bmatrix} \quad \Rightarrow \quad \mathbf{X}'\mathbf{X} = \begin{bmatrix} 6 & 15 & 3 \\ 15 & 38.5 & 7.5 \\ 3 & 7.5 & 3 \end{bmatrix} \quad \text{and} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 1163 \\ 3003 \\ 631 \end{bmatrix}$$

Notice that $\mathbf{X}'\mathbf{X}$ is symmetric (and will be in all cases—do you see why?). The least squares estimates of the regression coefficients are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 79/12 & -5/2 & -1/3 \\ -5/2 & 1 & 0 \\ -1/3 & 0 & 2/3 \end{bmatrix} \begin{bmatrix} 1163 \\ 3003 \\ 631 \end{bmatrix} = \begin{bmatrix} -61.417 \\ 95.5 \\ 33 \end{bmatrix}$$

Figure 12.34 shows R output from multiple regression using this (toy) data set. Notice that estimated regression coefficients exactly match our vector $\hat{\boldsymbol{\beta}}$.

```
Residuals:
    1      2      3      4      5      6
 2.417  4.417 -7.333 -6.333  4.917  1.917

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -61.417     17.966  -3.419 0.041887 *
x1            95.500      7.002  13.639 0.000853 ***
x2            33.000      5.717   5.772 0.010337 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.002 on 3 degrees of freedom
Multiple R-squared:  0.9865,    Adjusted R-squared:  0.9775
F-statistic: 109.7 on 2 and 3 DF,  p-value: 0.001567
```

Figure 12.34 R output for Example 12.28   ∎

Residuals, ANOVA, *F*, and R^2

The estimated regression coefficients can be used to obtain the predicted values and the residuals. Recall that the *i*th predicted value is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$. The vector of predicted values, $\hat{\mathbf{y}}$, is

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \cdots + \hat{\beta}_k x_{1k} \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 x_{n1} + \cdots + \hat{\beta}_k x_{nk} \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

If we define a matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, this relationship can be re-written as $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. The matrix $\mathbf{H}$ is amusingly called the **hat matrix** because it "puts a hat" on the vector $\mathbf{y}$.

The residual for the *i*th observation is defined by $e_i = y_i - \hat{y}_i$, and so the **residual vector** is

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y},$$

where $\mathbf{I}$ denotes the $n \times n$ identity matrix. Now the sums of squares encountered throughout this chapter can also be written in matrix form (more precisely, as squared lengths of particular vectors). Let $\bar{\mathbf{y}}$ denote an $n \times 1$ column vector whose every entry is the sample mean y value, $\bar{y}$. Then

$$\mathrm{SSE} = \sum e_i^2 = \|\mathbf{e}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2$$
$$\mathrm{SSR} = \sum (\hat{y}_i - \bar{y})^2 = \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2$$
$$\mathrm{SST} = \sum (y_i - \bar{y})^2 = \|\mathbf{y} - \bar{\mathbf{y}}\|^2$$

from which, as before, $s_e^2 = \mathrm{MSE} = \mathrm{SSE}/[n - (k + 1)]$ and $R^2 = \mathrm{SSR}/\mathrm{SST}$. The fundamental ANOVA identity $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$ can be obtained as follows:

$$\mathrm{SST} = \|\mathbf{y} - \bar{\mathbf{y}}\|^2 = (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}) = [(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \bar{\mathbf{y}})]'[(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \bar{\mathbf{y}})]$$
$$= \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 = \mathrm{SSE} + \mathrm{SSR}$$

The cross-terms in the matrix product are zero because of the normal equations (see Exercise 104). Equivalently, the middle two terms drop out because the vectors $\hat{\mathbf{y}} - \bar{\mathbf{y}}$ and $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ are *orthogonal*.

The model utility test of $H_0: \beta_1 = \cdots = \beta_k = 0$ uses the same F ratio as before:

$$f = \frac{\mathrm{MSR}}{\mathrm{MSE}} = \frac{\mathrm{SSR}/k}{\mathrm{SSE}/[n - (k + 1)]} = \frac{\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2/k}{\|\mathbf{e}\|^2/[n - (k + 1)]}$$

Example 12.29 (Example 12.28 continued) The predicted values and residuals are easily obtained:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 2.0 & 0 \\ 1 & 2.0 & 1 \\ 1 & 2.5 & 0 \\ 1 & 2.5 & 1 \\ 1 & 3.0 & 0 \\ 1 & 3.0 & 1 \end{bmatrix} \begin{bmatrix} -61.417 \\ 95.50 \\ 33 \end{bmatrix} = \begin{bmatrix} 129.583 \\ 162.583 \\ 177.333 \\ 210.333 \\ 225.083 \\ 258.083 \end{bmatrix} \qquad \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} 132 \\ 167 \\ 170 \\ 204 \\ 230 \\ 260 \end{bmatrix} - \begin{bmatrix} 129.583 \\ 162.583 \\ 177.333 \\ 210.333 \\ 225.083 \\ 258.083 \end{bmatrix} = \begin{bmatrix} 2.417 \\ 4.417 \\ -7.333 \\ -6.333 \\ 4.917 \\ 1.917 \end{bmatrix}$$

From these, $\text{SSE} = \|\mathbf{e}\|^2 = 2.417^2 + \cdots + 1.917^2 = 147.083$, $\text{MSE} = \text{SSE}/[n - (k + 1)] = 147.083/[6 - (2 + 1)] = 49.028$, and $s_e = \sqrt{49.028} = 7.002$. The total sum of squares is $\text{SST} = \|\mathbf{y} - \bar{\mathbf{y}}\|^2 = \sum (y_i - 193.83)^2 = 10{,}900.83$, and the regression sum of squares can be obtained most easily by subtraction: $\text{SSR} = \text{SST} - \text{SSE} = 10{,}900.83 - 147.083 = 10{,}753.75$. The coefficient of multiple determination is $R^2 = \text{SSR}/\text{SST} = .9865$. Finally, the model utility test statistic value is $f = \text{MSR}/\text{MSE} = (10{,}753.75/2)/49.028 = 109.67$, a massive F ratio that decisively rejects H_0 at any reasonable significance level. Notice that many of these quantities appear in the lower half of the R output in Figure 12.34.   ∎

Inference About Individual Parameters

In order to develop hypothesis tests and confidence intervals for the regression coefficients, the expected values and standard deviations of the estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are needed.

DEFINITION   Let $U_1, \ldots, U_m$ be rvs and $\mathbf{U}$ denote the $m \times 1$ column vector $[U_1, \ldots, U_m]'$. Then the **mean vector** of $\mathbf{U}$ is the $m \times 1$ column vector $\boldsymbol{\mu} = E(\mathbf{U})$ whose ith entry is $\mu_i = E(U_i)$. The **covariance matrix** of $\mathbf{U}$ is the $m \times m$ matrix whose (i, j)th entry is the covariance of U_i and U_j. That is,

$$\text{Cov}(\mathbf{U}) = \begin{bmatrix} \text{Cov}(U_1, U_1) & \text{Cov}(U_1, U_2) & \cdots & \text{Cov}(U_1, U_m) \\ \text{Cov}(U_2, U_1) & \text{Cov}(U_2, U_2) & \cdots & \text{Cov}(U_2, U_m) \\ \vdots & & \ddots & \\ \text{Cov}(U_m, U_1) & \text{Cov}(U_m, U_2) & \cdots & \text{Cov}(U_m, U_m) \end{bmatrix}$$

If we define the expected value of a matrix of rvs by the element-wise expectations of its entries, then it follows from the definition $\text{Cov}(U_i, U_j) = E[(U_i - \mu_i)(U_j - \mu_j)]$ that

$$\text{Cov}(\mathbf{U}) = E[(\mathbf{U} - \boldsymbol{\mu})(\mathbf{U} - \boldsymbol{\mu})'] \tag{12.17}$$

The diagonal entries of the covariance matrix are the variances of the rvs: $\text{Cov}(U_i, U_i) = V(U_i)$. Also, the covariance matrix is symmetric, since $\text{Cov}(U_i, U_j) = \text{Cov}(U_j, U_i)$.

For example, suppose U_1 and U_2 are rvs with means 10 and -4, standard deviations 2.5 and 2.3, and covariance -1.1. Then the mean vector and covariance matrix of $\mathbf{U} = [U_1, U_2]'$ are

$$E(\mathbf{U}) = \begin{bmatrix} E(U_1) \\ E(U_2) \end{bmatrix} = \begin{bmatrix} 10 \\ -4 \end{bmatrix} \quad \text{and} \quad \text{Cov}(\mathbf{U}) = \begin{bmatrix} V(U_1) & \text{Cov}(U_1, U_2) \\ \text{Cov}(U_2, U_1) & V(U_2) \end{bmatrix} = \begin{bmatrix} 2.5^2 & -1.1 \\ -1.1 & 2.3^2 \end{bmatrix}$$

Now consider the vector of random errors $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]'$. The linear regression model assumes that the ε_i's are independent (so covariance = 0 for each pair) with mean 0 and common variance σ^2. Under these assumptions, the mean vector of $\boldsymbol{\varepsilon}$ is $\mathbf{0}$ (an $n \times 1$ vector of 0's), while the covariance matrix of $\boldsymbol{\varepsilon}$ is $\sigma^2\mathbf{I}$ (an $n \times n$ matrix with σ^2 along the main diagonal and 0's everywhere else). It then follows from the model equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ that the (random) response vector $\mathbf{Y}$ satisfies

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{0} = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \text{Cov}(\mathbf{Y}) = \text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

To determine the sampling distribution of $\hat{\boldsymbol{\beta}}$, we will require the following proposition.

PROPOSITION   Let $\mathbf{U}$ be a random vector. If $\mathbf{A}$ is a matrix with constant entries and $\mathbf{V} = \mathbf{AU}$, then $E(\mathbf{V}) = \mathbf{A}E(\mathbf{U})$ and $\text{Cov}(\mathbf{V}) = \mathbf{A}\text{Cov}(\mathbf{U})\mathbf{A}'$.

Proof By the linearity of the expectation operator, $E(\mathbf{V}) = E(\mathbf{AU}) = \mathbf{A}E(\mathbf{U}) = \mathbf{A}\boldsymbol{\mu}$. Then, using (12.17),

$$
\begin{aligned}
\text{Cov}(\mathbf{V}) &= E[(\mathbf{V} - E(\mathbf{V}))(\mathbf{V} - E(\mathbf{V}))'] \quad \text{Equation (12.17)} \\
&= E[(\mathbf{AU} - \mathbf{A}\boldsymbol{\mu})(\mathbf{AU} - \mathbf{A}\boldsymbol{\mu})'] = E[\mathbf{A}(\mathbf{U} - \boldsymbol{\mu})(\mathbf{A}(\mathbf{U} - \boldsymbol{\mu}))'] \\
&= E[\mathbf{A}(\mathbf{U} - \boldsymbol{\mu})(\mathbf{U} - \boldsymbol{\mu})'\mathbf{A}'] \\
&= \mathbf{A}E[(\mathbf{U} - \boldsymbol{\mu})(\mathbf{U} - \boldsymbol{\mu})']\mathbf{A}' \qquad \text{linearity of expectation} \\
&= \mathbf{A}\text{Cov}(\mathbf{U})\mathbf{A}'
\end{aligned}
$$

∎

Let's apply this proposition to find the mean vector and covariance matrix of $\hat{\boldsymbol{\beta}}$. As an estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, so let $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{U} = \mathbf{Y}$. By linearity of expectation (the first part of the proposition),

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

That is, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ (for each j, $\hat{\beta}_j$ is unbiased for estimating β_j).

Next, the transpose of $\mathbf{A}$ is $\mathbf{A}' = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$; this relies on the fact that $\mathbf{X}'\mathbf{X}$ is symmetric, so $(\mathbf{X}'\mathbf{X})^{-1}$ is symmetric as well. Applying the second part of the proposition and the earlier observation that $\text{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$,

$$
\begin{aligned}
\text{Cov}(\hat{\boldsymbol{\beta}}) &= \mathbf{A}\text{Cov}(\mathbf{Y})\mathbf{A}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\sigma^2\mathbf{I}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

So, the variance of the regression coefficient $\hat{\beta}_j$ is the jth diagonal entry of the matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Exercise 101 asks you to demonstrate that this matrix formula matches the variance formulas presented earlier in the chapter for simple linear regression. Since σ is unknown, it must be estimated from the data, and the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is $s_e^2(\mathbf{X}'\mathbf{X})^{-1}$.

Example 12.30 (Example 12.29 continued) For the engine horsepower scenario we previously found the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ and the residual standard deviation $s_e = 7.002$. The estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$
7.002^2 \begin{bmatrix} 79/12 & -5/2 & -1/3 \\ -5/2 & 1 & 0 \\ -1/3 & 0 & 2/3 \end{bmatrix} = \begin{bmatrix} 322.766 & -122.569 & -16.343 \\ -122.569 & 49.028 & 0.0 \\ -16.343 & 0.0 & 32.685 \end{bmatrix}
$$

The estimated standard deviations of the three coefficients are $s_{\hat{\beta}_0} = \sqrt{322.766} = 17.966$, $s_{\hat{\beta}_1} = \sqrt{49.028} = 7.002$, and $s_{\hat{\beta}_2} = \sqrt{32.685} = 5.717$. Notice that these exactly match the standard errors given in the R output of Figure 12.34. These estimated standard errors form the basis for the variable utility tests and CIs for the β_j's presented in Section 12.7.

The covariance matrix also indicates that $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \approx -122.569$, meaning the estimates for the y-intercept and the slope on engine size are negatively correlated. In other words, if for a particular sample the estimated slope $\hat{\beta}_1$ is *greater* than its expectation (the true coefficient β_1), then typically the value of $\hat{\beta}_0$ will be *less* than β_0. This makes sense for (x, y) values in the first quadrant: rotating from the true line $y = \beta_0 + \beta_1 x$, if sample data results in a slope estimate that is too high (so the line is overly steep), the y-intercept estimate will naturally be too low.   ∎

What about the standard error of $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_k x_k^*$, our point estimate of the mean response $\mu_{Y|x^*}$ at a specified set of x values? The point estimate may be written as $\hat{Y} = \mathbf{x}^* \hat{\boldsymbol{\beta}}$, where $\mathbf{x}^*$ is the *row* vector $\mathbf{x}^* = [1, x_1^*, \ldots, x_k^*]$. Here, $\mathbf{x}^*$ is constant but $\hat{\boldsymbol{\beta}}$ (a vector of estimators) has sampling variability. Applying the earlier proposition,

$$V(\hat{Y}) = V(\mathbf{x}^* \hat{\boldsymbol{\beta}}) = \mathbf{x}^* V(\hat{\boldsymbol{\beta}})[\mathbf{x}^*]' = \mathbf{x}^* \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}[\mathbf{x}^*]' = \sigma^2 \mathbf{x}^* (\mathbf{X}'\mathbf{X})^{-1}[\mathbf{x}^*]' \Rightarrow$$
$$s_{\hat{Y}}^2 = s_e^2 \cdot \mathbf{x}^* (\mathbf{X}'\mathbf{X})^{-1}[\mathbf{x}^*]'$$

(It's easy to verify here that the expression for $s_{\hat{Y}}^2$ is a 1×1 matrix and, as such, may be treated as a scalar.) The square root of this expression gives the estimated standard error of $\hat{Y}$, which is required for confidence and prediction intervals.

The Hat Matrix, Leverage, and Outlier Detection

The foregoing proposition can also be used to find estimated standard deviations for the residuals. Recall that the $n \times n$ hat matrix is defined by $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. With the help of the matrix rules $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ and $(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}$, we find that $\mathbf{H}$ is symmetric, i.e., $\mathbf{H}' = \mathbf{H}$:

$$\mathbf{H}' = \left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \right]' = (\mathbf{X}')'[(\mathbf{X}'\mathbf{X})^{-1}]'\mathbf{X}' = \mathbf{X}[(\mathbf{X}'\mathbf{X})']^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}$$

Next, recall that the vector of predicted values is given by $\hat{\mathbf{Y}} = \mathbf{HY}$; here, we're treating the response vector $\mathbf{Y}$ as random, which implies that $\hat{\mathbf{Y}}$ is also a random vector. Thus

$$\begin{aligned}\text{Cov}(\hat{\mathbf{Y}}) &= \mathbf{H}\,\text{Cov}(\mathbf{Y})\mathbf{H}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\sigma^2\mathbf{I}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \sigma^2 \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \sigma^2 \mathbf{H}\end{aligned} \tag{12.18}$$

A similar calculation shows that the covariance matrix of the residuals is

$$\text{Cov}(\mathbf{Y} - \hat{\mathbf{Y}}) = \sigma^2(\mathbf{I} - \mathbf{H}) \tag{12.19}$$

The variances of $\hat{Y}_i$ and e_i are the diagonal entries of the matrices in (12.18) and (12.19), respectively. Of course, the value of σ^2 is generally unknown, so the estimate $s_e^2 = \text{MSE}$ is used instead. If we let h_{ii} denote the ith diagonal entry of $\mathbf{H}$, then (12.18) and (12.19) imply that the (estimated) standard deviations of $\hat{Y}_i$ and e_i are

$$s_{\hat{Y}_i} = s_e \cdot \sqrt{h_{ii}} \quad \text{and} \quad s_{e_i} = s_e \cdot \sqrt{1 - h_{ii}}$$

For the case of simple linear regression, it can be shown that these expressions match the standard error formulas given previously (Exercise 110).

The hat matrix is also important as a measure of the influence of individual observations. Because $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, $\hat{y}_i = h_{i1}y_1 + \cdots + h_{ii}y_i + \cdots + h_{in}y_n$, and therefore the ith diagonal element of $\mathbf{H}$ measures the impact of the ith observation y_i on its own predicted value $\hat{y}_i$. The h_{ii}'s are sometimes called the **leverages** to indicate their impact on the regression. An observation with very high leverage will tend to pull the regression toward it, and its residual will tend to be small. Notice, though, that $\mathbf{H}$ depends only on the values of the predictors (through the design matrix $\mathbf{X}$), so leverage measures only one aspect of influence.

Example 12.31 Students in a statistics class measured their height, foot length, and wingspan (measured fingertip to fingertip with hands outstretched) in inches. The accompanying table shows the measurements for 16 students; we encountered this data previously in Example 12.16. The last column has the leverages for the regression of wingspan on height and foot length.

Student	Height (x_1)	Foot (x_2)	Wingspan (y)	Leverage
1	63.0	9.0	62.0	0.239860
2	63.0	9.0	62.0	0.239860
3	65.0	9.0	64.0	0.228236
4	64.0	9.5	64.5	0.223625
5	68.0	9.5	67.0	0.196418
6	69.0	10.0	69.0	0.083676
7	71.0	10.0	70.0	0.262182
8	68.0	10.0	72.0	0.067207
9	68.0	10.5	70.0	0.187088
10	72.0	10.5	72.0	0.151959
11	73.0	11.0	73.0	0.143279
12	73.5	11.0	75.0	0.168719
13	70.0	11.0	71.0	0.245380
14	70.0	11.0	70.0	0.245380
15	72.0	11.0	76.0	0.128790
16	74.0	11.2	76.5	0.188340

Figure 12.35 shows a plot of $x_1 = $ height against $x_2 = $ foot length, along with the leverage for each point. Notice that the points at the extreme right and left of the plot have high leverage, and the points
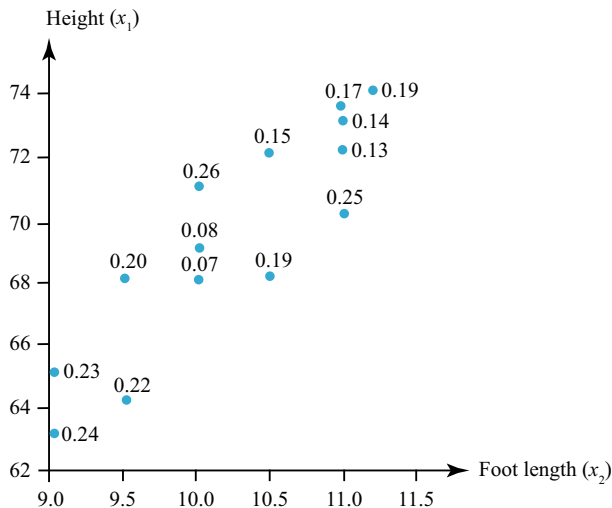

Figure 12.35 Plot of height and foot length showing leverage

near the center have low leverage. However, it is interesting that the point with highest leverage is not at the extremes of height or foot length. This is student number 7, with a 10-in. foot and height of 71 in., and the high leverage comes from the height being extreme relative to foot length. Indeed, when there are several predictors, high leverage often occurs when values of one predictor are extreme relative to the values of other predictors. For example, if height and weight are predictors, then an overweight or underweight subject would likely have high leverage.   ■

Together, standardized residuals and leverages can be used to identify unusual observations in a regression setting. This is particularly helpful in multiple regression, where outliers are more difficult to detect with the naked eye (e.g., student 7 in Example 12.31). By convention, the ith observation is said to have a *large residual* if $|e_i^*| > 2$ and a *very large residual* if $|e_i^*| > 3$, since these indicate that y_i is more than two (resp., three) standard deviations away from the value predicted by the estimated regression function.

As for the leverage values, it can be shown that

$$0 \le h_{ii} \le 1 \quad \text{and} \quad \sum_{i=1}^{n} h_{ii} = k+1$$

where k is the number of predictor variables in the regression model. (In fact, it isn't difficult to show directly that $\sum h_{ii} = 2$ for simple regression, i.e., $k = 1$.) This implies that the mean of the leverage values is $(k + 1)/n$, since there are n leverage values total (one for each observation). By convention, the ith observation is said to possess *high leverage* if $h_{ii} > 2(k + 1)/n$ and *very high leverage* if $h_{ii} > 3(k + 1)/n$.

More sophisticated tools for outlier detection in regression are also available. If the "influence" of an observation is defined in terms of the effect on the predicted values when the observation is omitted, then an *influential observation* is one that has both large leverage and a large residual. A popular measure that combines leverage and residual is *Cook's distance*; consult the book by Kutner et al. for more information. Many statistical software packages will provide the standardized residual, leverage, and Cook's distance for all n observations upon request; some will also flag observations with unusually high values (e.g., according to the criteria above).

Another Perspective on Least Squares
We previously used multivariate calculus—in particular, the normal equations (12.13)—to determine the least squares estimates of the regression coefficients. The matrix representation of regression allows an alternative derivation of these estimates that relies instead on linear algebra.

Let $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k$ denote the $k + 1$ columns of the design matrix $\mathbf{X}$. The principle of least squares says we should determine coefficients $b_0, b_1, \ldots, b_k$ that minimize

$$\sum (y_i - [b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}])^2 = \|\mathbf{y} - [b_0 \mathbf{1} + b_1 \mathbf{x}_1 + \cdots + b_k \mathbf{x}_k]\|^2$$

The expression in brackets is a (generic) linear combination of the vectors $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k$. Since $\| \cdot \|^2$ denotes Euclidean distance, we know from linear algebra that such a distance is minimized by finding the *projection* of $\mathbf{y}$ onto the vector space (i.e., the closest vector to $\mathbf{y}$ in the space) spanned by $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k$. Call this projection vector $\mathbf{p}$. Since $\mathbf{p}$ lies in span$\{\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k\}$ it must have the form $\mathbf{p} = b_0 \mathbf{1} + b_1 \mathbf{x}_1 + \cdots + b_k \mathbf{x}_k = \mathbf{Xb}$; our goal now is to find an explicit formula for the coefficients.

Use the property that if $\mathbf{p}$ is the projection of $\mathbf{y}$ onto span$\{\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k\}$, then the vector $\mathbf{y} - \mathbf{p}$ must be orthogonal to that space. That is, the vector $\mathbf{y} - \mathbf{p}$ must be perpendicular to each of $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k$, meaning that $\mathbf{1}'\mathbf{p} = 0$ and $\mathbf{x}_j'\mathbf{p} = 0$ for $j = 1, \ldots, k$. In matrix form, these $k + 1$ requirements can be written as $\mathbf{X}'(\mathbf{y} - \mathbf{p}) = \mathbf{0}$.

Put it all together: with $\mathbf{p} = \mathbf{Xb}$ and $\mathbf{X}'(\mathbf{y} - \mathbf{p}) = \mathbf{0}$,

$$\mathbf{X}'(\mathbf{y} - \mathbf{p}) = \mathbf{0} \Rightarrow \mathbf{X}'(\mathbf{y} - \mathbf{Xb}) = \mathbf{0} \Rightarrow \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{Xb} \Rightarrow \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

matching the previous formula (12.16). Incidentally, the projection vector itself is $\mathbf{p} = \mathbf{Xb} = \mathbf{Hy} = \hat{\mathbf{y}}$ (the vector of fitted values), and the vector orthogonal to the space is $\mathbf{y} - \mathbf{p} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{e}$ (the vector of residuals).

Exercises: Section 12.9 (99–110)

99. Consider fitting the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ to the following data:

x_1	x_2	y
-1	-1	1
-1	1	1
1	-1	0
1	1	4

    a. Determine $\mathbf{X}$ and $\mathbf{y}$, and express the normal equations in terms of matrices.

    b. Determine the $\hat{\boldsymbol{\beta}}$ vector, which contains the estimates for the three coefficients in the model.

    c. Determine $\hat{\mathbf{y}}$ and $\mathbf{e}$. Then calculate SSE, and use this to get the estimated variance MSE.

    d. Use MSE and $(\mathbf{X}'\mathbf{X})^{-1}$ to construct a 95% confidence interval for β_1.

    e. Carry out a t test for the hypothesis $H_0\colon \beta_1 = 0$ against a two-tailed alternative, and interpret the result.

    f. Form the analysis of variance table, and carry out the F test for the hypothesis $H_0\colon \beta_1 = \beta_2 = 0$. Find R^2 and interpret.

100. Consider the model $Y = \beta_0 + \beta_1 x_1 + \varepsilon$ for the following data:

x_1	y	x_1	y
$-.5$	1	.5	8
$-.5$	2	.5	9
$-.5$	2	.5	7
$-.5$	3	.5	8

    a. Determine the $\mathbf{X}$ and $\mathbf{y}$ matrices and express the normal equations in terms of matrices.

    b. Determine the $\hat{\boldsymbol{\beta}}$ vector, which contains the estimates for the two coefficients in the model.

    c. Determine $\hat{\mathbf{y}}$ and $\mathbf{e}$.

    d. Calculate SSE (by summing the squared residuals) and then the estimated variance MSE.

    e. Use MSE and $(\mathbf{X}'\mathbf{X})^{-1}$ to construct a 95% confidence interval for β_1.

    f. Carry out a t test of $H_0\colon \beta_1 = 0$ against a two-sided alternative.

    g. Carry out the F test of $H_0\colon \beta_1 = 0$. How is this related to part (f)?

101. Consider the simple linear regression model $Y = \beta_0 + \beta_1 x + \varepsilon$, so $k = 1$ and $\mathbf{X}$ consists of a column of 1's and a column of the values $x_1, \ldots, x_n$ of x.

    a. Determine $\mathbf{X}'\mathbf{X}$ and $(\mathbf{X}'\mathbf{X})^{-1}$ using the matrix inverse formula

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

    b. Determine $\mathbf{X}'\mathbf{y}$, then calculate the coefficient vector $\hat{\boldsymbol{\beta}}$. Compare your answers to the formulas given in Section 12.2. [Hint: $S_{xy} = \sum x_i y_i - n \cdot \bar{x} \cdot \bar{y}$, and similarly for S_{xx}.]

c.   Use $(\mathbf{X}'\mathbf{X})^{-1}$ to obtain expressions for the variances of the coefficients, and check your answers against the results given in Sections 12.3 and 12.4. [*Note:* $\hat{\beta}_0$ is the predicted value corresponding to $x^* = 0$, so the variance of $\hat{\beta}_0$ appears implicitly in Section 12.4.]

102. Suppose we have bivariate data (x_1, y_1), ..., (x_n, y_n). Consider the *centered model* $y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i$ for $i = 1, ..., n$.

   a.   Show that

   $$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & 0 \\ 0 & S_{xx} \end{bmatrix}$$

   b.   Determine $(\mathbf{X}'\mathbf{X})^{-1}$ and the coefficient vector $\hat{\boldsymbol{\beta}}$.

   c.   Determine the estimated standard errors of the regression coefficients.

   d.   Compare this exercise to the previous one. Why is it more efficient to have $x_{i1} = x_i - \bar{x}$ rather than $x_{i1} = x_i$ in the design matrix?

103. Consider the model $Y_i = \beta_0 + \varepsilon_i$ (so $k = 0$). Estimate β_0 from Equation (12.16). Find a simple expression for $s_{\hat{\beta}_0}$ and then the 95% confidence interval for β_0. [*Note:* Your result should be equivalent to the one-sample t confidence interval in Section 8.3.]

104. a.   Show that the normal equations are equivalent to $\mathbf{X}'\mathbf{e} = \mathbf{0}$. [*Hint:* Use the matrix representation of the normal equations in this section and substitute the formula for $\mathbf{b} = \hat{\boldsymbol{\beta}}$.]

   b.   Use part (a) to prove the ANOVA identity SST = SSE + SSR by showing that $(\hat{\mathbf{y}} - \bar{\mathbf{y}})'\mathbf{e} = 0$. [*Hint:* Part (a) also implies that each row of $\mathbf{X}'$ is orthogonal to $\mathbf{e}$; in particular, the first column of $\mathbf{X}$, a column of n 1's, satisfies $\mathbf{1}'\mathbf{e} = 0$.]

105. Suppose that we have $Y_1, ..., Y_m \sim N(\mu_1, \sigma)$, $Y_{m+1}, ..., Y_{m+n} \sim N(\mu_2, \sigma)$, and all $m + n$ observations are independent. These are the assumptions of the pooled

t procedure in Section 10.2. Let $k = 1$, $x_{11} = .5$, ..., $x_{m1} = .5$, $x_{m+1,1} = -.5$, ..., $x_{m+n,1} = -.5$. For convenience in inverting $\mathbf{X}'\mathbf{X}$ assume $m = n$.

   a.   Obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ from Equation (12.16). [*Hint:* Let $\bar{y}_1$ be the mean of the first m observations and $\bar{y}_2$ be the mean of the next n observations.]

   b.   Find simple expressions for $\hat{\mathbf{y}}$, SSE, s_e, and $s_{\hat{\beta}_1}$.

   c.   Use parts (a) and (b) to find a simple expression for the 95% CI for β_1. Show that your formula is equivalent to

   $$\hat{\beta}_1 \pm t_{.025,m+n-2}s_e\sqrt{\frac{1}{m} + \frac{1}{n}}$$

   $$= \bar{y}_1 - \bar{y}_2 \pm t_{.025,m+n-2} \cdot \sqrt{\frac{1}{m} + \frac{1}{n}} \cdot$$

   $$\sqrt{\frac{\sum_{i=1}^{m}(y_i - \bar{y}_1)^2 + \sum_{i=m+1}^{m+n}(y_i - \bar{y}_2)^2}{m+n-2}}$$

   which is the pooled variance confidence interval discussed in Section 9.2.

   d.   Let $m = 3$ and $n = 3$, with $y_1 = 117$, $y_2 = 119$, $y_3 = 127$, $y_4 = 129$, $y_5 = 138$, $y_6 = 139$. These are the prices in thousands for three houses in Brookwood and then three houses in Pleasant Hills. Apply parts (a), (b), and (c) to this data set.

106. The constant term β_0 is not always needed in the regression equation. For example, many physical principles imply that the response variable should be 0 when the explanatory variables are 0, so the constant term is not needed. Then it is preferable to omit β_0 and use the model $Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$. Here we focus on the special case $k = 1$.

   a.   Differentiate the appropriate sum of squares to derive the one normal equation for estimating β_1.

   b.   Express your normal equation in matrix form, where $\mathbf{X}$ consists of a single

column with the values of the predictor variable.

c. Apply part (b) to the data of Example 12.28, using hp for **y** and just engine size in **X**.

d. Explain why deletion of the constant term might be appropriate for the data set in part (c).

e. By fitting a regression model with a constant term added to the model of part (c), test the hypothesis that the constant is not needed.

107. a. Prove that the hat matrix **H** satisfies $\mathbf{H}^2 = \mathbf{H}$.

b. Prove Equation (12.19). [*Hint:* Look at the derivation of Equation (12.18).]

108. Use Eqs. (12.18) and (12.19) to show that each of the leverages is between 0 and 1, and therefore the variances of the predicted values and residuals are between 0 and σ^2.

109. The measurements here are similar to those in Example 12.31, except that here the students did the measuring at home, and the results suffered in accuracy.

Wingspan	Foot	Height
74	13.0	75
56	8.5	66
65	10.0	69
66	9.5	66
62	9.0	54
69	11.0	72
75	12.0	75
66	9.0	63
66	9.0	66
63	8.5	63

a. Regress wingspan on the other two variables. Carry out the test of model utility and the tests for the two individual regression coefficients of the predictors.

b. Obtain the diagonal elements of the hat matrix (leverages). Identify the point with the highest leverage. What is unusual about the point? Given the instructor's assertion that there were no students in the class less than five feet tall, would you say that there was an error? Give another reason that this student's measurements seem wrong.

c. For the other points with high leverages, what distinguishes them from the points with ordinary leverage values?

d. Examining the residuals, find another student whose data might be wrong.

e. Discuss the elimination of questionable points in order to obtain valid regression results.

110. Refer back to the centered simple regression model in Exercise 102.

a. Show that the leverage h_{ii}, the ith diagonal entry of the hat matrix **H**, is given by

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

b. Show that the sum of the leverages in simple regression is 2.

c. Use part (a) and the discussion of **H** in this section to confirm the following formulas from Sections 12.4 and 12.6:

$$V(\hat{Y}_i) = \sigma^2 \cdot \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right]$$

$$V(Y_i - \hat{Y}_i) = \sigma^2 \cdot \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right]$$

12.10   Logistic Regression

All of the regression models thus far have assumed a quantitative response variable *y*. (Section 12.8 discussed how to incorporate a categorical *explanatory* variable using one or more indicators, but *y* was still numerical.) In this final section, we describe procedures for modeling the relationship

between a categorical response variable and one or more predictors. For example, university administrators may wish to predict whether a student will graduate (a yes-or-no variable) as a function of high school GPA, SAT scores, and number of extracurricular activities. Medical researchers frequently construct models to determine the effect of treatment dosage and other factors (age, weight, and so on) on whether or not someone contracts a certain disease.

The Simple Logistic Regression Model

The simple *linear* regression model is appropriate for relating a quantitative response variable y to a quantitative predictor x. But suppose we have a dichotomous categorical response variable, whose "values" are success and failure. We can encode this with a Bernoulli rv Y, with possible values 1 and 0 corresponding to success and failure. As in previous chapters, let $p = P(S) = P(Y = 1)$ and $1 - p = P(F) = P(Y = 0)$. Frequently, the value of p will depend on the value of some quantitative variable x. For example, the probability that a car needs warranty service should depend on the car's mileage, or the probability of avoiding an infection might depend on the dosage in an inoculation. Instead of using just the symbol p for the success probability, we now use $p(x)$ to emphasize the dependence of this probability on the value of x. The simple linear regression equation $Y = \beta_0 + \beta_1 x + \varepsilon$ is no longer appropriate, for taking the mean value on each side of that equation would give

$$\mu_{Y|x} = 1 \cdot p(x) + 0 \cdot [1 - p(x)] = p(x) = \beta_0 + \beta_1 x$$

Whereas $p(x)$ is a probability and therefore must be between 0 and 1, $\beta_0 + \beta_1 x$ need not be in this range.

Instead of letting the mean value of y be a linear function of x, we now consider a model in which the mean response $p(x)$ is a particular nonlinear function of x. A function that has been found quite useful in many applications is the **logit function**

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \tag{12.20}$$

It is easy to see that the logit function is bounded between 0 and 1, so $0 < p(x) < 1$ as desired. Figure 12.36 shows a graph of $p(x)$ for particular values of β_0 and β_1 with $\beta_1 > 0$. As x increases, the probability of success increases. For $\beta_1 < 0$, the success probability would be a decreasing function of x.
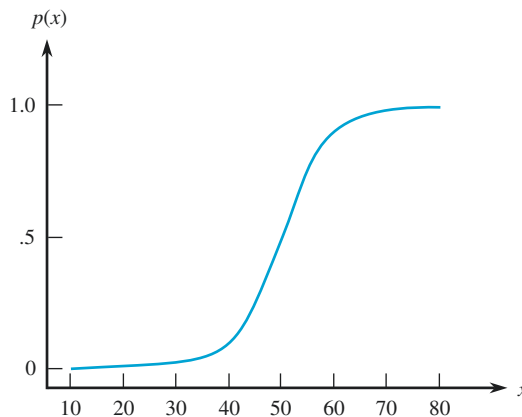

Figure 12.36 A graph of a logit function

Logistic regression means assuming that $p(x)$ is related to x by the logit function. Straightforward algebra shows that

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}$$

The expression on the left-hand side is called the **odds**. If, for example $p(60) = 3/4 = .75$, then $p(60)/(1 - p(60)) = .75/(1 - .75) = 3$ and when $x = 60$ a success is three times as likely as a failure. This is described by saying that the odds are 3 to 1 because the success probability is three times the failure probability. Taking natural logs of both sides, we see that the logarithm of the odds is a linear function of the predictor:

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

In particular, the slope parameter β_1 is the change in the log-odds associated with a one-unit increase in x. This implies that the odds itself changes by the multiplicative factor e^{β_1} when x increases by one unit. The quantity e^{β_1} is called the **odds ratio**, because it represents the ratio of the odds of success when the predictor variable equals $x + 1$ to the odds of success when the predictor variable equals x.

Example 12.32 It seems reasonable that the size of a cancerous tumor should be related to the likelihood that the cancer will spread (metastasize) to another site. The article "Molecular Detection of p16 Promoter Methylation in the Serum of Patients with Esophageal Squamous Cell Carcinoma" (*Cancer Res*. 2001: 3135–3138) investigated the spread of esophageal cancer to the lymph nodes. With x = size of a tumor (cm) and $Y = 1$ if the cancer does spread, consider the logistic regression model with $\beta_1 = .5$ and $\beta_0 = -2$ (values suggested by data in the article). Then

$$p(x) = \frac{e^{-2 + .5x}}{1 + e^{-2 + .5x}}$$

from which $p(2) = .27$ and $p(8) = .88$ (tumor sizes for patients in the study ranged from 1.7 to 9.0 cm). Because $e^{-2 + .5(6.77)} \approx 4$, the odds for a 6.77 cm tumor are 4, so that it is four times as likely as not that a tumor of this size will spread to the lymph nodes. Finally, for every 1-cm increase in tumor size, the odds of metastasis increase by a multiplicative factor of $e^{.5} \approx 1.65$, or 65%. Be careful here: the *probability* of metastasis is not increasing by 65%, but rather the odds; under the logistic regression model, the probability of an outcome does not increase linearly with x (see Figure 12.36). ∎

Fitting the Simple Logistic Regression Model

Fitting the logit model (12.20) to sample data requires that the parameters β_0 and β_1 be estimated. Rather than apply the principle of least squares from linear regression, the standard way to estimate logistic regression parameters is by the method of maximum likelihood. Suppose, for example, that $n = 5$ and that the observations made at x_2, x_4, and x_5 are successes whereas the other two observations are failures. Then the likelihood function is

$$\begin{aligned}
L(\beta_0, \beta_1) &= P(Y_1 = 0, Y_2 = 1, Y_3 = 0, Y_4 = 1, Y_5 = 1) \\
&= [1 - p(x_1)][p(x_2)][1 - p(x_3)][p(x_4)][p(x_5)] \\
&= \left[\frac{1}{1 + e^{\beta_0 + \beta_1 x_1}}\right]\left[\frac{e^{\beta_0 + \beta_1 x_2}}{1 + e^{\beta_0 + \beta_1 x_2}}\right]\left[\frac{1}{1 + e^{\beta_0 + \beta_1 x_3}}\right]\left[\frac{e^{\beta_0 + \beta_1 x_4}}{1 + e^{\beta_0 + \beta_1 x_4}}\right]\left[\frac{e^{\beta_0 + \beta_1 x_5}}{1 + e^{\beta_0 + \beta_1 x_5}}\right]
\end{aligned}$$

Unfortunately it is not at all straightforward to maximize this likelihood, and there are no nice formulas for the mles $\hat{\beta}_0$ and $\hat{\beta}_1$. The maximization process must be carried out using iterative numerical methods. The details are involved, but fortunately the most popular statistical software packages will do this on request and provide both quantitative and graphical indications of how well the model fits.

In particular, the mle $\hat{\beta}_1$ is typically provided along with its estimated standard deviation $s_{\hat{\beta}_1}$. For large n, the mle has approximately a normal distribution and the standardized variable $(\hat{\beta}_1 - \beta_1)/S_{\hat{\beta}_1}$ has approximately a standard normal distribution. This allows for calculation of a confidence interval for β_1 as well as for testing H_0: $\beta_1 = 0$, according to which the value of x has no impact on the likelihood of success.

Example 12.33 The following data resulted from a study commissioned by a large management consulting company to investigate the relationship between amount of job experience (x, in months) for a junior consultant and the likelihood of the consultant being able to perform a certain complex task. The value $y = 1$ indicates the consultant completed the task (success), whereas $y = 0$ corresponds to failure.

x	4	5	6	6	7	8	9	10	11	11	13	13	14	15	18
y	0	0	0	0	0	1	0	0	0	0	1	0	1	0	1

x	18	19	20	20	21	21	22	23	25	26	27	28	29	30	32
y	0	0	1	0	1	1	1	0	1	1	0	1	1	1	1

Figure 12.37 shows Minitab output for a logistic regression analysis. The estimates of the parameters β_0 and β_1 are $\hat{\beta}_0 = -3.21107$ and $\hat{\beta}_1 = 0.177717$, respectively. The resulting *estimated* logistic regression function, denoted $\hat{p}(x)$, is

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} = \frac{e^{-3.211 + 0.1777x}}{1 + e^{-3.211 + 0.1777x}}$$

The graph of $\hat{p}(x)$ is the curve shown in Figure 12.38; notice that the (estimated) probability of success increases as x increases. Remember that the logit curve is modeling the *mean y* value for each x value; we do not anticipate that it will intersect the points in the scatterplot.

```
Binary Logistic Regression: Success versus Months

Logistic Regression Table

                                        Odds     95% CI
Predictor      Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant   -3.21107    1.23540  -2.60  0.009
Months     0.177717  0.0657308   2.70  0.007   1.19   1.05   1.36
```

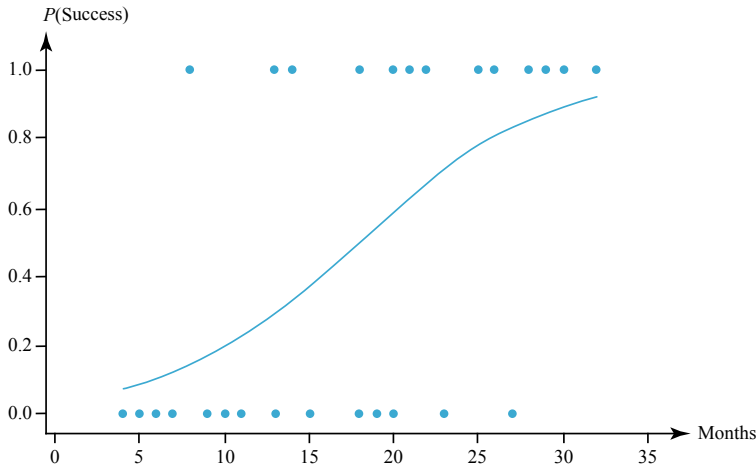**Figure 12.37** Logistic regression output from Minitab

Figure 12.38 Scatterplot with the fitted logistic regression function for Example 12.33

We may use $\hat{p}(x)$ to estimate the likelihood of a junior consultant completing the complex task, based upon her/his duration of job experience. For example,

$$\hat{p}(12) = \frac{e^{-3.211 + 0.1777(12)}}{1 + e^{-3.211 + 0.1777(12)}} = .254 \quad \text{and} \quad \hat{p}(24) = \frac{e^{-3.211 + 0.1777(24)}}{1 + e^{-3.211 + 0.1777(24)}} = .742$$

So, it is estimated that a consultant with just one year (12 months) of experience has about a .25 chance of successfully completing the task, compared to a probability of over .74 for someone with two years' experience.

The Minitab output includes $s_{\hat{\beta}_1}$ under SE Coef. For the "utility test" of $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, the test statistic value and two–tailed P-value are

$$z = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{.177717 - 0}{.0657308} = 2.70 \qquad P\text{-value} = 2P(Z \geq 2.70) = 2[1 - \Phi(2.70)] = .007$$

The null hypothesis is rejected at the .05 or .01 level, and we conclude that months of experience is a useful predictor of a junior consultant's ability to complete the task.

The estimated odds ratio is $e^{\hat{\beta}_1} = e^{0.1777} = 1.19$. A 95% CI for β_1 is given by

$$\hat{\beta}_1 \pm z_{.025}s_{\hat{\beta}_1} = .177717 \pm 1.96(.0657308) = (.04888, .30655)$$

from which a 95% CI for the true odds ratio, e^{β_1}, is $(e^{.04888}, e^{.30655}) = (1.05, 1.36)$. The estimated odds ratio and the CI all appear in the output. With a high degree of confidence, for each additional month of experience, the *odds* that a consultant can successfully complete the task increase by a multiplicative factor of between 1.05 and 1.36, i.e., increase by 5–36%. ∎

Some software packages report the value of the chi-squared statistic z^2 rather than z itself, along with the corresponding P-value for a two-tailed test.

Example 12.34   Here is data on launch temperature (°F) and the incidence of failure for O-rings in 23 space shuttle launches prior to the *Challenger* disaster of January 28, 1986.

Temperature	Failure	Temperature	Failure	Temperature	Failure
53	Y	68	N	75	N
57	Y	69	N	75	Y
58	Y	70	N	76	N
63	Y	70	N	76	N
66	N	70	Y	78	N
67	N	70	Y	79	N
67	N	72	N	81	N
67	N	73	N		

Figure 12.39 shows JMP output from a logistic regression analysis. We have chosen to let p denote the probability of an O-ring failure, since this is really the event of interest. Failures tended to occur at lower temperatures and successes at higher temperatures, so the graph of $\hat{p}(x)$ decreases as temperature (x) increases.
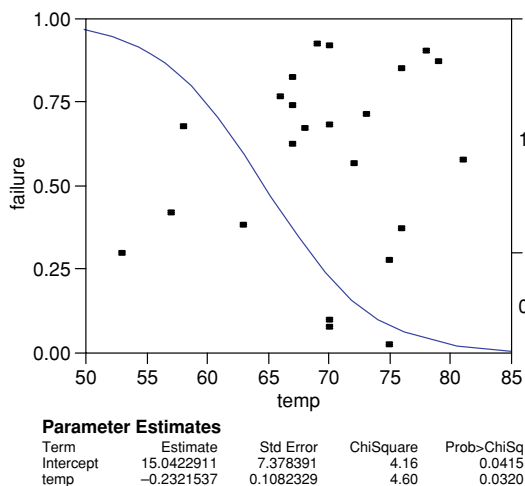

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	15.0422911	7.378391	4.16	0.0415
temp	−0.2321537	0.1082329	4.60	0.0320

Figure 12.39   Logistic regression output from JMP

The estimate of β_1 is $\hat{\beta}_1 = -.2322$, and the estimated standard deviation of $\hat{\beta}_1$ is $s_{\hat{\beta}_1} = .1082$. The value of z for testing $H_0: \beta_1 = 0$, which asserts that temperature does not affect the likelihood of O-ring failure, is $z = \hat{\beta}_1/s_{\hat{\beta}_1} = -.2322/.1082 = -2.15$. The P-value is $2P(Z \leq -2.15) = 2(.0158) = .032$. JMP reports the value of a chi-squared statistic computed as $(-2.15)^2 \approx 4.60$ (there is a slight disparity due to rounding in the z value). Either way, the P-value indicates that H_0 should be rejected at the .05 level and, hence, that temperature at launch time has a statistically significant effect on the likelihood of an O-ring failure. Specifically, for each 1 °F increase in launch temperature, we estimate that the odds of failure are multiplied by a factor of $e^{\hat{\beta}_1} = e^{-.2322} \approx .79$, i.e., the odds are estimated to decrease by 21%.

The launch temperature for the *Challenger* mission was only 31 °F. Because this value is much smaller than any temperature in the sample, it is dangerous to extrapolate the estimated relationship. Nevertheless, it appears that for a temperature this small, O-ring failure is almost a sure thing. The logistic regression gives the estimated probability at $x = 31$ as

$$\hat{p}(31) = \frac{e^{\beta_0 + \beta_1(31)}}{1 + e^{\beta_0 + \beta_1(31)}} = \frac{e^{15.0423 - .23215(31)}}{1 + e^{15.0423 - .23215(31)}} = .99961$$

and the odds associated with this probability are $.99961/(1 - .99961) \approx 2563$. Thus, if the logistic regression can be extrapolated down to 31°F, the probability of failure is .99961, the probability of success is .00039, and the predicted odds are 2563 to 1 against avoiding an O-ring failure.   ∎

Multiple Logistic Regression

Multiple logistic regression, a natural extension of simple logistic regression, postulates a model for relating a categorical response variable to more than one explanatory variable. The explanatory variables themselves may be true quantitative predictors or indicator variables coding categorical predictors. We continue to restrict attention to a binary response, such as yes/no or happy/sad, which may be coded as 1 or 0 (with 1 indicating the event of interest, i.e., a "success").

With predictors $x_1, \ldots, x_k$ in the model, let $p(x_1, \ldots, x_k)$ denote the true probability of the event of interest occurring and assume the following **multiple logit function** applies:

$$p(x_1, \ldots, x_k) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}} \qquad (12.21)$$

The multiple logit function (12.21) is the obvious extension of the simple logit function (12.20) to accommodate more than one explanatory variable. As in simple logistic regression, this logit function can be re-written in terms of the natural log of the odds of the event of interest:

$$\ln\left(\frac{p(x_1, \ldots, x_k)}{1 - p(x_1, \ldots, x_k)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Written this way, the coefficient β_j ($j = 1, \ldots, k$) is interpreted as the change in the log-odds of the event of interest associated with a one-unit increase in x_j, after adjusting for the effects of all the other predictors in the model. Equivalently, e^{β_j} is the multiplicative change in odds associated with a one-unit increase in x_j after accounting for the other $k - 1$ predictors, i.e., e^{β_j} is the odds ratio associated with x_j.

Inference procedures in multiple logistic regression are similar to those outlined for simple logistic regression. In particular, the point estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ for the unknown β_j's are based upon the principle of maximum likelihood, and each estimator has approximately a normal sampling distribution provided the sample size n is reasonably large. Several statistical software packages will provide point estimates and estimated standard errors for the coefficients, allowing for variable utility hypothesis tests as well as confidence intervals.

Example 12.35 The authors of the article "Building Social Capital in Forest Communities: Analysis of New Mexico's Collaborative Forest Restoration Program" (*Natural Resour. J.*, Fall 2007: 867–915) analyzed the factors that helped determine which proposals were funded by the

Collaborative Forest Restoration Program (CFRP, a federally funded grant program). Data was available on 219 proposals made in New Mexico between 2001 and 2006. The response variable of interest is

$$y = \begin{cases} 1 & \text{if the grant proposal was funded} \\ 0 & \text{if the grant proposal was not funded} \end{cases}$$

We will consider just a few of the predictor variables the authors used in their analysis: x_1 = amount of funding requested by the project (in \$1000), x_2 = percent of county residents living below the poverty threshold, and x_3 = 1 if the proposed treatment of private lands was cited by the review panel as a weakness of the project ($x_3 = 0$ otherwise).

Parameter estimates from software are $\hat{\beta}_0 = -1.216$, $\hat{\beta}_1 = .00156$, $\hat{\beta}_2 = .0327$, and $\hat{\beta}_3 = -2.002$. Consider a proposal requesting \$360,000 ($x_1 = 360$) that was not criticized for its proposed treatment of private lands ($x_3 = 0$) in a county with a 16.6% poverty rate ($x_2 = 16.6$); this exactly matches one of the proposals. Then the estimated log-odds of the project being funded are

$$-1.216 + .00156(360) + .0327(16.6) - 2.002(0) = -.11158$$

and the estimated probability of being funded is

$$\hat{p}(360, 16.6, 0) = \frac{e^{-.11158}}{1 + e^{-.11158}} = .4721$$

(For the record, that particular proposal *was* funded!)

Funding request amount had little practical impact on whether a project was funded: adjusting for poverty rate and private land use consideration, a \$1000 (one-unit) increase in requested funding actually *increased* the estimated odds of acceptance by $e^{.00156} = 1.0016$, i.e., by .16%. In contrast, criticism for private land treatment was a veritable death-knell: removing the effects of the other two variables, odds of acceptance when $x_3 = 1$ are $e^{-2.002} = .1351$ times the acceptance odds when $x_3 = 0$. In other words, if a proposal was criticized in this way, the odds of acceptance were reduced by more than 86%. ∎

A model utility test of H_0: $\beta_1 = \cdots = \beta_k = 0$ versus H_a: not all β's are zero is based on the likelihood ratio test statistic Λ presented in Section 9.5; most statistical software packages will include the test statistic value and P-value when multiple logistic regression is performed. (The test is based on the large-sample approximation mentioned at the end of Section 9.5, whereby $-2\ln(\Lambda)$ has approximately a chi-squared distribution with k df.)

The logit functions in (12.20) and (12.21) are not the only choices for modeling the probability of success. Two other popular options are the *probit* and *complimentary log–log* functions, both of which are implemented in many software packages. The relative suitability of these functions to fitting a particular data set can be assessed using various automated "goodness-of-fit" procedures, including the *deviance test* and the *Hosmer-Lemeshow test*. Consult the text by Kutner et al. listed in the bibliography for more information.

Exercises: Section 12.10 (111–120)

111. A major electronics retailer sensibly believes that customers are more likely to redeem an emailed coupon if it's worth more money. With $x =$ coupon discount amount (\$), and $Y = 1$ if a customer redeems the coupon, consider a logistic regression model with $\beta_0 = -3.75$ and $\beta_1 = 0.1$.

 a. Calculate and interpret both $p(10)$ and $p(50)$.
 b. Calculate the odds that a \$10 coupon is redeemed, then repeat for \$50.
 c. Interpret β_1 in this context.
 d. According to this model, for what discount amount is there a 50–50 chance the coupon will be redeemed?

112. In Example 12.32, the probability of cancer metastasizing was given by the logistic regression model with $\beta_0 = -2$ and $\beta_1 = 0.5$.

 a. Tabulate values of x, $p(x)$, the odds $p(x)/[1 - p(x)]$, and the log-odds for $x = 2, 3, 4, \ldots, 9$. (In the cited article, tumor sizes ranged from 1.7 to 9.0 cm.)
 b. Explain what happens to the odds when x is increased by 1. Your explanation should involve the .5 that appears in the formula for $p(x)$.
 c. Support your answer to (b) algebraically, starting from the formula for $p(x)$.
 d. For what value of x are the odds 1? 5? 10?

113. Adolescents are getting less sleep that ever before, and this can have serious behavioral repercussions. The article "Dose-Dependent Associations Between Sleep Duration and Unsafe Behaviors Among US High-School Students" (*JAMA Pediatr.* 2018: 1187–1189) reported a large-scale study of American teenagers. The investigators fit a simple logistic regression model with the response variable $y = 1$ if a teenager had driven drunk in the last 30 days (and 0 otherwise), and $x =$ typical number of hours of sleep per night. Information in the article suggests $\hat{\beta}_1 = -.1998$ and $s_{\hat{\beta}_1} = .0986$.

 a. Test whether sleep has an effect on the likelihood of driving drunk among American teenagers, at the .05 significance level.
 b. Calculate a 95% confidence interval for e^{β_1}.
 c. Interpret the confidence interval from part (b) in terms of a one-hour *decrease* in sleep.

114. The pharmaceutical industry has increasingly developed "nanoformulations" for drug delivery, but quality control at such a small scale is tricky. The article "Quality by Design Approach Using Multiple Linear and Logistic Regression Modeling Enables Microemulsion Scale Up" (*Molecules* 2019) describes one study to determine how $x =$ oil concentration (g/100 mL) affects whether a development run meets a certain critical quality attribute (CQA) with respect to polydispersity. Here, $y = 1$ if the CQA was achieved and $= 0$ if not.

x	6.0	6.0	2.0	4.0	6.0	2.0	2.0	4.0	2.0	6.0
y	0	0	1	1	0	1	1	1	1	0

x	2.0	2.0	6.0	2.0	4.0	6.0	2.0	6.0	2.0	2.0
y	1	1	0	1	0	0	1	0	1	1

x	4.5	2.0	2.0	6.0	2.0	2.0	2.0	4.5	4.0	2.0
y	0	1	1	0	1	1	1	1	0	1

Software reports coefficients $\hat{\beta}_0 = 11.13$ and $\hat{\beta}_1 = -2.68$ with estimated standard errors 5.96 and 1.42, respectively.

 a. Write out the estimated logit function, and use it to estimate $p(2)$, $p(4)$, and $p(6)$.
 b. Does the data provide convincing statistical evidence that oil concentration affects the chance of meeting this particular CQA? Test at the .05 significance level.

c. Construct a 95% CI for β_1, then use this to give an interval estimate for e^{β_1}. Interpret the latter interval.

d. What does e^{β_0} represent in this context? Does that interpretation seem appropriate here? Why or why not?

115. Kyphosis, or severe forward flexion of the spine, may persist despite corrective spinal surgery. A study carried out to determine risk factors for kyphosis reported the following ages (months) for 40 subjects at the time of the operation; the first 18 subjects did have kyphosis and the remaining 22 did not.

Kyphosis	12	15	42	52	59	73
	82	91	96	105	114	120
	121	128	130	139	139	157
No kyphosis	1	1	2	8	11	18
	22	31	37	61	72	81
	97	112	118	127	131	140
	151	159	177	206		

a. Use software to fit a logistic regression model to this data.

b. Interpret the coefficient $\hat{\beta}_1$. [*Hint:* It might be more sensible to work in terms of $e^{\hat{\beta}_1}$.]

c. Test whether age has a statistically significant impact on the presence of kyphosis.

116. Exercise 16 of Chapter 1 presented data on the noise level (dBA) for 77 individuals working at a particular office. In fact, each person was also asked whether the noise level in the office was acceptable or not.

Acceptable	55.3 55.3 55.3 55.9 55.9 55.9 55.9 56.1
	56.1 56.1 56.1 56.1 56.1 56.8 56.8 57.0
	57.0 57.0 57.8 57.8 57.8 57.9 57.9 57.9
	58.8 58.8 58.8 59.8 59.8 59.8 62.2 62.2
	65.3 65.3 65.3 65.3 68.7 69.0 73.0 73.0
Unacceptable	63.8 63.8 63.8 63.9 63.9 63.9 64.7 64.7
	64.7 65.1 65.1 65.1 67.4 67.4 67.4 67.4
	68.7 68.7 68.7 70.4 70.4 71.2 71.2 73.1
	73.1 74.6 74.6 74.6 74.6 79.3 79.3 79.3
	79.3 79.3 83.0 83.0 83.0

a. Use software to fit a logistic regression model to this data.

b. Interpret the coefficient $\hat{\beta}_1$. [*Hint:* It might be more sensible to work in terms of $e^{\hat{\beta}_1}$.]

c. Construct and interpret a 95% confidence interval for e^{β_1}.

117. The article "Consumer Attitudes Toward Genetic Modification and Other Possible Production Attributes for Chicken" (*J. Food Distr. Res.* 2005: 1–11) reported a survey of 498 randomly selected consumers concerning their views on genetically modified (GM) food. The researchers' goal was to model the response variable $Y = 1$ if a consumer wants GM chicken products labeled (and 0 otherwise) as a function of $x_1 = $ consumer's age (yr), $x_2 = $ income ($1000s), sex ($x_3 = 1$ if female), and whether there are children in the consumer's household ($x_4 = 1$ if yes). Estimated model parameter values are $\hat{\beta}_0 = .8247$, $\hat{\beta}_1 = .0073$, $\hat{\beta}_2 = .0041$, $\hat{\beta}_3 = .9910$, and $\hat{\beta}_4 = .0224$.

a. Estimate the likelihood that a consumer wants GM chicken products labeled if that person is a 35-year-old female with $65,000 annual income and no children.

b. Repeat part (a) for a 35-year-old male (keep other features the same).

c. Interpret the coefficient on age.

d. Interpret the coefficient on the sex indicator variable.

118. Road trauma is the leading cause of death and injury among young people in Australia. The article "The Journey from Traffic Offender to Severe Road Trauma Victim: Destiny or Preventive Opportunity?" (*PLoS ONE*, April 22, 2015) reported a study to determine factors that might help predict future serious accidents. The article included estimated odds ratios and 95% CIs for true odds ratios for several variables. The response variable here has value $y = 1$ if a subject was in an accident leading to intensive care admission or death, and 0 otherwise.

	Est. OR	OR 95% CI
x_1 = age/10	1.02	(1.01, 1.03)
x_2 = 1 if male, 0 if female	1.18	(0.98, 1.42)
x_3 = years with a driver's license	0.99	(0.98, 0.99)
x_4 = number of prior traffic offenses	1.10	(1.08, 1.11)

a. Which of these four explanatory variables were associated with a *decreased* likelihood of later severe road trauma? How can you tell?

b. Which of these four explanatory variables were *not* statistically significant predictors in this model? How can you tell?

c. Interpret the 95% CI provided for e^{β_4}.

119. Whale-watching is big business in Alaska, particularly around salmon release sites where whales tend to congregate. The article "Humpback Whales Feed on Hatchery-Released Juvenile Salmon" (*Roy. Soc. Open Sci.* 2017) reported a study to determine what factors help predict the likelihood of spotting a humpback whale when visiting one of these sites. The following data on x_1 = days after final salmon release, x_2 = duration of visit (min), and whether a whale was sighted are from a recent year at Little Port Walter.

x_1	x_2	Whale?	x_1	x_2	Whale?
2	15	Y	7	15	N
1	15	N	7	15	N
1	15	N	8	15	N
2	15	N	8	15	N
3	15	N	9	15	N
3	15	N	9	15	N
4	15	N	10	15	N
5	30	N	12	15	N
5	15	N	12	15	N
6	15	N	13	15	N
6	15	N	13	35	Y

(The full study investigated five sites across several years before, during, and after salmon release.)

a. Use software to fit a multiple logistic regression model to this data, and confirm that the estimated log-odds function is $-5.68 - .096x_1 + .210x_2$.

b. What does the negative sign for the coefficient $-.096$ signify? What does the positive sign for the coefficient .210 signify? [*Hint:* The latter should not be surprising.]

c. Estimate the probability of spotting a humpback whale during a 30-minute tour one week (i.e., seven days) after the final salmon release.

d. The estimated standard errors of the coefficients are $s_{\hat{\beta}_1} = .253$ and $s_{\hat{\beta}_2} = .120$. Perform variable utility tests at the .1 significance level.

e. Interpret both $e^{-.096}$ and $e^{.210}$ in this context.

120. The article "Developing Coal Pillar Stability Chart Using Logistic Regression" (*J. Rock Mech. Mining Sci.* 2013: 55–60) includes the following data on x_1 = height–width ratio, x_2 = strength–stress ratio, and y = 1 (stable) or 0 (not stable) for 29 pillars used to stabilize current and former mines in India.

x_1	1.80	1.65	2.70	3.67	1.41	1.76	2.10	2.10
x_2	2.40	2.54	0.84	1.68	2.41	1.93	1.77	1.50
y	1	1	1	1	1	1	1	1

x_1	4.57	3.59	8.33	2.86	2.58	2.90	3.89	0.80
x_2	2.43	5.55	2.58	2.00	3.68	1.13	2.49	1.37
y	1	1	1	1	1	1	1	0

x_1	0.60	1.30	0.83	0.57	1.44	2.08	1.50	1.38
x_2	1.27	0.87	0.97	0.94	1.00	0.78	1.03	0.82
y	0	0	0	0	0	0	0	0

x_1	0.94	1.58	1.67	3.00	2.21
x_2	1.30	0.83	1.05	1.19	0.86
y	0	0	0	0	0

a. Fit a multiple logistic regression model to this data, and report the estimated logit equation.

b. Perform the two variable utility tests $H_0: \beta_1 = 0$ and $H_0: \beta_2 = 0$, each at the .1 significance level.

c. Calculate the predicted probability of stability for pillar 8 ($x_1 = 2.10$, $x_2 = 1.50$).

d. Calculate the predicted probability of stability for pillar 28 ($x_1 = 3.00$, $x_2 = 1.19$).

Supplementary Exercises: (121–138)

121. In anticipation of future floods, insurance companies must quantify the relationship between water depth and the amount of flood damage that will occur. The Federal Insurance Administration provided the following information on x = depth of flooding (in feet above first-floor level) and y = flood damage (as a percentage of structural value) for homes with no basements.

Flood level (x)	Flood damage (y)	Flood level (x)	Flood damage (y)
0	7	8	44
1	10	9	45
2	14	10	46
3	26	11	47
4	28	12	48
5	29	13	49
6	41	14	50
7	43		

a. Create a scatterplot of the data, and briefly describe what you see.

b. Does a straight-line relationship seem appropriate for this data? Why or why not?

122. The article "Exhaust Emissions from Four-Stroke Lawn Mower Engines" (*J. Air Water Manage. Assoc.* 1997: 945–952) reported data from a study in which both a baseline gasoline mixture and a reformulated gasoline were used. Consider the following observations on age (year) and NO_x emissions (g/kWh):

Engine	1	2	3	4	5
Age	0	0	2	11	7
Baseline	1.72	4.38	4.06	1.26	5.31
Reformulated	1.88	5.93	5.54	2.67	6.53

Engine	6	7	8	9	10
Age	16	9	0	12	4
Baseline	.57	3.37	3.44	.74	1.24
Reformulated	.74	4.94	4.89	.69	1.42

a. Construct a scatterplot of baseline vs. reformulated NO_x emissions. Comment on what you find.

b. Construct scatterplots of NO_x emissions versus age. What appears to be the nature of the relationship between these two variables?

123. The presence of hard alloy carbides in high chromium white iron alloys results in excellent abrasion resistance, making them suitable for materials handling in the mining and materials processing industries. The accompanying data on x = retained austenite content (%) and y = abrasive wear loss (mm^3) in pin wear tests with garnet as the abrasive was read from a plot in the article "Microstructure-Property Relationships in High Chromium White Iron Alloys" (*Internat. Mater. Rev.* 1996: 59–82). Refer to the accompanying SAS output.

x	4.6	17.0	17.4	18.0	18.5	22.4	26.5	30.0	34.0
y	.66	.92	1.45	1.03	.70	.73	1.20	.80	.91

x	38.8	48.2	63.5	65.8	73.9	77.2	79.8	84.0
y	1.19	1.15	1.12	1.37	1.45	1.50	1.36	1.29

a. What proportion of observed variation in wear loss can be attributed to the simple linear regression model relationship?

b. What is the value of the sample correlation coefficient?

c. Test the utility of the simple linear regression model using $\alpha = .01$.

d. Estimate the true average wear loss when content is 50% and do so in a way that conveys information about reliability and precision.

e. What value of wear loss would you predict when content is 30%, and what is the value of the corresponding residual?

```
                              Analysis of variance

Source          DF              Sum of squares        Mean square        F Value        Prob > F
Model           1                   0.63690            0.63690            15.444          0.0013
Error           15                  0.61860            0.04124
C Total         16                  1.25551
                Root MSE            0.20308            R-square            0.5073
                Dep Mean            1.10765            Adj R-sq            0.4744
                C.V.               18.33410

                              Parameter estimates

Variable     DF     Parameter estimate     Standard error     T for H0: Parameter = 0     Prob > |T|
INTERCEP     1          0.787218              0.09525879               8.264                  0.0001
AUSTCONT     1          0.007570              0.00192626               3.930                  0.0013
```

124. An investigation was carried out to study the relationship between speed (ft/s) and stride rate (number of steps taken/s) among female marathon runners. Resulting summary quantities included $n = 11$, $\sum(\text{speed}) = 205.4$, $\sum(\text{speed})^2 = 3880.08$, $\sum(\text{rate}) = 35.16$, $\sum(\text{rate})^2 = 112.681$, and $\sum(\text{speed})(\text{rate}) = 660.130$.

   a. Calculate the equation of the least squares line that you would use to predict stride rate from speed. [*Hint:* $\bar{x} = \sum x_i/n$ and similarly for $\bar{y}$; $S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n$ and similarly for S_{xx} and S_{yy}.]

   b. Calculate the equation of the least squares line that you would use to predict speed from stride rate.

   c. Calculate the coefficient of determination for the regression of stride rate on speed of part (a) and for the regression of speed on stride rate of part (b). How are these related?

   d. How is the product of the two slope estimates related to the value calculated in (c)?

125. Suppose that x and y are positive variables and that a sample of n pairs results in $r \approx 1$. If the sample correlation coefficient is computed for the (x, y^2) pairs, will the resulting value also be approximately 1? Explain.

126. In Section 12.4, we presented a formula for the variance $V(\hat{\beta}_0 + \hat{\beta}_1 x^*)$ and a CI for $\beta_0 + \beta_1 x^*$. Taking $x^* = 0$ gives $\sigma_{\hat{\beta}_0}^2$ and a CI for β_0. Use the data of Example 12.18 to calculate the estimated standard deviation

of $\hat{\beta}_0$ and a 95% CI for the *y*-intercept of the true regression line.

127. In biofiltration of wastewater, air discharged from a treatment facility is passed through a damp porous membrane that causes contaminants to dissolve in water and be transformed into harmless products. The accompanying data on $x =$ inlet temperature (°C) and $y =$ removal efficiency (%) was the basis for a scatterplot that appeared in the article "Treatment of Mixed Hydrogen Sulfide and Organic Vapors in a Rock Medium Biofilter" (*Water Environ. Res.* 2001: 426–435).

Obs	Temp	Removal %	Obs	Temp	Removal %
1	7.68	98.09	17	8.55	98.27
2	6.51	98.25	18	7.57	98.00
3	6.43	97.82	19	6.94	98.09
4	5.48	97.82	20	8.32	98.25
5	6.57	97.82	21	10.50	98.41
6	10.22	97.93	22	16.02	98.51
7	15.69	98.38	23	17.83	98.71
8	16.77	98.89	24	17.03	98.79
9	17.13	98.96	25	16.18	98.87
10	17.63	98.90	26	16.26	98.76
11	16.72	98.68	27	14.44	98.58
12	15.45	98.69	28	12.78	98.73
13	12.06	98.51	29	12.25	98.45
14	11.44	98.09	30	11.69	98.37
15	10.17	98.25	31	11.34	98.36
16	9.64	98.36	32	10.97	98.45

Calculated summary quantities are $\sum x_i = 384.26$, $\sum y_i = 3149.04$, $S_{xx} = 485.00$, $S_{xy} = 36.71$, and $S_{yy} = 3.50$.

   a. Does a scatterplot of the data suggest appropriateness of the simple linear regression model?

   b. Fit the simple linear regression model, obtain a point prediction of removal

efficiency when temperature = 10.50, and calculate the value of the corresponding residual.

c. Roughly what is the size of a typical deviation of points in the scatterplot from the least squares line?

d. What proportion of observed variation in removal efficiency can be attributed to the model relationship?

e. Estimate the slope coefficient in a way that conveys information about reliability and precision, and interpret your estimate.

f. Personal communication with the authors of the article revealed that one additional observation was not included in their scatterplot: (6.53, 96.55). What impact does this additional observation have on the equation of the least squares line and the values of s and R^2?

128. Normal hatchery processes in aquaculture inevitably produce stress in fish, which may negatively impact growth, reproduction, flesh quality, and susceptibility to disease. Such stress manifests itself in elevated and sustained corticosteroid levels. The article "Evaluation of Simple Instruments for the Measurement of Blood Glucose and Lactate, and Plasma Protein as Stress Indicators in Fish" (*J. World Aquacult. Soc.* 1999: 276–284) described an experiment in which fish were subjected to a stress protocol and then removed and tested at various times after the protocol had been applied. The accompanying data on x = time (min) and y = blood glucose level (mmol/L) was read from a plot.

x	2	2	5	7	12	13	17	18	23	24	26	28
y	4.0	3.6	3.7	4.0	3.8	4.0	5.1	3.9	4.4	4.3	4.3	4.4

x	29	30	34	36	40	41	44	56	56	57	60	60
y	5.8	4.3	5.5	5.6	5.1	5.7	6.1	5.1	5.9	6.8	4.9	5.7

Use the methods developed in this chapter to analyze the data, and write a brief report summarizing your conclusions (assume that

the investigators are particularly interested in glucose level 30 min after stress).

129. The article "Evaluating the BOD POD for Assessing Body Fat in Collegiate Football Players" (*Med. Sci. Sports Exerc.* 1999: 1350–1356) reports on a new air displacement device for measuring body fat. The customary procedure utilizes the hydrostatic weighing device, which measures the percentage of body fat by means of water displacement. Here is representative data read from a graph in the paper.

a. Use various methods to decide whether it is plausible that the two techniques measure on average the same amount of fat.

b. Use the data to develop a way of predicting an HW measurement from a BOD POD measurement, and investigate the effectiveness of such predictions.

BOD	2.5	4.0	4.1	6.2	7.1	7.0
HW	8.0	6.2	9.2	6.4	8.6	12.2

BOD	8.3	9.2	9.3	12.0	12.2
HW	7.2	12.0	14.9	12.1	15.3

BOD	12.6	14.2	14.4	15.1	15.2
HW	14.8	14.3	16.3	17.9	19.5

BOD	16.3	17.1	17.9	17.9
HW	17.5	14.3	18.3	16.2

130. Reconsider the situation of Exercise 123, in which x = retained austenite content using a garnet abrasive and y = abrasive wear loss were related via the simple linear regression model $Y = \beta_0 + \beta_1 x + \varepsilon$. Suppose that for a second type of abrasive, these variables are also related via the simple linear regression model $Y = \gamma_0 + \gamma_1 x + \varepsilon$ and that $V(\varepsilon) = \sigma^2$ for both types of abrasive. If the data set consists of n_1 observations on the first abrasive and n_2 on the second and if SSE_1 and SSE_2 denote the two error sums of squares, then a pooled estimate of σ^2 is $\hat{\sigma}^2 = (SSE_1 + SSE_2)/(n_1 + n_2 - 4)$. Let SS_{x1} and SS_{x2} denote $\sum (x_i - \bar{x})^2$ for the data on the first and second abrasives, respectively. A test of $H_0: \beta_1 - \gamma_1 = 0$ (equal slopes) is based on the statistic

$$T = \frac{\hat{\beta}_1 - \hat{\gamma}_1}{\hat{\sigma}\sqrt{\dfrac{1}{SS_{x1}} + \dfrac{1}{SS_{x2}}}}$$

When H_0 is true, T has a t distribution with $n_1 + n_2 - 4$ df. Suppose the 15 observations using the alternative abrasive give $SS_{x2} = 7152.5578$, $\hat{\gamma}_1 = .006845$, and $SSE_2 = .51350$. Using this along with the data of Exercise 123, carry out a test at level .05 to see whether expected change in wear loss associated with a 1% increase in austenite content is identical for the two types of abrasive.

131. Show that the ANOVA version of the model utility test discussed in Section 12.3 (with test statistic $F = MSR/MSE$) is in fact a likelihood ratio test for $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$. [*Hint*: We have already pointed out that the least squares estimates of β_0 and β_1 are the mle's. What is the mle of β_0 when H_0 is true? Now determine the mle of σ^2 both in Ω (when β_1 is not necessarily 0) and in Ω_0 (when H_0 is true).]

132. Show that the t ratio version of the model utility test is equivalent to the ANOVA F statistic version of the test. Equivalent here means that rejecting $H_0: \beta_1 = 0$ when either $t \geq t_{\alpha/2,n-2}$ or $t \leq -t_{\alpha/2,n-2}$ is the same as rejecting H_0 when $f \geq F_{\alpha,1,n-2}$.

133. When a scatterplot of bivariate data shows a pattern resembling an exponentially increasing or decreasing curve, the following *multiplicative* exponential model is often used: $Y = \alpha e^{\beta x} \cdot \varepsilon$.

    a. What does this multiplicative model imply about the relationship between $Y' = \ln(Y)$ and x? [*Hint*: take logs on both sides of the model equation and let $\beta_0 = \ln(\alpha)$, $\beta_1 = \beta$, $\varepsilon' = \ln(\varepsilon)$, and suppose that ε has a lognormal distribution.]

    b. The accompanying data resulted from an investigation of how road pulse duration (y, in ms, a measure of structural stress) varied with asphalt depth (x, in mm) in a simulation of large trucks driving 40 mph ("Comparative Study of Asphalt Pavement Responses Under FWD and Moving Vehicular Loading," *J. Transp. Engr.* 2016).

x	40	40	190	190	267	267	420	420
y	25	36	53	55	78	91	168	201

    Fit the simple linear regression model to this data, and check model adequacy using the residuals.

    c. Is a scatterplot of the data consistent with the exponential regression model? Fit this model by first carrying out a simple linear regression analysis using $\ln(y)$ as the response variable and x as the explanatory variable. How good a fit is the simple linear regression model to the "transformed" data (i.e., the $(x, \ln(y))$ pairs)? What are point estimates of the parameters α and β?

    d. Obtain a 95% prediction interval for pulse duration when asphalt thickness is 250 mm. [*Hint*: first obtain a PI for $\ln(y)$ based on the simple linear regression carried out in (c).]

134. No tortilla chip aficionado likes soggy chips, so it is important to identify characteristics of the production process that produce chips with an appealing texture. The following data on x = frying time (s) and y = moisture content (%) appeared in the article "Thermal and Physical Properties of Tortilla Chips as a Function of Frying Time" (*J. Food Process. Preserv.* 1995: 175–189).

x	5	10	15	20	25	30	45	60
y	16.3	9.7	8.1	4.2	3.4	2.9	1.9	1.3

    a. Construct a scatterplot of the data and comment.

    b. Construct a scatterplot of the pairs $(\ln(x), \ln(y))$ (i.e., transform both x and y by logs) and comment.

    c. Consider the *multiplicative* power model $Y = \alpha x^{\beta}\varepsilon$. What does this model

imply about the relationship between $y' = \ln(y)$ and $x' = \ln(x)$ (assuming that ε has a lognormal distribution)?

d. Obtain a prediction interval for moisture content when frying time is 25 s. [*Hint*: first carry out a simple linear regression of y' on x' and calculate an appropriate prediction interval.]

135. Forest growth and decline phenomena throughout the world have attracted considerable public and scientific interest. The article "Relationships Among Crown Condition, Growth, and Stand Nutrition in Seven Northern Vermont Sugarbushes" (*Canad. J. Forest Res.* 1995: 386–397) included a scatterplot of y = mean crown dieback (%), one indicator of growth retardation, and x = soil pH (higher pH corresponds to less acidic soil), from which the following observations were taken:

x	3.3	3.4	3.4	3.5	3.6	3.6	3.7	3.7	3.8	3.8
y	7.3	10.8	13.1	10.4	5.8	9.3	12.4	14.9	11.2	8.0

x	3.9	4.0	4.1	4.2	4.3	4.4	4.5	5.0	5.1
y	6.6	10.0	9.2	12.4	2.3	4.3	3.0	1.6	1.0

a. Construct a scatterplot of the data. What model is suggested by the plot?

b. Use a statistical software package to fit the model suggested in (a) and test its utility.

c. Use the software package to obtain a prediction interval for crown dieback when soil pH is 4.0, and also a confidence interval for expected crown dieback in situations where the soil pH is 4.0. How do these two intervals compare to each other? Is this result consistent with what you learned in simple linear regression? Explain.

d. Use the software package to obtain a PI and CI when $x = 3.4$. How do these intervals compare to the corresponding intervals obtained in (c)? Is this result consistent with what you learned in simple linear regression? Explain.

136. The article "Validation of the Rockport Fitness Walking Test in College Males and Females" (*Res. Q. Exerc. Sport* 1994: 152–158) recommended the following estimated regression equation for relating y = VO_2max (L/min, a measure of cardiorespiratory fitness) to the predictors x_1 = gender (female = 0, male = 1), x_2 = weight (lb), x_3 = 1-mile walk time (min), and x_4 = heart rate at the end of the walk (beats/min):

$$y = 3.5959 + .65661x_1 + .0096x_2 - .0996x_3 - .0080x_4$$

a. How would you interpret the estimated coefficient $-.0996$?

b. How would you interpret the estimated coefficient $.6566$?

c. Suppose that an observation made on a male whose weight was 170 lb, walk time was 11 min, and heart rate was 140 beats/min resulted in VO_2max = 3.15. What would you have predicted for VO_2max in this situation, and what is the value of the corresponding residual?

d. Using SSE = 30.1033 and SST = 102.3922, what proportion of observed variation in VO_2max can be attributed to the model relationship?

e. Assuming a sample size of $n = 20$, carry out a test of hypotheses to decide whether the chosen model specifies a useful relationship between VO_2max and at least one of the predictors.

137. Investigators carried out a study to see how various characteristics of concrete are influenced by x_1 = % limestone powder and x_2 = water–cement ratio, resulting in the accompanying data ("Durability of Concrete with Addition of Limestone Powder," *Mag. Concr. Res.* 1996: 131–137).

x_1	x_2	28-day comp str. (MPa)	Adsorbability (%)
21	.65	33.55	8.42
21	.55	47.55	6.26
7	.65	35.00	6.74
7	.55	35.90	6.59
28	.60	40.90	7.28
0	.60	39.10	6.90
14	.70	31.55	10.80
14	.50	48.00	5.63
14	.60	42.30	7.43

a. Consider first compressive strength as the dependent variable y. Fit a first-order model, and determine R_a^2.

b. Determine the adjusted R^2 value for a model including the interaction term and also for the complete second-order model. Of the three models in parts (a)–(b), which seems preferable?

c. Use the "best" model from part (b) to predict compressive strength when % limestone = 14 and water–cement ratio = .60.

d. Repeat parts (a)–(b) with adsorbability as the response variable. That is, fit three models: the first-order model, one with first-order terms plus an interaction, and the complete second-order model.

138. A sample of $n = 20$ companies was selected, and the values of y = stock price and $k = 15$ predictor variables (such as quarterly dividend, previous year's earnings, and debt ratio) were determined. When the multiple regression model using these 15 predictors was fit to the data, $R^2 = .90$ resulted.

a. Does the model appear to specify a useful relationship between y and the predictor variables? Carry out a test using significance level .05. [*Hint:* The F critical value for 15 numerator and 4 denominator df is 5.86.]

b. Based on the result of part (a), does a high R^2 value by itself imply that a model is useful? Under what circumstances might you be suspicious of a model with a high R^2 value?

c. With n and k as given previously, how large would R^2 have to be for the model to be judged useful at the .05 level of significance?

Chi-Squared Tests

13

Introduction

In the simplest type of situation considered in this chapter, each observation in a sample is classified as belonging to one of a finite number of categories—for example, blood type could be one of the four categories O, A, B, or AB. With p_i denoting the probability that any particular observation belongs in category i, we wish to test a null hypothesis that completely specifies the values of all the p_i's (such as H_0: $p_1 = .45$, $p_2 = .35$, $p_3 = .15$, $p_4 = .05$). Other times, the null hypothesis specifies that the p_i's depend on some smaller number of parameters without specifying the values of these parameters; the values of any unspecified parameters must then be estimated from the sample data. In either case, the test statistic will be a measure of the discrepancy between the observed numbers in the categories and the expected numbers when H_0 is true. This method, called a *chi-squared test* and presented in Section 13.1, can also be applied to test the null hypothesis that the sample comes from a particular probability distribution.

Chi-squared tests for two different situations are presented in Section 13.2. In the first, the null hypothesis states that the p_i's are the same for several different populations. The second type of situation involves taking a sample from a single population and cross-classifying each individual with respect to two different categorical factors (such as religious preference and political party registration). The null hypothesis in this situation is that the two factors are independent within the population.

13.1 Goodness-of-Fit Tests

Recall that a binomial experiment consists of n independent trials in which each trial can result in one of two possible outcomes, S (for success) and F (for failure). The probability of success is assumed to be constant from trial to trial, and n is fixed at the outset of the experiment. A **multinomial experiment** generalizes the binomial experiment by allowing each trial to result in one of k possible outcomes, where $k \geq 2$. For example, suppose a store accepts three different types of credit cards: Visa, MasterCard, and American Express. A multinomial experiment would result from observing the type of credit card used—Visa, MC, or Amex—by each of 50 randomly selected customers who pay with a credit card.

DEFINITION   A **multinomial experiment** satisfies the following conditions:

1. The experiment consists of a sequence of n trials, where n is fixed in advance of the experiment.
2. Each trial can result in one of the same k possible outcomes (also called categories).
3. The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.
4. The probability that a trial results in category i is p_i, which remains constant from trial to trial.

The parameters p_1, …, p_k must of course satisfy $p_i \geq 0$ and $\sum p_i = 1$.

If the experiment consists of selecting n individuals or objects from a population and categorizing each one, then p_i is interpreted as the proportion of the population falling in the ith category; such an experiment will be approximately multinomial provided that n is much smaller than the population size. In the aforementioned example, $k = 3$ (number of categories = number of credit cards accepted), $n = 50$ (number of trials = number of customers), and p_i denotes the proportion of *all* credit card purchases made with type i (1 = Visa, 2 = MC, 3 = Amex).

The null hypothesis of interest at this point will specify the value of each p_i. For example, suppose the store manager believes 50% of all credit card customers use Visa, 30% MasterCard, and the remaining 20% American Express. This belief can be expressed as the assertion

$$H_0: p_1 = .5, \ p_2 = .3, \ p_3 = .2$$

The alternative hypothesis will state that H_0 is not true—i.e., that at least one of the p_i's has a value different from that asserted by H_0 (in which case at least two must be different, since they sum to 1). The symbol p_{i0} will represent the value of p_i claimed by the null hypothesis. In the example just given, $p_{10} = .5$, $p_{20} = .3$, and $p_{30} = .2$. (The symbol p_{10} is read "p one naught" and not "p ten.")

Before the multinomial experiment is performed, the number of trials that will result in the ith category ($i = 1, 2, \ldots,$ or k) is a random variable—just as the number of successes and the number of failures in a binomial experiment are random variables. This random variable will be denoted by N_i and its observed value by n_i. Since each trial results in exactly one of the k categories, $\sum N_i = n$, and the same is true of the n_i's. As an example, an experiment with $n = 50$ and $k = 3$ might yield $N_1 = 22$, $N_2 = 13$, and $N_3 = 15$. The N_i's (or n_i's) are called the **observed counts**.

When the null hypothesis $H_0: p_1 = p_{10},\ldots,p_k = p_{k0}$ is true, the expected number of trials resulting in category i is

$$E(N_i) = (\text{total number of trials}) \, (\text{hypothesized probability of category } i) = np_{i0}$$

These are the **expected counts** under H_0. For the case $H_0: p_1 = .5$, $p_2 = .3$, $p_3 = .2$ and $n = 50$, $E(N_1) = 25$, $E(N_2) = 15$, and $E(N_3) = 10$ when H_0 is true. The expected counts, like the observed counts, sum to n. It is customary to display both sets of counts, with the expected counts under H_0 in parentheses below the observed counts. The counts in the credit card situation under discussion would be displayed in tabular format as

Credit card	Visa	MC	Amex
Observed count	22	13	15
Expected count	(25)	(15)	(10)

A test procedure requires assessing the discrepancy between the observed and expected counts, with H_0 being rejected when the discrepancy is sufficiently large. The test statistic, originally proposed by Karl Pearson around 1900, is

$$\sum_{\text{all categories}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} = \sum_{i=1}^{k} \frac{(N_i - np_{i0})^2}{np_{i0}} \tag{13.1}$$

The numerator of each term in the sum is the squared difference between observed and expected counts. The more these differ within any particular category, the larger will be the contribution to the overall sum. The reason for including the denominator term will be explained shortly. Since the observed counts (the N_i's) are random variables, their values depend on the specific sample collected, and the test statistic (13.1) will vary in value from sample to sample. Larger values of the test statistic indicate a greater discrepancy between the observed and expected counts, making us more apt to reject H_0. The approximate sampling distribution of (13.1) is given in the following theorem.

PEARSON'S CHI-SQUARED THEOREM	When $H_0: p_1 = p_{10}, \ldots, p_k = p_{k0}$ is true, the statistic $$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_{i0})^2}{np_{i0}}$$ has approximately a chi-squared distribution with $k - 1$ df. This approximation is reasonable provided that $np_{i0} \geq 5$ for every i ($i = 1, 2, \ldots, k$).

The chi-squared distribution was introduced in Chapter 6 and used in Chapter 8 to obtain a confidence interval for the variance of a normal population. Recall that the chi-squared distribution has a single parameter v, called the number of degrees of freedom (df) of the distribution. Analogous to the critical value $t_{\alpha,v}$ for the t distribution, $\chi^2_{\alpha,v}$ is the value such that α of the area under the χ^2 curve with v df lies to the right of $\chi^2_{\alpha,v}$ (see Figure 13.1). Selected values of $\chi^2_{\alpha,v}$ are given in Appendix Table A.5. Notice that, unlike a z or t curve, the chi-squared distribution is positively skewed and only takes on nonnegative values.
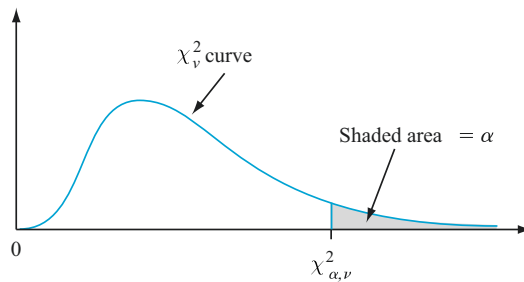
Figure 13.1 A critical value for a chi-squared distribution

The fact that df $= k - 1$ in the preceding theorem is a consequence of the restriction $\sum N_i = n$: although there are k observed counts, once any $k - 1$ are known, the remaining one is uniquely determined. That is, there are only $k - 1$ "freely determined" cell counts, and thus $k - 1$ df.

CHI-SQUARED GOODNESS-OF-FIT TEST

Null hypothesis:	$H_0: p_1 = p_{10}, \ldots, p_k = p_{k0}$
Alternative hypothesis:	H_a: at least one p_i does not equal p_{i0}
Test statistic value:	$\chi^2 = \sum\limits_{i=1}^{k} \dfrac{(n_i - np_{i0})^2}{np_{i0}}$

Rejection Region for Level α Test   P-value Calculation

$\chi^2 \geq \chi^2_{\alpha, k-1}$ area under χ^2_{k-1} curve to the right of the calculated χ^2

The term "goodness-of-fit" refers to the idea that we wish to see how well the observed counts of a categorical variable "fit" a set of hypothesized population proportions. Appendix Table A.5 provides upper-tail critical values at five α levels for each different v. Because this is not sufficiently granular for accurate P-value information, we have also included Appendix Table A.10, analogous to Table A.7, that facilitates making more precise P-value statements.

Example 13.1 If we focus on two different characteristics of an organism, each controlled by a single gene, and cross a pure strain having genotype AABB with a pure strain having genotype aabb (capital letters denoting dominant alleles and small letters recessive alleles), the resulting genotype will be AaBb. If these first-generation organisms are then crossed among themselves (a *dihybrid* cross), there will be four phenotypes depending on whether a dominant allele of either type is present. Mendel's laws of inheritance imply that these four phenotypes should have probabilities 9/16, 3/16, 3/16, and 1/16 of arising in any given dihybrid cross.

The article "Inheritance of Fruit Attributes in Chili Pepper" (*Indian J. Hort.* 2019: 86–93) reports the phenotype counts resulting from a dihybrid cross of two chili pepper varietals popular in India (WBC-Sel-5 and GVC-101). There are $k = 4$ categories corresponding to the four possible fruit-bearing phenotypes, with the null hypothesis being

$$H_0: p_1 = \frac{9}{16}, \quad p_2 = \frac{3}{16}, \quad p_3 = \frac{3}{16}, \quad p_4 = \frac{1}{16}$$

Since the total sample size was $n = 63$, the expected cell counts are 63(9/16) = 35.44, 63(3/16) = 11.81, 11.81, and 63(1/16) = 3.94. Observed and expected counts are given in Table 13.1. (Although one expected count is slightly less than 5, Pearson's chi-squared theorem should still apply reasonably well to this scenario.)

Table 13.1 Observed and expected cell counts for Example 13.1

$i = 1$ Single, drooping	$i = 2$ Single, erect	$i = 3$ Cluster, drooping	$i = 4$ Cluster, erect
32	8	16	7
(35.44)	(11.81)	(11.81)	(3.94)

The contribution to the χ^2 test statistic value from the first cell is

$$\frac{(n_1 - np_{10})^2}{np_{10}} = \frac{(32 - 35.44)^2}{35.44} = .333$$

Cells 2, 3, and 4 contribute 1.230, 1.484, and 2.382, respectively, so $\chi^2 = .333 + 1.230 + 1.484 + 2.382 = 5.43$. The expected value of χ^2 under H_0 is roughly $v = k - 1 = 3$ and the standard deviation is approximately $\sqrt{2v} = 2.5$. So our test statistic value is only about one standard deviation larger than what we'd expect if the null hypothesis was true, seemingly not highly contradictory to H_0.

More formally, a test with significance level .10 at 3 df requires $\chi^2_{.10,3}$, the number in the 3 df row and .10 column of Appendix Table A.5. This critical value is 6.251. Since $5.43 < 6.251$, H_0 cannot be rejected even at this rather large level of significance. (The $v = 3$ column of Appendix Table A.10 confirms that P-value $> .10$; software provides a P-value of .143.) The data is reasonably consistent with Mendel's laws. ∎

Why not simply use $\sum (N_i - np_{i0})^2$ as the test statistic, rather than the more complicated statistic (13.1)? Suppose, for example, that $np_{10} = 100$ and $np_{20} = 10$. Then if $n_1 = 95$ and $n_2 = 5$, the two categories contribute the same squared deviations to $\sum(n_i - np_{i0})^2$. Yet n_1 is only 5% less than what would be expected when H_0 is true, whereas n_2 is 50% less. To take *relative* magnitudes of the deviations into account, we divide each squared deviation by the corresponding expected count and then combine.

χ^2 for Completely Specified Probability Distributions

Frequently researchers wish to determine whether observed data is consistent with a particular probability distribution. When the distribution and all of its parameters are completely specified, Pearson's chi-squared test can be applied to this scenario. Later in this section, we examine the case when the parameters must be estimated from the available data.

Example 13.2 In a famous genetics article ("The Progeny in Generations F_{12} to F_{17} of a Cross Between a Yellow-Wrinkled and a Green-Round Seeded Pea," *J. Genet.* 1923: 255–331), the early statistician G. U. Yule analyzed data resulting from crossing garden peas. The dominant alleles in the experiment were Y = yellow color and R = round shape, resulting in the double dominant YR. Yule examined 269 four-seed pods resulting from a dihybrid cross and counted the number of YR seeds in each pod.

Let X denote the number of YR's in a randomly selected peapod, so possible X values are 0, 1, 2, 3, 4. Based on the discussion in Example 13.1, the Mendelian laws are operative and genotypes of individual seeds within a pod are independent of one another. Thus X has a $\mathrm{Bin}(4, \frac{9}{16})$ distribution. If for $i = 1, 2, 3, 4, 5$ we define $p_i = P(X = i - 1)$, then we wish to test $H_0: p_1 = p_{10}, \ldots, p_5 = p_{50}$, where

$$p_{i0} = P(i - 1 \text{ YR's among 4 seeds when } H_0 \text{ is true})$$

$$= \binom{4}{i-1} \left(\frac{9}{16}\right)^{i-1} \left(1 - \frac{9}{16}\right)^{4-(i-1)} \quad i = 1, 2, 3, 4, 5$$

Substituting into this binomial pmf gives hypothesized probabilities .0366, .1884, .3634, .3115, and .1001. Yule's data and the expected cell counts $np_{i0} = 269p_{i0}$ are in Table 13.2.

Table 13.2 Observed and expected cell counts for Example 13.2

$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$
16	45	100	82	26
(9.86)	(50.68)	(97.75)	(83.78)	(26.93)

The test statistic value is

$$\chi^2 = \frac{(16 - 9.86)^2}{9.86} + \cdots + \frac{(26 - 26.93)^2}{26.93} = 3.823 + \cdots + .032 = 4.582$$

Since $4.582 < \chi^2_{.01,k-1} = \chi^2_{.01,4} = 13.277$, H_0 is not rejected at level .01. In fact, software provides a P-value of .333, so H_0 should not be rejected at any reasonable significance level. ∎

The χ^2 test can also be used to test whether a sample comes from a specific underlying continuous distribution. Let X denote the variable being sampled and suppose the hypothesized pdf of X is $f_0(x)$. As in the construction of a frequency distribution in Chapter 1, subdivide the measurement scale of X into k disjoint intervals $(-\infty, a_1)$, $[a_1, a_2)$, …, $[a_{k-1}, \infty)$. The cell probabilities for $i = 2, …, k - 1$ specified by H_0 are then

$$p_{i0} = P(a_{i-1} \leq X < a_i) = \int_{a_{i-1}}^{a_i} f_0(x)dx$$

and similarly for the two extreme intervals. The intervals should be chosen so that $np_{i0} \geq 5$ for $i = 1$, …, k; often they are selected so that the p_{i0}'s are equal. Once the p_{i0}'s are calculated, the underlying distribution is, in a sense, irrelevant—the chi-squared test will determine whether data is consistent with *any* probability distribution that places probability p_{i0} on the ith specified interval.

Example 13.3 To see whether time of birth is uniformly distributed throughout a 24-hour day, we can divide a day into one-hour periods starting at midnight ($k = 24$ intervals). The null hypothesis states that $f(x)$ is the uniform pdf on the interval [0, 24], so that $p_{i0} = 1/24$ for all i. A random sample of 1000 births from the CDC's 2018 Natality Public Use File resulted in cell counts of 34 (midnight to 12:59 a.m.), 28, 37, 29, 31, 28, 32, 38, 73, 50, 43, 52, 58, 58, 46, 43, 51, 35, 46, 32, 53, 31, 35, and 37 (11:00 p.m. to 11:59 p.m.). Each expected "cell count" is $1000 \cdot 1/24 = 41.67$, and the resulting value of χ^2 is 74.432. Since $\chi^2_{.01,23} = 41.637$, the computed value is highly significant, and the null hypothesis is resoundingly rejected. In particular, babies were far more likely to be born in the 8:00 a.m.–8:59 a.m. window (observed count = 73) that in any other hour of the day. ∎

Example 13.4 The developers of a new online tax program want it to satisfy three criteria: (1) actual time to complete a tax return is normally distributed; (2) the mean completion time is 90 min; (3) 90% of all users will finish their tax returns within 120 min (2 h). A pilot test of the program will utilize 120 volunteers, resulting in 120 completion time observations. This data will be used to test whether the performance criteria are met, using a chi-squared test with $k = 8$ intervals.

Calculating normal probabilities requires both μ and σ. The target value $\mu = 90$ min is given; the 90th percentile of a normal distribution is $\mu + 1.28\sigma$, and the criterion $\mu + 1.28\sigma = 120$ min implies $\sigma = 23.44$ min. To divide the *standard* normal scale into eight equally likely intervals, we look for

the .125 quantile, .25 quantile, etc., in the z table. From Table A.3 these values, which form the boundary points of our intervals, have z-scores equal to

$$-1.15 \quad -.675 \quad -.32 \quad 0 \quad .32 \quad .675 \quad 1.15$$

For $\mu = 90$ and $\sigma = 23.44$, these boundary points become

$$63.04 \quad 74.18 \quad 82.50 \quad 90.00 \quad 97.50 \quad 105.82 \quad 116.96$$

(Completion times obviously cannot be negative. The area to the left of $x = 0$ under this curve is negligible, so this issue is not of concern here.) If we define p_i = the probability a randomly selected completion time falls in the ith interval defined by the above boundary points, then the goal is to test $H_0: p_1 = .125, \ldots, p_8 = .125$.

Suppose the observed counts are as shown in the accompanying table; the expected count for each interval is $np_{i0} = (120)(.125) = 15$.

Lower endpoint of interval	0	63.04	74.18	82.50	90.00	97.50	105.82	116.96
Observed count	21	17	12	16	10	15	19	10
Expected count	(15)	(15)	(15)	(15)	(15)	(15)	(15)	(15)

The resulting test statistic value is

$$\chi^2 = \frac{(21 - 15)^2}{15} + \cdots + \frac{(10 - 15)^2}{15} = 7.73$$

The corresponding P-value (using Table A.10 at df = $8 - 1 = 7$) exceeds .100; statistical software gives P-value = .357. Thus we have no reason to reject H_0; the 120 observations are consistent with a $N(90, 23.44)$ population distribution, as desired. ∎

Goodness-of-Fit Tests for Composite Hypotheses

The goodness-of-fit test based on Pearson's chi-squared theorem involves a *simple* null hypothesis, in the sense that each p_{i0} is a specified number, so that the expected cell counts when H_0 is true are completely determined. But in some situations, H_0 states only that the p_i's are functions of other parameters $\theta_1, \ldots, \theta_m$ without specifying the values of these θ_j's.

For example, a population may be in equilibrium with respect to proportions of the three genotypes AA, Aa, and aa. With p_1, p_2, and p_3 denoting these proportions (probabilities), one may wish to test $H_0: p_1 = \theta^2, p_2 = 2\theta(1 - \theta), p_3 = (1 - \theta)^2$, where θ represents the proportion of gene A in the population. This hypothesis is *composite* because knowing that H_0 is true does not uniquely determine the cell probabilities and expected cell counts, but only their general form. More generally, the null hypothesis now states that each p_i is a function of a small number of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ with the θ_j's otherwise unspecified:

$$H_0: p_1 = \pi_1(\boldsymbol{\theta}), \ldots, p_k = \pi_k(\boldsymbol{\theta})$$

$$H_a: \text{the hypothesis } H_0 \text{ is not true}$$

(13.2)

In the genotype example, $m = 1$ (there is only one θ), $\pi_1(\theta) = \theta^2$, $\pi_2(\theta) = 2\theta(1-\theta)$, and $\pi_3(\theta) = (1-\theta)^2$. To carry out a χ^2 test, the unknown θ's must first be estimated.

In the case $k = 2$, there is really only a single rv, N_1 (since $N_1 + N_2 = n$), which has a binomial distribution. The joint probability that $N_1 = n_1$ and $N_2 = n_2$ is then

$$P(N_1 = n_1, N_2 = n_2) = \binom{n}{n_1} p_1^{n_1} p_2^{n_2} \propto p_1^{n_1} p_2^{n_2}$$

where $p_1 + p_2 = 1$ and $n_1 + n_2 = n$. For general k, the joint distribution of $N_1, \ldots, N_k$ is the multinomial distribution (Section 5.1) with

$$P(N_1 = n_1, \ldots, N_k = n_k) \propto p_1^{n_1} p_2^{n_2} \cdots \cdots p_k^{n_k}$$

which, when H_0 is true, becomes

$$P(N_1 = n_1, \ldots, N_k = n_k) \propto [\pi_1(\boldsymbol{\theta})]^{n_1} \cdots \cdots [\pi_k(\boldsymbol{\theta})]^{n_k} \tag{13.3}$$

METHOD OF MULTINOMIAL ESTIMATION

Let n_1, n_2, $\ldots$, n_k denote the observed values of N_1, $\ldots$, N_k. Then $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are those values of the θ_j's that maximize Expression (13.3), that is, the maximum likelihood estimates *with respect to the multinomial model*.

Example 13.5 In humans there is a blood group, the MN group, that is composed of individuals having one of three blood types: M, MN, or N. Type is determined by two alleles, and there is no dominance, so the three possible genotypes give rise to three phenotypes. A population consisting of individuals in the MN group is in equilibrium if

$$P(\text{M}) = p_1 = \theta^2 \quad P(\text{MN}) = p_2 = 2\theta(1-\theta) \quad P(\text{N}) = p_3 = (1-\theta)^2$$

for some θ. Suppose a sample from such a population yielded the results shown in Table 13.3.

Table 13.3 Observed counts for Example 13.5

Type	M	MN	M	
Observed count	125	225	150	$n = 500$

Then Expression (13.3) becomes

$$[\pi_1(\theta)]^{n_1} [\pi_2(\theta)]^{n_2} [\pi_3(\theta)]^{n_3} = [\theta^2]^{n_1} [2\theta(1-\theta)]^{n_2} [(1-\theta)^2]^{n_3}$$
$$= 2^{n_2} \cdot \theta^{2n_1 + n_2} \cdot (1-\theta)^{n_2 + 2n_3}$$

Maximizing this with respect to θ (or, equivalently, maximizing the natural logarithm of this quantity, which is easier to differentiate) yields

$$\hat{\theta} = \frac{2n_1 + n_2}{[(2n_1 + n_2) + (n_2 + 2n_3)]} = \frac{2n_1 + n_2}{2n}$$

With $n_1 = 125$ and $n_2 = 225$, $\hat{\theta} = 475/1000 = .475$.   ∎

Once $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ has been estimated by $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$, the *estimated* expected cell count for the ith category is $n\hat{p}_i = n\pi_i(\hat{\boldsymbol{\theta}})$. These are now used in place of the np_{i0}'s in Expression (13.1) to specify a χ^2 statistic. The following theorem was proved by R. A. Fisher in 1924 as a generalization of Pearson's chi-squared test.

FISHER'S CHI-SQUARED THEOREM

Under general "regularity" conditions on $\theta_1, \ldots, \theta_m$ and the $\pi_i(\boldsymbol{\theta})$'s, if $\theta_1, \ldots, \theta_m$ are estimated by maximizing the multinomial expression (13.3), the rv

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n\hat{P}_i)^2}{n\hat{P}_i} = \sum_{i=1}^{k} \frac{[N_i - n\pi_i(\hat{\boldsymbol{\theta}})]^2}{n\pi_i(\hat{\boldsymbol{\theta}})}$$

has approximately a chi-squared distribution with $k - 1 - m$ df when H_0 of (13.2) is true. An approximately level α test of H_0 versus H_a is then to reject H_0 if $\chi^2 \geq \chi^2_{\alpha, k-1-m}$.

In practice, the test can be used if $n\pi_i(\hat{\boldsymbol{\theta}}) \geq 5$ for every i.

Notice that the number of degrees of freedom is reduced by the number of θ_j's estimated.

Example 13.6 (Example 13.5 continued) With $\hat{\theta} = .475$ and $n = 500$, the estimated expected cell counts are

$$n\hat{p}_1 = n\pi_1(\hat{\theta}) = n \cdot \hat{\theta}^2 = 500 \cdot (.475)^2 = 112.81$$
$$n\hat{p}_2 = n\pi_2(\hat{\theta}) = n \cdot 2\hat{\theta}(1 - \hat{\theta}) = 500 \cdot 2(.475)(.525) = 249.38$$
$$n\hat{p}_3 = n\pi_3(\hat{\theta}) = n \cdot (1 - \hat{\theta})^2 = 500 \cdot (.525)^2 = 137.81$$

Notice that these estimated expected counts sum to $n = 500$. Then

$$\chi^2 = \frac{(125 - 112.81)^2}{112.81} + \frac{(225 - 249.38)^2}{249.38} + \frac{(150 - 137.81)^2}{137.81} = 4.78$$

Since $\chi^2_{.05, k-1-m} = \chi^2_{.05, 3-1-1} = \chi^2_{.05, 1} = 3.843$ and $4.78 \geq 3.843$, H_0 is rejected at the .05 significance level (software provides P-value $= .029$). Therefore, the data does not conform to the proposed equilibrium model. Even using the θ estimate that "fits" the data best, the expected counts under the null model are too discordant with the observed counts.   ∎

Example 13.7 Consider a series of games between two teams, I and II, that terminates as soon as one team has won four games (with no possibility of a tie)—this is the "best of seven" format used for many professional league play-offs. A simple probability model for such a series assumes that

outcomes of successive games are independent and that the probability of team I winning any particular game is a constant θ. We arbitrarily designate team I the better team, so that $\theta \geq .5$. Any particular series can terminate after 4, 5, 6, or 7 games. Let $\pi_1(\theta)$, $\pi_2(\theta)$, $\pi_3(\theta)$, $\pi_4(\theta)$ denote the probability of termination in 4, 5, 6, and 7 games, respectively. Then

$$\begin{aligned}
\pi_1(\theta) &= P(\text{I wins in 4 games}) + P(\text{II wins in 4 games}) \\
&= \theta^4 + (1-\theta)^4 \\
\pi_2(\theta) &= P(\text{I wins 3 of the first 4 and the fifth}) \\
&\quad + P(\text{I loses 3 of the first 4 and the fifth}) \\
&= \binom{4}{3}\theta^3(1-\theta)\cdot\theta + \binom{4}{1}\theta(1-\theta)^3\cdot(1-\theta) \\
&= 4\theta(1-\theta)\left[\theta^3 + (1-\theta)^3\right] \\
\pi_3(\theta) &= 10\theta^2(1-\theta)^2[\theta^2 + (1-\theta)^2] \\
\pi_4(\theta) &= 20\theta^3(1-\theta)^3
\end{aligned}$$

The *Mathematics Magazine* article "Seven-Game Series in Sports" by Groeneveld and Meeden tested the fit of this model to results of National Hockey League playoffs during the period 1943–1967, when league membership was stable. The data appears in Table 13.4.

Table 13.4   Observed and expected counts for the simple model

$i = 1$ 4 games	$i = 2$ 5 games	$i = 3$ 6 games	$i = 4$ 7 games	
15	26	24	18	$n = 83$
(16.351)	(24.153)	(23.240)	(19.256)	

The estimated expected cell counts are $83\pi_i(\hat{\theta})$, where $\hat{\theta}$ is the value of θ that maximizes the multinomial expression

$$\left\{\theta^4 + (1-\theta)^4\right\}^{15}\cdot\left\{4\theta(1-\theta)\left[\theta^3 + (1-\theta)^3\right]\right\}^{26}\cdot\left\{10\theta^2(1-\theta)^2\left[\theta^2 + (1-\theta)^2\right]\right\}^{24}\cdot\left\{20\theta^3(1-\theta)^3\right\}^{18}$$

$$(13.4)$$

Standard calculus methods fail to yield a nice formula for the maximizing value $\hat{\theta}$, so it must be computed using numerical methods. The result is $\hat{\theta} = .654$, from which $\pi_i(\hat{\theta})$ and the estimated expected cell counts in Table 13.4 were computed. The resulting test statistic value is $\chi^2 = .360$, much lower than the critical value $\chi^2_{.10,k-1-m} = \chi^2_{.10,4-1-1} = \chi^2_{.10,2} = 4.605$. There is thus no reason to reject the simple model as applied to NHL playoff series, at least for that early era.

The cited article also considered World Series data for the period 1903–1973. For the preceding model, $\chi^2 = 5.97 \geq 4.605$, so the model does not seem appropriate. The suggested reason for this is that for this simple model it can be shown that

$$P(\text{series lasts exactly six games} \mid \text{series lasts at least six games}) \geq .5, \qquad (13.5)$$

whereas of the 38 best-of-seven series that actually lasted at least six games, only 13 lasted exactly six. The following alternative model is then introduced:

$$\pi_1(\theta_1, \theta_2) = \theta_1^4 + (1 - \theta_1)^4 \qquad\qquad \pi_3(\theta_1, \theta_2) = 10\theta_1^2(1 - \theta_1)^2\theta_2$$
$$\pi_2(\theta_1, \theta_2) = 4\theta_1(1 - \theta_1)[\theta_1^3 + (1 - \theta_1)^3] \quad \pi_4(\theta_1, \theta_2) = 10\theta_1^2(1 - \theta_1)^2(1 - \theta_2)$$

The first two π_i's are identical to the simple model, while θ_2 is the conditional probability of (13.5), which can now be any number between zero and one. The values of $\hat{\theta}_1$ and $\hat{\theta}_2$ that maximize the multinomial expression analogous to (13.4) are determined numerically as $\hat{\theta}_1 = .614$ and $\hat{\theta}_2 = .342$. A summary appears in Table 13.5, and $\chi^2 = .384$. Two parameters are estimated, so df $= k - 1 - m = 4 - 1 - 2 = 1$ and $.384 < \chi^2_{.10,1} = 2.706$, indicating a good fit of the data to this new model.

Table 13.5 Observed and expected counts for the more complex model

4 games	5 games	6 games	7 games
12	16	13	25
(10.85)	(18.08)	(12.68)	(24.39)

∎

One of the regularity conditions on the θ_j's in Fisher's theorem is that they be functionally independent of one another. That is, no single θ_j can be determined from the values of other θ_j's, so that m is the number of functionally independent parameters estimated. A general rule for degrees of freedom in a chi-squared test is the following.

GENERAL χ^2 DF RULE
$$\chi^2 \text{df} = \begin{pmatrix} \text{number of freely} \\ \text{determined cell counts} \end{pmatrix} - \begin{pmatrix} \text{number of independent} \\ \text{parameters estimated} \end{pmatrix}$$

This rule will be used in connection with several different chi-squared tests in the next section.

χ^2 for Probability Distributions with Parameter Values Unspecified

In Examples 13.2–13.4, we considered goodness-of-fit tests to assess whether quantitative data was consistent with a particular distribution, such as $\text{Bin}(4, \frac{9}{16})$ or $N(90, 23.44)$. Pearson's chi-squared theorem could be applied because all model parameters were completely specified. But quite often researchers wish to determine whether their data conforms to *any* member of a particular family—any Poisson distribution, any Weibull distribution, etc. To use the χ^2 test to see whether the distribution is Poisson, for example, the parameter μ must be estimated. In addition, because there are actually an infinite number of possible values of a Poisson variable, these values must be grouped so that there are a finite number of cells.

Example 13.8 Table 13.6 presents count data on X = the number of egg pouches produced by *B. alexandrina* snails that were subjected to both parasitic infection and drought stress (meant to simulate the effects of climate change), as reported in the article "One Stimulus, Two Responses: Host and Parasite Life-History Variation in Response to Environmental Stress" (*Evolution* 2016: 2640–2646).

Table 13.6 Observed counts for Example 13.8

Cell	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
No. of egg pouches	0	1	2	3	≥ 4
Observed count	44	2	5	1	9

Denoting the sample values by $x_1, \ldots, x_{61}$, 44 of the x_i's were 0, two were 1, and so on. The nine observed counts in the last cell were 4, 5, 6, 6, 7, 11, 13, 15, and 17, but these have been collapsed into a single "≥ 4" category in order to ensure that all expected counts will be at least 5.

The authors considered fitting a Poisson distribution to the data; let μ denote the Poisson parameter. The estimate of μ required for Fisher's χ^2 procedure is obtained by maximizing the multinomial expression (13.3). The cell probabilities are

$$\pi_i(\mu) = P(X = i - 1) = \frac{e^{-\mu}\mu^{i-1}}{(i-1)!} \quad i = 1, 2, 3, 4$$

$$\pi_5(\mu) = P(X = 4) = 1 - \sum_{x=0}^{3} \frac{e^{-\mu}\mu^x}{x!}$$

so the right-hand side of (13.3) becomes

$$\left[\frac{e^{-\mu}\mu^0}{0!}\right]^{44} \left[\frac{e^{-\mu}\mu^1}{1!}\right]^2 \left[\frac{e^{-\mu}\mu^2}{2!}\right]^5 \left[\frac{e^{-\mu}\mu^3}{3!}\right]^1 \left[1 - \sum_{x=0}^{3} \frac{e^{-\mu}\mu^x}{x!}\right]^9 \tag{13.6}$$

There is no nice formula for the maximizing value of μ in Expression (13.6), so it must be obtained numerically. ∎

While maximizing Expression (13.6) with respect to μ is challenging, there is an alternative way to estimate μ: *apply the method of maximum likelihood from Chapter 7 to the full sample* $X_1, \ldots, X_n$. Because parameter estimates are usually much more difficult to compute from the multinomial likelihood function (13.3) than from the full-sample likelihood, they are often computed using this latter method. Using Fisher's critical value $\chi^2_{\alpha, k-1-m}$ then results in an *approximate* level α test.

Example 13.9 (Example 13.8 continued) The likelihood of the observed sample $x_1, \ldots, x_{61}$ under a Poisson(μ) model is

$$L(\mu) = p(x_1; \mu) \cdots p(x_{61}; \mu) = \frac{e^{-\mu}\mu^{x_1}}{x_1!} \cdot \cdots \cdot \frac{e^{-\mu}\mu^{x_{61}}}{x_{61}!} = \frac{e^{-61\mu}\mu^{\Sigma x_i}}{x_1! \cdot \cdots \cdot x_{61}!} = \frac{e^{-61\mu}\mu^{99}}{x_1! \cdot \cdots \cdot x_{61}!}$$

The value of μ for which this is maximized—i.e., the maximum likelihood estimate of μ—is $\hat{\mu} = \sum x_i / n = 99/61 = 1.623$. Using $\hat{\mu} = 1.623$, the estimated expected cell counts are computed from $n\pi_i(\hat{\mu})$, where $n = 61$. For example,

$$n\pi_1(\hat{\mu}) = 61 \cdot \frac{e^{-1.623}(1.623)^0}{0!} = (61)(.1973) = 12.036$$

Similarly, $n\pi_2(\hat{\mu}) = 19.534$, $n\pi_3(\hat{\mu}) = 15.852$, and $n\pi_4(\hat{\mu}) = 8.576$, from which the last count is $n\pi_5(\hat{\mu}) = 61 - [12.036 + \cdots + 8.576] = 5.002$. Notice that, as planned, all of the estimated expected cell counts are ≥ 5, as required for the accuracy of chi-squared tests. Then

$$\chi^2 = \frac{(44 - 12.036)^2}{12.036} + \cdots + \frac{(9 - 5.002)^2}{5.002} = 117.938$$

Since $m = 1$ and $k = 5$, at level .05 we need $\chi^2_{.05,5-1-1} = \chi^2_{.05,3} = 7.815$. Because $117.938 > 7.815$, we strongly reject H_0 at the .05 significance level (in fact, with such a ridiculously large test statistic value, H_0 is rejected at any reasonable α).

The largest contributor to the χ^2 statistic here is the number of 0's in the sample: 44 were observed, but under a Poisson model only 12.036 are expected. This excess of zeros often occurs with "count" data, and statisticians have developed *zero-inflated* versions of the Poisson and other distributions to accommodate this reality. ∎

When model parameters are estimated using full-sample maximum likelihood, it is known that the *true* level α critical value falls between $\chi^2_{\alpha,k-1-m}$ and $\chi^2_{\alpha,k-1}$. So, applying Fisher's chi-squared method to situations such as Example 13.9 will occasionally lead to incorrectly rejecting H_0, though in practice this is uncommon. Sometimes even the maximum likelihood estimates based on the full sample are quite difficult to compute. This is the case, for example, for the two-parameter generalized negative binomial distribution (Exercise 17). In such situations, method-of-moments estimates are often used and the resulting χ^2 value compared to $\chi^2_{\alpha,k-1-m}$, although it is not known to what extent the use of moments estimators affects the true critical value.

In theory, the chi-squared test can also be used to test whether a sample comes from a specified family of continuous distributions, such as the exponential family or the normal family. However, when the parameter values are not specified, the goodness-of-fit test is rarely used for this purpose. Instead, practitioners use one or more of the test procedures mentioned in connection with probability plots in Section 4.6, such as the Shapiro–Wilk or Anderson–Darling test. For example, the Shapiro–Wilk procedure tests the hypotheses

H_0: the population from which the sample was drawn is normal
H_a: H_0 is not true

A *P*-value is calculated based on a test statistic similar to the correlation coefficient associated with the points in a normal probability plot. These procedures are generally considered superior to the chi-squared tests of this section for continuous families; in particular, they do not rely on creating arbitrary class intervals.

More on the Goodness-of-Fit Test

When one or more expected counts are less than 5, the chi-squared distribution does not necessarily accurately approximate the sampling distribution of the test statistic (13.1). This can occur either because the sample size n is small or because one of the hypothesized proportions p_{i0} is small. If a larger sample is not available, a common practice is to sensibly "merge" some of the categories with small expected counts, so that the expected count for the merged category is larger. This might occur, for example, if we examined the political party affiliation of a sample of graduate students, categorized as Republican, Democrat, Libertarian, Green, Peace and Freedom, and Independent. The counts for the nonmajor parties might be quite small, in which case we could combine them to form, say, three categories: Republican, Democrat, and Other. The downside of this technique is that information has been discarded—two students belonging to different political parties (e.g., Libertarian and Green) are no longer distinguishable.

Although the chi-squared test was developed to handle situations in which $k > 2$, it can also be used when $k = 2$. The null hypothesis in this case can be stated as $H_0: p_1 = p_{10}$, since the relations $p_2 = 1 - p_1$ and $p_{20} = 1 - p_{10}$ make the inclusion of $p_2 = p_{20}$ in H_0 redundant. The alternative hypothesis is $H_a: p_1 \neq p_{10}$. These hypotheses can also be tested using a two-tailed one-proportion z test with test statistic

$$Z = \frac{(N_1/n) - p_{10}}{\sqrt{\dfrac{p_{10}(1 - p_{10})}{n}}} = \frac{\hat{P}_1 - p_{10}}{\sqrt{\dfrac{p_{10}p_{20}}{n}}}$$

In fact, the two test procedures are completely equivalent. This is because it can be shown that $Z^2 = \chi^2$ (see Exercise 12) and $z_{\alpha/2}^2 = \chi_{\alpha,1}^2$, so that $\chi^2 \geq \chi_{\alpha,1}^2$ if and only if $|Z| \geq z_{\alpha/2}$.[1] In other words, the two-tailed z test from Chapter 9 rejects H_0 if and only if the chi-squared goodness-of-fit test does. However, if the alternative hypothesis is either $H_a: p_1 > p_{10}$ or $H_a: p_1 < p_{10}$, the chi-squared test cannot be used. One must then revert to an upper- or lower-tailed z test.

As is the case with all test procedures, one must be careful not to confuse statistical significance with practical significance. A computed χ^2 that leads to the rejection of H_0 may be a result of a very large sample size rather than any practical differences between the hypothesized p_{i0}'s and true p_i's. Thus if $p_{10} = p_{20} = p_{30} = \frac{1}{3}$, but the true p_i's have values .330, .340, and .330, a large value of χ^2 is sure to arise with a sufficiently large n. Before rejecting H_0, the $\hat{p}_i$'s should be examined to see whether they suggest a model different from that of H_0 from a practical point of view.

Exercises: Section 13.1 (1–18)

1. What conclusion would be appropriate for an upper-tailed chi-squared test in each of the following situations?
   a. $\alpha = .05$, df = 4, $\chi^2 = 12.25$
   b. $\alpha = .01$, df = 3, $\chi^2 = 8.54$
   c. $\alpha = .10$, df = 2, $\chi^2 = 4.36$
   d. $\alpha = .01$, $k = 6$, $\chi^2 = 10.20$

2. Say as much as you can about the P-value for an upper-tailed chi-squared test in each of the following situations:
   a. $\chi^2 = 7.5$, df = 2
   b. $\chi^2 = 13.0$, df = 6
   c. $\chi^2 = 18.0$, df = 9
   d. $\chi^2 = 21.3$, $k = 5$
   e. $\chi^2 = 5.0$, $k = 4$

3. A statistics department at a large university maintains a tutoring center for students in its introductory service courses. The center has been staffed with the expectation that 40% of its clients would be from the business statistics

course, 30% from engineering statistics, 20% from the statistics course for social science students, and the other 10% from the course for agriculture students. A random sample of $n = 120$ clients revealed 52, 38, 21, and 9 from the four courses. Does this data suggest that the percentages on which staffing was based are not correct? State and test the relevant hypotheses using $\alpha = .05$.

4. It is hypothesized that when homing pigeons are disoriented in a certain manner, they will exhibit no preference for any direction of flight after takeoff (so that the direction X should be uniformly distributed on the interval from $0°$ to $360°$). To test this, 120 pigeons were disoriented, let loose, and the direction of flight of each was recorded; the resulting data follows. Use the chi-squared test at level .10 to see whether the data supports the hypothesis.

---

[1]The fact that $z_{\alpha/2}^2 = \chi_{\alpha,1}^2$ is a consequence of the relationship between the standard normal distribution and the chi-squared distribution with 1 df: if $Z \sim N(0, 1)$, then by definition Z^2 has a chi-squared distribution with $v = 1$. See Section 6.3.

Direction	0–<45°	45–<90°	90–<135°
Frequency	12	16	17

Direction	135–<180°	180–<225°	225–<270°
Frequency	15	13	20

Direction	270–<315°	315–<360°	
Frequency	17	10	

5. An information retrieval system has ten storage locations. Information has been stored with the expectation that the long-run proportion of requests for location i is given by the expression $p_i = (5.5 - |i - 5.5|)/30$. A sample of 200 retrieval requests gave the following frequencies for locations 1–10, respectively: 4, 15, 23, 25, 38, 31, 32, 14, 10, and 8. Use a chi-squared test at significance level .10 to decide whether the data is consistent with the a priori proportions (use the P-value approach).

6. The article "Racial Stereotypes in Children's Television Commercials" (*J. Adver. Res.* 2008) reported the following frequencies with which characters of different ethnicities appeared in recorded commercials aired on Philadelphia television stations.

Ethnicity	African–American	Asian	Caucasian	Hispanic
Frequency	57	11	330	6

Census data at the time reported the population proportions for these four ethnic groups was .177, .032, .734, and .057, respectively. Does the data suggest that the true proportions in commercials are different from the census proportions? Carry out a test of appropriate hypotheses using a significance level of .01.

7. A retail bookstore manager is re-evaluating weekday staffing by looking at recent sales. The accompanying table summarizes a sample of 92 recent weekday sales.

Weekday	Monday	Tuesday	Wednesday	Thursday	Friday
Number of sales	22	13	16	17	24

Assuming these sales are representative of all weekday sales at the bookstore, does the data indicate that such sales are *not* evenly distributed throughout the week?

8. *Benford's Law*, introduced in Chapter 3 Exercise 19, postulates that the lead digits (1, 2, …, 9) in a large data set should follow the rule

$$P(\text{lead digit is } d) = \log_{10}\left(\frac{d+1}{d}\right)$$

(So, for example, the proportion of numbers with a leading 1 is predicted to be $\log_{10}(2) \approx .3$, and the probabilities decrease as d increases.) The author of the article "Benford's Law Applies to Online Social Networks" (*PLoS One* 2015) used Twitter's API to randomly sample 78,226 Twitter users and record the number of followers each person has. The lead digits of those counts are summarized below.

Lead digit	1	2	3	4	5	6	7	8	9
Frequency	26,286	14,395	9923	7246	5737	4641	3834	3348	2816

Does the data indicate that the lead digits of the variable "number of Twitter followers" indeed conforms to Benford's Law? Test at the .05 significance level.

9. The response time of a computer system to a request for a certain type of information is hypothesized to have an exponential distribution with parameter $\lambda = 1$ [so if $X = $ response time, the pdf of X under H_0 is $f_0(x) = e^{-x}$ for $x > 0$].

a. If you had observed $X_1, X_2, \ldots, X_n$ and wanted to use the chi-squared test with five class intervals having equal probability under H_0, what would be the resulting class intervals?

b. Carry out the chi-squared test using the following data resulting from a random sample of 40 response times:

.10	.99	1.14	1.26	3.24	.12	.26	.80
.79	1.16	1.76	.41	.59	.27	2.22	.66
.71	2.21	.68	.43	.11	.46	.69	.38
.91	.55	.81	2.51	2.77	.16	1.11	.02
2.13	.19	1.21	1.13	2.93	2.14	.34	.44

10. a. Show that another expression for the chi-squared statistic (13.1) is

$$\chi^2 = \sum_{i=1}^{k} \frac{N_i^2}{np_{i0}} - n$$

Why is it more efficient to compute χ^2 using this formula?

b. When the null hypothesis is $H_0: p_1 = \cdots = p_k = 1/k$ (i.e., $p_{i0} = 1/k$ for all i), how does the formula of part (a) simplify? Use the simplified expression to calculate χ^2 for the pigeon/direction data in Exercise 4.

11. a. Having obtained a random sample from a population, you wish to use a chi-squared test to decide whether the population distribution is standard normal. If you base the test on six class intervals having equal probability under H_0, what should the class intervals be?

b. If you wish to use a chi-squared test to test H_0: the population distribution is normal with $\mu = .5$, $\sigma = .002$ and the test is to be based on six equiprobable (under H_0) class intervals, what should these intervals be?

c. Use the chi-squared test with the intervals of part (b) to decide, based on the following 45 bolt diameters, whether bolt diameter is a normally distributed variable with $\mu = .5$ in., $\sigma = .002$ in.

```
.4974  .4976  .4991  .5014  .5008  .4993
.4994  .5010  .4997  .4993  .5013  .5000
.5017  .4984  .4967  .5028  .4975  .5013
.4972  .5047  .5069  .4977  .4961  .4987
.4990  .4974  .5008  .5000  .4967  .4977
.4992  .5007  .4975  .4998  .5000  .5008
.5021  .4959  .5015  .5012  .5056  .4991
.5006  .4987  .4968
```

12. Let p_1 denote the proportion of successes in a particular population. The test statistic value in Chapter 9 for testing $H_0: p_1 = p_{10}$ was $z = (\hat{p}_1 - p_{10})/\sqrt{p_{10}p_{20}/n}$, where $p_{20} = 1 - p_{10}$. Show that for the case $k = 2$, Pearson's chi-squared statistic value

satisfies $\chi^2 = z^2$. [*Hint*: First show that $(n_1 - np_{10})^2 = (n_2 - np_{20})^2$.]

13. Consider a large population of families in which each family has exactly three children. If the sexes of the three children in any family are independent of one another, the number of male children in a randomly selected family will have a binomial distribution based on three trials.

a. Suppose a random sample of 160 families yields the following results. Test the relevant hypotheses by proceeding as in Example 13.5.

Number of Male Children	0	1	2	3
Frequency	14	66	64	16

b. Suppose a random sample of families resulted in observed frequencies of 15, 20, 12, and 3, respectively. Would the chi-squared test be based on the same number of degrees of freedom as the test in part (a)? Explain.

14. A certain type of flashlight is sold with the four batteries included. A random sample of 150 flashlights is obtained, and the number of defective batteries in each is determined, resulting in the following data:

Number defective	0	1	2	3	4
Frequency	26	51	47	16	10

Let X be the number of defective batteries in a randomly selected flashlight. Test the null hypothesis that the distribution of X is $\text{Bin}(4, \theta)$. That is, with $p_i = P(i \text{ defectives})$, test

$$H_0: p_i = \binom{4}{i} \theta^i (1 - \theta)^{4-i}$$
$$i = 0, 1, 2, 3, 4$$

[*Hint*: To obtain the mle of θ, write the multinomial likelihood (the function to be maximized) as $\theta^u (1 - \theta)^v$, where the exponents u and v are linear functions of the cell counts. Then take the natural log,

differentiate with respect to θ, equate the result to 0, and solve for $\hat{\theta}$.]

15. An article in *Annals of Mathematical Statistics* reports the following data on the number of borers (i.e., insects that bore into wood) in each of 120 groups of borers. Does the Poisson pmf provide a plausible model for the distribution of the number of borers in a group? [*Hint*: Add the frequencies for 7, 8, ..., 12 to establish a single category "≥ 7."]

Number of Borers	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	24	16	16	18	15	9	6	5	3	4	3	0	1

16. Modeling the proliferation of *E. coli* in farm animals is critical to a safe food supply. The article "Ecological and Genetic Determinants of Plasmid Distribution in *E. coli*" (*Environ. Biol.* 2016: 4230–4239) describes a study of bacterial replication in grazing cattle with low frequencies of antibiotic-resistant genes. The following data is provided on X = number of replicons for 527 bacterial isolates:

No. of replicons	0	1	2	3	4
Frequency	139	184	154	34	16

   a. The article's authors examined whether X follows a Poisson distribution. Use the data to determine the maximum likelihood estimate of the parameter μ.
   b. Perform a χ^2 test at the .05 significance level by treating the last category as "≥ 4" so that the hypothesized probabilities sum to 1. [*Hint:* Refer back to Example 13.9.]

17. The following data on X = number of corrosion defects per segment is consistent with data on one of the pipelines described in the article "The Negative Binomial Distribution as a Model for External Corrosion Defect Counts in Buried Pipelines" (*Corr. Sci.* 2015: 114–131):

x	0	1	2	3	4	5	6	7	8
Freq.	17	15	17	19	9	5	4	5	3

x	9	10	11	12	13	14	15	16	17
Freq.	0	3	0	1	1	2	1	0	1

The authors propose a generalized negative binomial model for X, which has pmf $nb(x; r, p) = k(r, x) \times p^r (1 - p)^x$ for $x = 0, 1, 2, \ldots$, where $k(r, 0) = 1$ and

$$k(r, x) = \frac{r(r+1) \cdots (r+x-1)}{x!}$$
$$\text{for } x \geq 1$$

Based on these $n = 103$ randomly selected segments, the authors estimate the negative binomial parameters to be $\hat{r} = 1.272$ and $\hat{p} = .258$. Test the hypothesis that the data is consistent with a generalized negative binomial distribution at the .10 significance level. [*Suggestion:* To ensure that all expected counts are ≥ 5, define "cells" by $x = 0, 1, \ldots, 6, 7–8$, and ≥ 9.]

18. Each headlight on an automobile undergoing an annual vehicle inspection can be focused either too high (H), too low (L), or properly (N). Checking the two headlights simultaneously (and not distinguishing between left and right) results in the six possible outcomes HH, LL, NN, HL, HN, and LN. If the probabilities (population proportions) for the single headlight focus direction are $P(H) = \theta_1$, $P(L) = \theta_2$, and $P(N) = 1 - \theta_1 - \theta_2$ and the two headlights are focused independently of each other, the probabilities of the six outcomes for a randomly selected car are the following:

$$p_1 = \theta_1^2 \quad p_2 = \theta_2^2 \quad p_3 = (1 - \theta_1 - \theta_2)^2$$
$$p_4 = 2\theta_1\theta_2 \quad p_5 = 2\theta_1(1 - \theta_1 - \theta_2)$$
$$p_6 = 2\theta_2(1 - \theta_1 - \theta_2)$$

Use the accompanying data to test the null hypothesis

$$H_0: p_1 = \pi_1(\theta_1, \theta_2), \ldots, p_6 = \pi_6(\theta_1, \theta_2)$$

where the $\pi_i(\theta_1, \theta_2)$'s are given previously.

Outcome	HH	LL	NN	HL	HN	LN
Frequency	49	26	14	20	53	38

[*Hint*: Write the multinomial likelihood as a function of θ_1 and θ_2, take the natural log, then obtain $\partial/\partial\theta_1$ and $\partial/\partial\theta_2$, equate them to 0, and solve for $\hat{\theta}_1, \hat{\theta}_2$.]

13.2   Two-Way Contingency Tables

In the previous section, we discussed inferential methods for a single categorical variable (e.g., genotype), as well as for a quantitative variable whose values have been partitioned into disjoint categories. We now study problems involving two categorical variables. There are two commonly encountered situations in which such data arises:

1. There are I populations of interest, and each population is divided into the same J categories. A sample is taken from the ith population ($i = 1, \ldots, I$), and the number of individuals in each of the J categories is recorded. For example, customers of each of $I = 3$ department store chains might have available the same $J = 5$ payment categories: cash, check, credit card, debit card, and Apple Pay.
2. There is a single population of interest, with each individual in the population categorized with respect to two different factors. There are I categories associated with the first factor and J categories associated with the second factor. A single sample is taken, and individuals are "cross-classified" by the two factors. As an example, customers making a department store purchase might be classified according to both the department in which the purchase was made (with $I = 6$ departments) and according to method of payment (with the same $J = 5$ methods as above).

In both cases (1) and (2), the data can be summarized by reporting the counts for each combination: (store chain, payment method) for (1) and (department, payment method) for (2). Let N_{ij} denote the number of individuals in the sample(s) falling in the (i, j)th category. A table displaying the n_{ij}'s (observed counts) is called a **two-way contingency table**; a prototype is shown in Table 13.7.

Table 13.7   A two-way contingency table

	1	2	...	j	...	J
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}
2	n_{21}					$\vdots$
$\vdots$	$\vdots$					
i	n_{i1}	...		n_{ij}	...	
$\vdots$	$\vdots$					
I	n_{I1}	...				n_{IJ}

In situations of the first type, we want to investigate whether the proportions in the different categories (columns) are the same for all populations (rows). The null hypothesis states that the I populations are *homogeneous* with respect to these J categories. In the second situation, we

investigate whether the categories of the two factors occur *independently* of one another in the population. It turns out, interestingly, that the two methods of analysis are actually identical (same calculations, test statistic formula, and null sampling distribution)—it doesn't matter if our two-way table is the result of stratified sampling (the first case) or simple random sampling (the second case).

Testing for Homogeneity

The test of homogeneity generalizes the two-proportion z test of Chapter 9 to the comparison of two or more populations with respect to two or more categories. We assume that each individual in every one of the I populations belongs in exactly one of J categories. A sample of n_i individuals is taken from the ith population. Let $n = \sum n_i$, the total sample size, and

$$N_{ij} = \text{ the number of individuals in the } i\text{th sample who fall into category } j$$

$$N_{\cdot j} = \sum_{i=1}^{I} N_{ij} = \begin{array}{l} \text{the total number of individuals among} \\ \text{the } n \text{ sampled who fall into category } j \end{array}$$

As before, upper-case letters denote rvs and lower-case letters the observed values. The n_{ij}'s are recorded in a contingency table with I rows and J columns (Table 13.7). The sum of the n_{ij}'s in the ith row is n_i, whereas the sum of the entries in the jth column is $n_{\cdot j}$.

Let

$$p_{ij} = \begin{array}{l} \text{the proportion of the individuals in} \\ \text{population } i \text{ who fall into category } j \end{array}$$

Thus, for population 1, the J proportions are $p_{11}, p_{12}, \ldots, p_{1J}$ (which sum to 1) and similarly for the other populations. The **null hypothesis of homogeneity** states that the proportion of individuals in category j is the same for each population and that this is true for every category; that is, for every j, $p_{1j} = p_{2j} = \cdots = p_{Ij}$.

When H_0 is true, we can use $p_1, p_2, \ldots, p_J$ to denote the population proportions in the J different categories; these proportions are common to all I populations. The expected number of individuals in the ith sample who fall in the jth category when H_0 is true is then $E(N_{ij}) = n_i \cdot p_j$. To estimate $E(N_{ij})$, we must first estimate p_j, the proportion in category j. Among the total sample of n individuals, $N_{\cdot j}$ fall into category j, so we use $\hat{P}_j = N_{\cdot j}/n$ as the estimator (this is the maximum likelihood estimator of p_j). Substitution of $\hat{P}_j$ for p_j in $n_i \cdot p_j$ yields a simple formula for estimated expected counts under H_0:

$$\hat{E}_{ij} = \text{ estimator of the expected count in cell } (i,j)$$

$$= n_i \cdot \frac{N_{\cdot j}}{n} = \frac{(i\text{th row total})(j\text{th column total})}{n} \tag{13.7}$$

The test statistic will have the same form (13.1) as in previous chi-squared tests. The number of degrees of freedom comes from the general χ^2 df rule of the previous section. In each row of Table 13.7 there are $J - 1$ freely determined cell counts (each sample size n_i is fixed), so there are a total of $I(J - 1)$ freely determined cells. Parameters $p_1, \ldots, p_J$ are estimated, but because $\sum p_j = 1$, only $J - 1$ of these are independently determined. Thus df $= I(J - 1) - (J - 1) = (I - 1)(J - 1)$.

CHI-SQUARED TEST OF HOMOGENEITY

Null hypothesis: H_0: $p_{1j} = p_{2j} = \cdots = p_{Ij}$ $j = 1, 2, \ldots, J$
Alternative hypothesis: H_a: H_0 is not true
Test statistic value:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed } - \text{ estimated expected})^2}{\text{estimated expected}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

Rejection region: $\chi^2 \geq \chi^2_{\alpha, (I-1)(J-1)}$

P-value: area under the $\chi^2_{(I-1)(J-1)}$ curve to the right of χ^2

Estimated expected counts are calculated using Expression (13.7). The test can safely be applied as long as all estimated expected counts are ≥ 5.

Example 13.10 A company packages a particular product in cans of three different sizes, each one using a different production line. Most cans conform to specifications, but a quality control engineer has identified the following reasons for nonconformance: (1) blemish on can; (2) crack in can; (3) improper pull tab location; (4) pull tab missing; (5) other. A sample of nonconforming units is selected from each of the three lines, and each unit is categorized according to reason for nonconformity, resulting in the following contingency table data:

		Reason for Nonconformity					
		Blemish	Crack	Location	Missing	Other	n_i
	1	34	65	17	21	13	150
	2	23	52	25	19	6	125
Production line	3	32	28	16	14	10	100
	Total	89	145	58	54	29	375

Does the data suggest that the proportions falling in various nonconformance categories are not the same for the three lines? The parameters of interest are various proportions, and the relevant hypotheses are

H_0: the production lines are homogeneous with respect to the five nonconformance categories; that is, $p_{1j} = p_{2j} = p_{3j}$ for $j = 1, \ldots, 5$
H_a: the production lines are not homogeneous with respect to the categories

The estimated expected frequencies (assuming homogeneity) must now be calculated using (13.7). Consider the first nonconformance category for the first production line. When the lines are homogeneous, the estimated expected number among the 150 selected units that are blemished is

$$\hat{e}_{11} = \frac{(\text{first row total})(\text{first column total})}{\text{total of sample sizes}} = \frac{(150)(89)}{375} = 35.60$$

The contribution of the cell in the upper-left corner to χ^2 is then

$$\frac{(\text{observed } - \text{ estimated expected})^2}{\text{estimated expected}} = \frac{(34 - 35.60)^2}{35.60} = .072$$

The other contributions are calculated in a similar manner. Figure 13.2 shows Minitab output for the chi-squared test. The observed count is the top number in each cell, and directly below it is the estimated expected count. The contribution of each cell to χ^2 appears below the counts, and the test statistic value is $\chi^2 = 14.159$. All estimated expected counts are at least 5, so merging categories is unnecessary. The test is based on $(3 - 1)(5 - 1) = 8$ df. Appendix Table A.10 shows that the values that capture upper-tail areas of .08 and .075 under the 8 df curve are 14.06 and 14.26, respectively. Thus the P-value is between .075 and .08; Minitab gives P-value = .079. The null hypothesis of homogeneity should not be rejected at the usual significance levels of .05 or .01, but it would be rejected for the higher α of .10.

```
Expected counts are printed below observed counts

             blem     crack      loc    missing     other     Total
1              34        65       17         21        13       150
            35.60     58.00    23.20      21.60     11.60

2              23        52       25         19         6       125
            29.67     48.33    19.33      18.00      9.67

3              32        28       16         14        10       100
            23.73     38.67    15.47      14.40      7.73

Total          89       145       58         54        29       375

Chisq = 0.072 + 0.845 + 1.657 + 0.017 + 0.169 + 1.498 + 0.278 +
        1.661 + 0.056 + 1.391 + 2.879 + 2.943 + 0.018 + 0.011 +
        0.664 = 14.159
df = 8, p = 0.079
```

Figure 13.2 Minitab output for the chi-squared test of Example 13.10

It's worth exploring the specific differences (i.e., lack of homogeneity) indicated by the χ^2 test. The **segmented bar chart** in Figure 13.3 displays the distribution of nonconformances for each of the three production lines. Line 2 appears to have a higher proportion of improper pull tab locations than the other two lines, while Line 3 has a disproportionately large number of cans with blemishes.
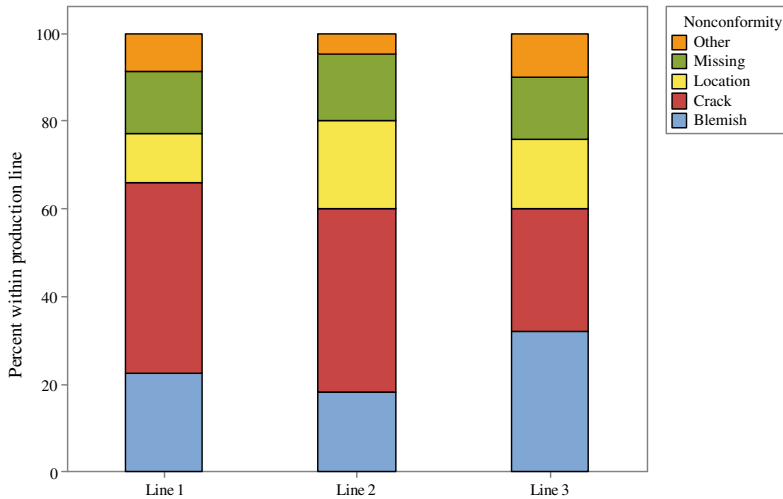
Figure 13.3 Segmented bar chart for Example 13.10   ■

When $I = 2$ and $J = 2$ (two populations and two categories), data in a two-way table may also be analyzed using the two-proportion z procedure of Chapter 9; we associate $j = 1$ with "success" and $j = 2$ with "failure." In this case, the chi-squared test of homogeneity is equivalent to the z test of $H_0: p_1 = p_2$ versus $H_a: p_1 \neq p_2$: the test statistic values are related by $(z)^2 = \chi^2$, and the P-values will be identical. The two-proportion z test allows us to consider one-sided alternatives ($p_1 > p_2$ and $p_1 < p_2$), while the chi-squared test does not. The benefit of the chi-squared test of homogeneity is that we may compare more than two populations and/or consider a response variable with more than two categories, as we did in Example 13.10.

Testing for Independence

We focus now on the relationship between two different factors in a single population. The number of categories of the first factor will be denoted by I and the number of categories of the second factor by J. Each individual in the population is assumed to belong in exactly one of the I categories associated with the first factor and exactly one of the J categories associated with the second factor. For example, the population of interest might consist of all individuals who regularly watch the national news on television, with the first factor being preferred network (ABC, CBS, NBC, PBS, CNN, Fox News, or MSNBC, so $I = 7$) and the second factor political views (liberal, moderate, conservative, giving $J = 3$).

For a sample of n individuals taken from the population, let N_{ij} denote the number among the n who fall into the (i, j)th category pair. The observed n_{ij}'s can be displayed in a two-way contingency table like Table 13.7. In the case of homogeneity for I populations, the row totals were fixed in advance, and only the J column totals were random. Now only the total sample size is fixed, and both the $N_{i.}$'s (row totals) and $N_{.j}$'s (column totals) are random variables. To state the hypotheses of interest, let

$$p_{ij} = \text{the proportion of individuals in the population who}$$
$$\text{belong in category } i \text{ of factor 1 and category } j \text{ of factor 2}$$

$$= P(\text{a randomly selected individual falls in both category}$$
$$i \text{ of factor 1 and category } j \text{ of factor 2})$$

Then

$$p_{i\cdot} = \sum_{j=1}^{J} p_{ij} = P(\text{a randomly selected individual falls in category } i \text{ of factor } 1)$$

$$p_{\cdot j} = \sum_{i=1}^{I} p_{ij} = P(\text{a randomly selected individual falls in category } j \text{ of factor } 2)$$

The **null hypothesis of independence** says that an individual's category with respect to factor 1 is independent of the category with respect to factor 2. Recall that two events A and B are independent if $P(A \cap B) = P(A) \cdot P(B)$; using the above notation, this becomes $p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$ for every pair (i, j).

The expected count in cell (i, j) is $n \cdot p_{ij}$, so when H_0 is true, $E(N_{ij}) = n \cdot p_{i\cdot} \cdot p_{\cdot j}$. To obtain a chi-squared statistic, we must therefore estimate the $p_{i\cdot}$'s $(i = 1, \ldots, I)$ and $p_{\cdot j}$'s $(j = 1, \ldots, J)$. The maximum likelihood estimators are

$$\hat{P}_{i\cdot} = \frac{N_{i\cdot}}{n} = \text{sample proportion for category } i \text{ of factor } 1 \quad \text{and}$$

$$\hat{P}_{\cdot j} = \frac{N_{\cdot j}}{n} = \text{sample proportion for category } j \text{ of factor } 2$$

This gives estimated expected cell counts identical to those in the case of homogeneity:

$$\hat{E}_{ij} = n \cdot \hat{P}_{i\cdot} \cdot \hat{P}_{\cdot j} = n \cdot \frac{N_{i\cdot}}{n} \cdot \frac{N_{\cdot j}}{n} = \frac{N_{i\cdot} \cdot N_{\cdot j}}{n}$$
$$= \frac{(i\text{th row total})(j\text{th column total})}{n} \tag{13.8}$$

Thus the test statistic is also identical to that used in testing for homogeneity. Perhaps surprisingly, so is the number of degrees of freedom! This is because the number of freely determined cell counts is $IJ - 1$, since only the total n is fixed in advance. There are I estimated $p_{i\cdot}$'s, but only $I - 1$ are independently estimated since $\sum p_{i\cdot} = 1$, and similarly $J - 1$ $p_{\cdot j}$'s are independently estimated, so $I + J - 2$ parameters are independently estimated. The df rule now yields df $= (IJ - 1) - (I + J - 2) = IJ - I - J + 1 = (I - 1)(J - 1)$, identical to the df for the test of homogeneity.

CHI-SQUARED TEST OF INDEPENDENCE

Null hypothesis: $H_0: p_{ij} = p_{i\cdot} \cdot p_{\cdot j} \quad i = 1, \ldots, I; \quad j = 1, \ldots, J$

Alternative hypothesis: $H_a: H_0$ is not true

Test statistic value:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{estimated expected})^2}{\text{estimated expected}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

Rejection region: $\chi^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$

P-value: area under the $\chi^2_{(I-1)(J-1)}$ curve to the right of χ^2

Estimated expected counts are calculated using Expression (13.8). The test can safely be applied as long as all estimated expected counts are ≥ 5.

Example 13.11 Do faculty perceive lack of diversity as a problem at their universities? Each individual in a survey of 1312 accounting faculty members across the USA was asked, "Do you think educational institutions need to improve diversity among their faculty?" (*Issues Account. Educ.* 2007). Respondents were then classified into three categories: Caucasian men, Caucasian women, and minorities. Observed and estimated expected counts (in parentheses) are given in Table 13.8. The estimated expected counts were calculated using Expression (13.8).

Table 13.8 Observed and estimated expected counts for Example 13.11

		White Men	White Women	Minority
	Yes	355 (411.70)	279 (249.91)	158 (130.39)
"Do you think…?"	No	310 (255.23)	129 (154.93)	52 (80.84)
	No response	17 (15.07)	6 (9.15)	6 (4.77)

All but one estimated expected count is ≥ 5; the value $\hat{e}_{33} = 4.77$ is close enough to 5 that a χ^2 analysis will still be accurate. In words, the hypotheses being tested for the population of all accounting faculty members in the USA are

H_0: diversity attitude and race/sex classification are independent
H_a: diversity attitude and race/sex classification are not independent

From Table 13.8, the test statistic value is

$$\chi^2 = \frac{(355 - 411.70)^2}{411.70} + \cdots + \frac{(6 - 4.77)^2}{4.77} = 45.065$$

and because $45.065 \geq \chi^2_{.01,(3-1)(3-1)} = \chi^2_{.01,4} = 13.277$, the hypothesis of independence is rejected at the .01 significance level. (The *P*-value is 0 to several decimal places.) The data suggests that a faculty member's attitude toward diversity is *not* independent of race/sex.

A segmented bar chart (Figure 13.4), along with the observed and expected counts, allows us to explore further. We see that Caucasian males were much more likely than expected to say diversity
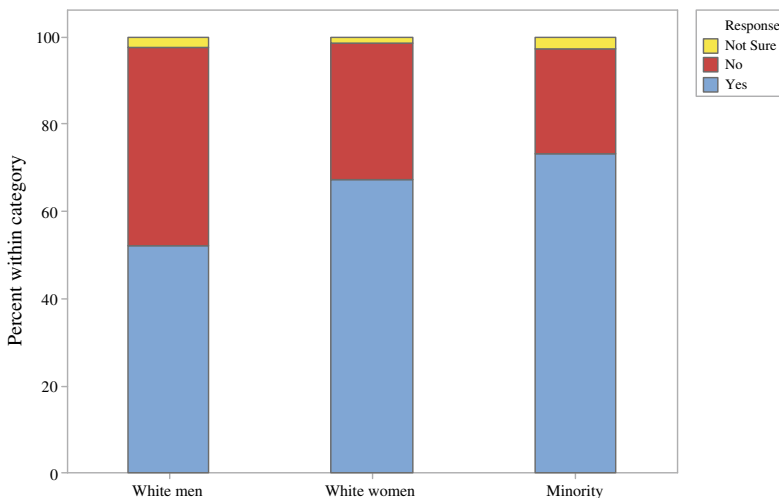
Figure 13.4 Segmented bar chart for Example 13.11

doesn't need to be improved (observed No's = 310, expected No's = 255.23), while minority faculty said "Yes" to the question of diversity improvement more often than expected if H_0 were true (observed Yes's = 158, expected = 130.39). Caucasian women's responses were somewhere in-between. ∎

Ordinal Factors and Logistic Regression

Sometimes a factor has *ordinal* categories, meaning that there is a natural ordering. For example, there is a natural ordering to freshman, sophomore, junior, senior. In such situations we can use a method that often has greater power to detect relationships by adapting the logistic regression model of Chapter 12.

Consider the case in which the first (row) factor is ordinal and the other (column) factor has two categories. Denote by X the level of the ordinal factor, which will be the predictor in the model. Let Y designate the column, so Y will be the response variable in the model. It is convenient for purposes of logistic regression to label the two columns as $Y = 0$ (failure, $j = 1$) and $Y = 1$ (success, $j = 2$), corresponding to the usual notation for Bernoulli trials. In terms of logistic regression, $p(x)$ is the probability of success given that $X = x$:

$$p(x) = P(Y = 1|X = x) = P(j = 2|i = x) = \frac{p_{x2}}{p_{x1} + p_{x2}}$$

Then the logistic model of Chapter 12 says that there exist parameters β_0, β_1 satisfying

$$e^{\beta_0 + \beta_1 x} = \frac{p(x)}{1 - p(x)} = \frac{p_{x2}}{p_{x1}}$$

In terms of the odds of success in a row (estimated by the ratio of the two counts), the model says that the odds change proportionally (by the fixed multiple e^{β_1}, the odds ratio) from row to row. For example, suppose a test is given in grades 1, 2, 3, and 4 with successes and failures as follows:

Grade	Failed	Passed	Estimated odds
1	45	45	1
2	30	60	2
3	18	72	4
4	10	80	8

Here the model fits perfectly, with odds ratio $e^{\beta_1} = 2$, so $\beta_1 = \ln(2)$ and $\beta_0 = -\ln(2)$. If a table with I rows and 2 columns has roughly a common odds ratio from row to row, then the logistic model should be a good fit if the rows are labeled with consecutive integers.

We focus on β_1 because the relationship between the two variables hinges on this parameter. The hypothesis of no relationship is equivalent to $H_0: \beta_1 = 0$, which is usually tested against a two-tailed alternative.

Example 13.12 Is there a relationship between TV watching and physical fitness? For an answer we refer to the article "Television Viewing and Physical Fitness in Adults" (*Res. Quart. Exerc. Sport* 1990: 315–320). Subjects were asked about their TV viewing habits and were classified as physically

fit if they scored in the excellent or very good category on a step test. Table 13.9 shows the results in the form of a 4×2 table.

Table 13.9 TV versus fitness results

	TV Time	Unfit	Fit
$i = 1$	0 h	147	35
$i = 2$	1–2 h	629	101
$i = 3$	3–4 h	222	28
$i = 4$	5+ h	34	4

The rows need to be given specific numeric values for computational purposes, and it is convenient to make these just 1, 2, 3, 4, because consecutive integers correspond to the assumption of a common odds ratio from row to row. (The *columns* may need to be labeled as 0 and 1 for input to software.) The logistic regression results from R are shown in Figure 13.5, where the estimated coefficient $\hat{\beta}_1$ is given as $-.2907$, for an odds ratio of $e^{-.2907} \approx .75$. This means that, for each increase in TV watching category, the odds of being fit decline to about 3/4 of the previous value.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2132     0.2675   -4.535 5.75e-06 ***
TV           -0.2907     0.1256   -2.315   0.0206 *
```

Figure 13.5 Logistic regression output (from R) for TV versus fitness

The *P*-value of .0206 associated with the z test of H_0: $\beta_1 = 0$ indicates that we should reject H_0 at the .05 level and can conclude that there is a relationship between TV watching and fitness. Of course, the existence of a relationship does not imply anything about one causing the other, because this was an observational study and not a randomized comparative experiment.

A chi-squared test of the same data, which treats both variables as unordered and does not exploit the ordinal nature of the TV viewing variable, yields $\chi^2 = 6.161$ with 3 df, *P*-value = .104. So with this test we would not conclude that there is a relationship, even at the 10% level. There is an advantage in using logistic regression for this kind of data. ∎

The analysis of two ordinal variables, each with more than two levels, can also be handled with logistic regression, but it requires a procedure called *ordinal logistic regression* that allows an ordinal response variable. When one factor is ordinal and the other is not, the analysis can be done with *multinomial* (also called nominal or polytomous) *logistic regression*, which allows a nonordinal response variable.

Models and methods for analyzing data in which each individual is categorized with respect to three or more factors (multidimensional contingency tables) are discussed in several of the references in the bibliography.

Exercises: Section 13.2 (19–31)

19. Reconsider the tax holiday data of Exercise 60(b) in Chapter 10. Use a χ^2 statistic to test the hypothesis of equal population proportions. The χ^2 statistic should be the square of the z statistic in that exercise. How are the P-values related?

20. Should illegal downloading of intellectual property (music, images, etc.) be punished? This question was asked of 501 teenagers in a study published by KRC Research (January, 2008). The teenagers were also asked whether they were familiar with the laws against illegal downloading.

		Familiar with the law?	
		Yes	No
Should illegal downloads be punished?	Yes	209	140
	No	46	106

    Are familiarity with the law and attitude toward illegal downloading independent factors within the teenage population? Test at the 5% significance level. If these factors are not independent, describe the nature of the association.

21. Brushing your teeth helps prevent cavities, doesn't it? Consider the following data from a survey and subsequent dental exam of Italian 12-year-olds ("Influence of Occlusal Disorders, Food Intake and Oral Hygiene Habits on Dental Caries on Adolescents," *Dentistry* 2016).

Brushing Freq.	Cavities	No cavities
Never	11	7
Once a day	24	21
2 times a day	99	77
3 times a day	107	117
4 times a day	42	30

    a. Test whether brushing frequency and the presence/absence of cavities are independent in the population of Italian 12-year-olds at the .05 significance level.

    b. Discuss the results of part (a): what are some possible explanations for this potentially surprising finding?

22. The authors of the article cited in the previous exercise also considered the relationship between children's dental health and their parents' education level. In the accompanying table, fathers' education levels are translated from the Italian educational system into the rough US equivalent.

Father's Education	Cavities	No cavities
<High school	51	22
High school	88	56
Some college	76	55
Higher ed. degree	40	40

    Does the data indicate that parental education is related to the prevalence of cavities in children? State the appropriate null and alternative hypotheses, compute the value of χ^2, and obtain information about the P-value. How would you then answer the question posed?

23. Do vacation habits vary by sex? The 2006 Expedia Vacation Deprivation Survey interviewed 968 Canadian adults (*Ipsos Insight* May 18, 2006). The accompanying table shows each person cross-classified by sex and the number of vacation days the person "usually [takes] each year."

	Number of vacation days						
Sex	None	1–5	6–10	11–15	16–20	20–25	>25
Female	42	25	79	94	70	58	79
Male	51	21	67	111	71	82	118

    Is there evidence at the $\alpha = .05$ significance level to conclude that the distribution of the number of vacation days taken is different for the two sexes?

24. How universal is the notion of "green light good, red light bad"? The article "Effects of Personal Experiences on the Interpretation

of the Meaning of Colours Used in the Displays and Controls in Electric Control Panels" (*Ergonomics* 2015: 1974–1982) reports the results of a survey of 144 people with occupations related to electrical equipment and 206 people in unrelated fields. Each person was asked to identify the correct meaning of colored panel lights; the accompanying data shows answers for the color red.

	Red Light Meaning?		
Occupation	Emergency situation	Normal situation	Other/ unknown
Elec. Equip.	86	40	18
Other	185	5	16

Does the data indicate a difference in how those with electrical equipment experience and those without understanding the meaning of a red panel light? Test at the .01 significance level. Discuss your findings.

25. The article "Student-Faculty Interaction in Research Universities" (*Res. High. Educ.* 2009: 437–459) reported that 20.4% of 3168 students from lower-class families said they frequently talked with faculty outside class about course material. The corresponding percentages for the 16,774 middle-class students and 8188 upper-class students were 18.6% and 20.2%, respectively. Does this data suggest that social class of a student is independent of whether or not he/she frequently talked with faculty outside class about course material?

    a. Carry out an appropriate test of hypotheses. [*Hint:* Think about how to lay out the data as a two-way table first.]

    b. In light of the sample sizes used in this study, why is the result in (a) not surprising?

26. Show that the chi-squared statistic for the test of independence can be written in the form

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{N_{ij}^2}{\hat{E}_{ij}} \right) - n$$

Why is this formula more efficient computationally than the defining formula for χ^2?

27. Suppose a random sample of students were categorized with respect to political views (liberal, moderate, conservative), marijuana usage (never, rarely, frequently), and religious affiliation (Christian, Jewish, Muslim, and other). The data could be displayed in four different two-way tables, one corresponding to each category of the third factor. With $p_{ijk} = P$(political category i, marijuana category j, and religious category k), the null hypothesis of independence of all three factors states that $p_{ijk} = p_{i\cdot\cdot} \cdot p_{\cdot j\cdot} \cdot p_{\cdot\cdot k}$. Let n_{ijk} denote the observed frequency in cell (i, j, k). Show how to estimate the expected cell counts assuming that H_0 is true ($\hat{e}_{ijk} = n\hat{p}_{ijk}$, so the $\hat{p}_{ijk}$'s must be determined). Then use the general df rule to determine the number of degrees of freedom for the chi-squared statistic.

28. Suppose that in a particular state consisting of four distinct regions, a random sample of n_k voters is obtained from the kth region for $k = 1, 2, 3, 4$. Each voter is then classified according to which candidate (1, 2, or 3) he/she prefers and according to voter registration (1 = Dem., 2 = Rep., 3 = Other). Let p_{ijk} denote the proportion of voters in region k who belong in candidate category i and registration category j. The null hypothesis of homogeneous regions is H_0: $p_{ij1} = p_{ij2} = p_{ij3} = p_{ij4}$ for all i, j (i.e., the proportion within each candidate/registration combination is the same for all four regions). Assuming that H_0 is true, determine $\hat{p}_{ijk}$ and $\hat{e}_{ijk}$ as functions of the observed n_{ijk}'s, and use the general df rule to obtain the number of degrees of freedom for the chi-squared test.

29. Consider the accompanying 2×3 table displaying the sample proportions that fell

in various combinations of categories (e.g., 13% of those in the sample were in the first category of both factors).

	1	2	3
1	.13	.19	.28
2	.07	.11	.22

a. Suppose the sample consisted of $n = 100$ people. Perform the chi-squared test for independence with significance level .10.
b. Repeat part (a) assuming that the sample size was $n = 1000$.
c. What is the smallest sample size n for which these observed proportions would result in rejection of the independence hypothesis at the .10 level?

30. Use logistic regression to test the relationship between cavities and father's education in Exercise 22. Compare the P-value with what was found in Exercise 22. (Remember that $\chi_1^2 = z^2$.) Explain why you expected the logistic regression to give a smaller P-value.

31. A random sample of 100 faculty at a university gives the results shown below for professorial rank versus sex.

Rank	Male	Female
Professor	25	9
Assoc Prof	20	8
Asst Prof	18	20

a. Test for a relationship at the 5% level using a chi-squared statistic.
b. Test for a relationship at the 5% level using logistic regression.
c. Compare the P-values in parts (a) and (b). Is this in accord with your expectations? Explain.
d. Interpret your results. Assuming that today's assistant professors are tomorrow's associate professors and professors, do you see implications for the future?

Supplementary Exercises: (32–43)

32. The report "Majoring in Money: How American College Students Manage Their Finances" (*Sallie Mae* 2016) includes the following data on whether students in different age groups have at least one credit card. Data was based on a survey of randomly selected US college students.

Age	n	Credit card(s)?
18–20	348	43%
21–22	258	63%
23–24	187	71%

Does the data provide convincing evidence that, among US college students, credit card ownership rate varies by age group? Test at the $\alpha = .01$ significance level. [*Hint:* Think about how to lay out a contingency table.]

33. The report cited in the previous exercise also asked students with credit cards how much they pay off each month.

	Male	Female
Full balance	146	131
Minimum payment	17	21
Other	52	78

Perform a χ^2 test, and report your results at the .05 significance level. Be clear about what hypotheses you're testing!

34. The article "Psychiatric and Alcoholic Admissions Do Not Occur Disproportionately Close to Patients' Birthdays" (*Psych. Rep.* 1992: 944–946) focuses on the existence of any relationship between date of patient admission for treatment of alcoholism and patient's birthday. Assuming a 365-day year (i.e., excluding leap year), in the absence of any relation, a patient's admission date is equally likely to be any one of the 365 possible days. The investigators established four different admission categories: (1) within

7 days of birthday, (2) between 8 and 30 days, inclusive, from the birthday, (3) between 31 and 90 days, inclusive, from the birthday, and (4) more than 90 days from the birthday. A sample of 200 patients gave observed frequencies of 11, 24, 69, and 96 for categories 1, 2, 3, and 4, respectively. State and test the relevant hypotheses using a significance level of .01.

35. A Gallup survey (August 9, 2019) asked adults who consume alcoholic beverages for their favorite type. The following table shows responses separated by the region where each respondent lives:

	Liquor	Wine	Beer
East	63	84	105
Midwest	86	73	156
South	197	174	186
West	106	120	131

Does the data suggest that adult beverage preferences vary by region? Test at the .05 significance level. Discuss your findings.

36. Qualifications of male and female head and assistant college athletic coaches were compared in the article "Sex Bias and the Validity of Believed Differences Between Male and Female Interscholastic Athletic Coaches" (*Res. Q. Exerc. Sport* 1990: 259–267). Each person in random samples of 2225 male coaches and 1141 female coaches was classified according to number of years of coaching experience to obtain the accompanying two-way table. Is there enough evidence to conclude that the proportions falling into the experience categories are different for men and women? Use $\alpha = .01$.

		Years of Experience			
Sex	1–3	4–6	7–9	10–12	13+
Male	202	369	482	361	811
Female	230	251	238	164	258

37. The authors of the article "Predicting Professional Sports Game Outcomes from Intermediate Game Scores" (*Chance* 1992: 18–22) used a chi-squared test to determine whether there was any merit to the idea that basketball games are not settled until the last quarter, whereas baseball games are over by the seventh inning. They also considered football and hockey. Data was collected for 189 basketball games, 92 baseball games, 80 hockey games, and 93 football games. The games analyzed were sampled randomly from all games played during the 1990 season for baseball and football and for the 1990–1991 season for basketball and hockey. For each game, the late-game leader was determined, and then it was noted whether the late-game leader actually ended up winning the game. The resulting data is summarized in the accompanying table.

Sport	Late-Game Leader Wins	Late-Game Leader Loses
Basketball	150	39
Baseball	86	6
Hockey	65	15
Football	72	21

The authors state, "Late-game leader is defined as the team that is ahead after three quarters in basketball and football, two periods in hockey, and seven innings in baseball. The chi-square value on three degrees of freedom is 10.52 ($P < .015$)."

a. State the relevant hypotheses and reach a conclusion using $\alpha = .05$.

b. Do you think that your conclusion in part (a) can be attributed to a single sport being an anomaly?

38. A study in the *Journal of Marketing Research* investigated the relationship between facility conditions at gas stations and aggressiveness in the pricing of gasoline. The accompanying data is based on a random sample of $n = 441$ stations.

		Observed pricing policy		
		Aggressive	Neutral	Nonaggressive
Condition	Substandard	24	15	17
	Standard	52	73	80
	Modern	58	86	36

a. Does the data suggest that an association exists between these two variables? Test at the $\alpha = .01$ level.

b. If a statistically significant association exists, describe that association carefully and in context.

39. The Associated Press (Dec. 7, 2005) reported on an international survey about the treatment of terrorist suspects. Random samples of 1000 adults from each of several nations were asked, "Do you feel the use of torture against suspected terrorists to obtain information about terrorists activities is justified?" Data consistent with the article appears in the accompanying table.

Country	Okay to torture terror suspects?				
	Never	Rarely	Sometimes	Often	Not sure
Italy	600	140	140	90	30
France	400	250	200	120	30
South Korea	100	330	470	60	40
Spain	540	160	140	70	90
USA	360	230	270	110	30

Does the data suggest that attitudes toward the treatment of terrorist suspects differed between these five nations in 2005? State and test the relevant hypotheses at the $\alpha = .01$ level. Comment on any specific trends.

40. The likelihood ratio test of Chapter 9 provides an alternative to Pearson's chi-squared statistic. Let $L(p_1, \ldots, p_k) = C \cdot p_1^{n_1} \cdots p_k^{n_k}$ denote the multinomial likelihood function (C will be irrelevant in what follows). The likelihood ratio test statistic is

$$\Lambda = \frac{L(p_{10}, \ldots, p_{k0})}{L(\hat{p}_1, \ldots, \hat{p}_k)},$$

where $\hat{P}_i = N_i/n$, the sample proportion of observations in the ith category (the mle for p_i). The key result required for the test is that for large n, $-2\ln(\Lambda)$ has approximately a χ^2_{k-1} distribution.

a. Using the information provided, simplify the test statistic $-2\ln(\Lambda)$ as much as possible.

b. If a roulette wheel is working properly, spins should land on the colors black, red, and green in proportions 18/38, 18/38, and 2/38, respectively. Suppose that 190 spins resulted in 96 black, 76 red, and 18 green. Use the likelihood ratio test to determine whether the sample data is compatible with the theoretical probabilities.

c. Use Pearson's chi-squared goodness-of-fit test for the data in part (b). How do the results of the two tests compare?

41. The NCAA basketball tournament begins with 64 teams that are apportioned into four regional tournaments, each involving 16 teams. The 16 teams in each region are then ranked (seeded) from 1 to 16. During the 12-year period from 1991 to 2002, the top-ranked team won its regional tournament 22 times, the second-ranked team won 10 times, the third-ranked team won 5 times, and the remaining 11 regional tournaments were won by teams ranked lower than 3. Let P_{ij} denote the probability that the team ranked i in its region is victorious in its game against the team ranked j. Once the P_{ij}'s are available, it is possible to compute the probability that any particular seed wins its regional tournament (a complicated calculation because the number of outcomes in the sample space is quite large). The paper "Probability Models for the NCAA Regional Basketball Tournaments" (*Amer. Statist.* 1991: 35–38) proposed several different models for the P_{ij}'s.

a. One model postulated $P_{ij} = .5 - \lambda(i - j)$ with $\lambda = \frac{1}{32}$ (from which $P_{16,1} = \frac{1}{32}$, $P_{16,2} = \frac{2}{32}$, etc.). Based on this, P(seed #1 wins) = .27477, P(seed #2 wins) = .20834, and P(seed #3 wins) = .15429. Does this model appear to provide a good fit to the data?

b. A more sophisticated model has $P_{ij} = .5 + .2813625(z_i - z_j)$, where the z's are measures of relative strengths related to standard normal percentiles [percentiles for successive highly seeded teams are closer together than is the case for teams seeded lower, and $.2813625$ ensures that the range of probabilities is the same as for the model in part (a)]. The resulting probabilities of seeds 1, 2, or 3 winning their regional tournaments are .45883, .18813, and .11032, respectively. Assess the fit of this model.

42. Have you ever wondered whether soccer players suffer adverse effects from hitting "headers"? The authors of the article "No Evidence of Impaired Neurocognitive Performance in Collegiate Soccer Players" (*Amer. J. Sports Med.* 2002: 157–162) investigated this issue from several perspectives.

   a. The paper reported that 45 of the 91 soccer players in their sample had suffered at least one concussion, 28 of 96 nonsoccer athletes had suffered at least one concussion, and only 8 of 53 student controls had suffered at least one concussion. Analyze this data and draw appropriate conclusions.

   b. For the soccer players, the sample correlation coefficient calculated from the values of $x =$ soccer exposure (total number of competitive seasons played prior to enrollment in the study) and $y =$ score on an immediate memory recall test was $r = -.220$. Interpret this result.

   c. Here is summary information on scores on a controlled oral word association test for the soccer and nonsoccer athletes:

   $$n_1 = 26, \bar{x}_1 = 37.50, s_1 = 9.13,$$
   $$n_2 = 56, \bar{x}_2 = 39.63, s_2 = 10.19$$

   Analyze this data and draw appropriate conclusions.

   d. Considering the number of prior nonsoccer concussions, the values of mean $\pm$ SD for the three groups were soccer players, $.30 \pm .67$; nonsoccer athletes, $.49 \pm .87$; and student controls, $.19 \pm .48$. Analyze this data and draw appropriate conclusions.

43. Do the successive digits in the decimal expansion of π behave as though they were selected from a random number table (or came from a computer's random number generator)?

   a. Let p_0 denote the long-run proportion of digits in the expansion that equal 0, and define $p_1, \ldots, p_9$ analogously. What hypotheses about these proportions should be tested, and what is df for the chi-squared test?

   b. H_0 of part (a) would not be rejected for the nonrandom sequence 012 … 901 … 901 …. Consider nonoverlapping groups of two digits, and let p_{ij} denote the long-run proportion of groups for which the first digit is i and the second digit is j. What hypotheses about these proportions should be tested, and what is df for the chi-squared test?

   c. Consider nonoverlapping groups of 5 digits. Could a chi-squared test of appropriate hypotheses about the p_{ijklm}'s be based on the first 100,000 digits? Explain.

   d. The paper "Are the Digits of π an Independent and Identically Distributed Sequence?" (*Amer. Statist.* 2000: 12–16) considered the first 1,254,540 digits of π, and reported the following P-values for group sizes of 1, …, 5 digits: .572, .078, .529, .691, .298. What would you conclude?

Nonparametric Methods

14

Introduction

In this chapter we consider some inferential methods that are different in important ways from those considered earlier. Recall that many of the confidence intervals and test procedures developed in Chapters 8, 9, 10, 11 and 12 were based on some sort of a normality assumption. As long as such an assumption is at least approximately satisfied, the actual confidence and significance levels will be at least approximately equal to the "nominal" levels, those prescribed by the experimenter through the choice of particular t or F critical values. However, if there is a substantial violation of the normality assumption, the actual levels may differ considerably from the nominal levels (e.g., the use of $t_{.025}$ in a confidence interval formula may actually result in a confidence level of only 88% rather than the nominal 95%, more than doubling the error rate). Here we develop *nonparametric* or *distribution-free* procedures that are valid for a wide variety of underlying distributions rather than being tied to normality. We have actually already introduced several such methods: the bootstrap intervals and permutation tests are valid without restrictive assumptions on the underlying distribution(s).

Section 14.1 details inference procedures for population quantiles—the population median, 90th percentile, and so on—that apply to any continuous distribution. In Section 14.2, we present alternatives to the one-sample t procedures that do not require population normality (although they do make some less-restrictive distributional assumptions). The most popular nonparametric methods are so-called *rank-based tests*, wherein the original raw data is replaced by their ranks (1 for the smallest observation, 2 for the next smallest, etc.). Sections 14.3 and 14.4 describe rank-based alternatives to two-sample t procedures, one-way ANOVA, and randomized block ANOVA.

14.1 Exact Inference for Population Quantiles

The inferential methods presented so far in this book—including t tests and the analysis of variance—have largely focused on one or more population means. However, in some situations other summary measures are more relevant. For example, house prices in any particular city or region are famously right-skewed: a small number of very large, expensive homes inflates the mean cost. Realtors or buyers looking to quantify the "typical" house price in an area might be better served to estimate the population *median* price, rather than the mean price. Or, an internet service provider may plan to charge an extra fee to the 5% of customers with the heaviest data usage, meaning that the population 95th percentile is of interest to the company.

As in Chapter 4, we will use the notation η_p to denote the $(100p)$th percentile (aka the pth quantile) of a probability distribution, i.e., the value that separates the lowest $(100p)\%$ of values from the rest. So, for instance, a population upper quartile (75th percentile) will be denoted by $\eta_{.75}$. The population median, $\eta_{.5}$, will be more frequently denoted by $\tilde{\mu}$ as in earlier sections. Here, we first develop a general confidence interval method for η_p. Then, we present a hypothesis testing procedure for $\tilde{\mu}$ which may be applied to the analysis of both one-sample and paired data.

A CI for a Population Quantile

Inferences on population quantiles depend, perhaps not surprisingly, on the quantiles of the sample data. To that end, let $X_1, \ldots, X_n$ represent a random sample from some continuous population distribution of interest. The **order statistics** $Y_1, \ldots, Y_n$ as defined in Chapter 5 are

$$Y_1 = \text{the smallest among } X_1, X_2, \ldots, X_n \text{ (i.e., the sample minimum)}$$
$$Y_2 = \text{the second smallest among } X_1, X_2, \ldots, X_n$$
$$\vdots$$
$$Y_n = \text{the largest among } X_1, X_2, \ldots, X_n \text{ (the sample maximum)}$$

Because the population distribution is assumed continuous, with probability one there will be no ties among the X_i's and, hence, $Y_1 < Y_2 < \cdots < Y_n$. Note, though, that no other assumptions (such as normality) are made about the population—the methods presented here apply to a broad range of distributions. Confidence intervals for population quantiles rely on the following proposition.

PROPOSITION   Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with pth quantile η_p $(0 < p < 1)$ and let $Y_1, \ldots, Y_n$ represent the corresponding order statistics. Then for any two integers r and s satisfying $1 \leq r < s \leq n$,

$$P(Y_r \leq \eta_p \leq Y_s) = \sum_{k=r}^{s-1} \binom{n}{k} p^k (1-p)^{n-k} \tag{14.1}$$

Proof   For any integer k between 1 and $n-1$, η_p will fall between the consecutive order statistics Y_k and Y_{k+1} if and only if exactly k of the X_i's are $\leq \eta_p$. Now consider X_i a success if $X_i \leq \eta_p$ and a failure otherwise. Since the X_i's are independent, the number of successes among them (n independent trials) is a binomial rv with parameters n and $P(\text{success}) = P(X_i \leq \eta_p) = p$. Hence

$$P(Y_k \leq \eta_p \leq Y_{k+1}) = P(\text{exactly } k \text{ of the } X_i\text{'s are} \leq \eta_p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Therefore, for integers $r < s$,

$$P(Y_r \leq \eta_p \leq Y_s) = P(\{Y_r \leq \eta_p \leq Y_{r+1}\} \cup \{Y_{r+1} \leq \eta_p \leq Y_{r+2}\} \cup \cdots \cup \{Y_{s-1} \leq \eta_p \leq Y_s\})$$

$$= P(Y_r \leq \eta_p \leq Y_{r+1}) + \cdots + P(Y_{s-1} \leq \eta_p \leq Y_s) = \sum_{k=r}^{s-1} \binom{n}{k} p^k (1-p)^{n-k} \qquad \blacksquare$$

Suppose now that a confidence interval for η_p is desired. The preceding proposition indicates that if the interval (y_r, y_s) is used, then the associated confidence level is the binomial probability on the right-hand side of Expression (14.1). Due to the discrete nature of this probability, it will not be possible to achieve every confidence level. In practice, r and s are selected so that the confidence level from (14.1) is as close as possible to, but no lower than, the desired level.

Example 14.1 Let's determine a 90% confidence interval for the population upper quartile $\eta_{.75}$ based on a sample of size $n = 20$ from *any* continuous population distribution. Logically, the interval should straddle the sample 75th percentile, which is very roughly the $.75(20) = 15$th ordered observation. The required indices r and s can then by determine by trial and error using (14.1). For instance,

$$P(Y_{12} \leq \eta_{.75} \leq Y_{18}) = \sum_{k=12}^{18-1} \binom{20}{k}(.75)^k(.25)^{20-k} = .8678$$

Hence the interval (y_{12}, y_{18}) is slightly too "narrow," in the sense that the associated confidence level is just shy of 90%. However, a similar calculation shows $P(Y_{12} \leq \eta_{.75} \leq Y_{19}) = .9348 > .90$, and this is as close as we can get to .90 without going under. So, the suggested CI is (y_{12}, y_{19}).

With 93.48% confidence, the parameter $\eta_{.75}$ lies between y_{12} and y_{19}, the 12th-smallest and 19th-smallest (i.e., second largest) ordered values from a sample of size $n = 20$. Again, this interval is valid for *any* (continuous) population; the X_i's may come from a normal, Weibull, or any other continuous distribution. ∎

Expression (14.1) can be modified slightly to obtain one-sided bounds for the pth quantile. If an upper confidence bound is desired, delete Y_r from the left-hand side of (14.1) and substitute $r = 0$ into the binomial calculation on the right-hand side. Similarly, eliminating Y_s and substituting $s - 1 = n$ in the binomial calculation results in a lower confidence bound.

Determining the indices r and s to achieve a desired confidence level can clearly be tedious. Notice that if the desired (two-sided) confidence level is $100(1 - \alpha)\%$, then on the right-hand side of (14.1) r and $s - 1$ are effectively the $\alpha/2$ and $1 - \alpha/2$ quantiles of the Bin(n, p) distribution. Using the normal approximation to the binomial from Chapter 4 with a continuity correction, r and s can then be approximated by

$$\begin{array}{ll} r - .5 \approx \mu - z_{\alpha/2}\sigma \\ (s - 1) + .5 \approx \mu + z_{\alpha/2}\sigma \end{array} \Rightarrow \begin{array}{l} r \approx (np + .5) - z_{\alpha/2}\sqrt{np(1-p)} \\ s \approx (np + .5) + z_{\alpha/2}\sqrt{np(1-p)} \end{array}$$

In Example 14.1, even though the normal approximation is of questionable accuracy (n is fairly small), the preceding expressions give $r \approx 12.3$ and $s \approx 18.7$, which round to the correct integers found in the example.

Hypothesis Testing for a Population Median

A binomial calculation similar to the one presented in Expression (14.1) can also be used to calculate the P-value for a hypothesis test concerning a population quantile. Here we focus on the population median, because this is the most common quantile of interest, but the ideas can easily be generalized to any other percentile. Consider the null hypothesis H_0: $\tilde{\mu} = \tilde{\mu}_0$, where $\tilde{\mu}_0$ is the null value of the population median. If H_0 is true, we expect roughly half of the sample observations to fall below $\tilde{\mu}_0$ and the other half above it. A test procedure is based on counting how many of the x_i's exceed $\tilde{\mu}_0$.

Example 14.2 Example 1.17 presented $n = 57$ observations on total nitrogen (TN) load (kg/day) from a particular location in Chesapeake Bay. The data is extremely right-skewed (see Figure 1.17). The sample median of these 57 observations is $\tilde{x} = 92.2$. Let's test the hypothesis that the *population* median TN load exceeds 60 kg/day.

With $\tilde{\mu}$ = true median TN load in this part of Chesapeake Bay, the null and alternative hypotheses are

$$H_0: \tilde{\mu} = 60$$
$$H_a: \tilde{\mu} > 60$$

Of the 57 TN values in the sample, 36 exceed 60 kg/day and the other 21 are below this value. If H_0 is true, we'd expect 28.5 measurements on either side of 60, so the data appears to somewhat contradict the null hypothesis.

To compute a *P*-value, let's determine the probability that 36 *or more* of the observations in a sample of size 57 would exceed 60 kg/day, if that is truly the population median. If H_0 is true, the number of X_i's that exceed 60 has a binomial distribution with $n = 57$ and $p = .5$, since by definition $P(X_i > \tilde{\mu}) = P(X_i < \tilde{\mu}) = .5$. Reflecting the upper-tailed alternative hypothesis, the *P*-value is

$$P\text{-value} = P(36 \text{ } or \text{ } more \text{ } X_i\text{'s exceed } 60, \text{ when } \tilde{\mu} = 60)$$
$$= \sum_{k=36}^{57} \binom{57}{k} (.5)^k (1 - .5)^{57-k} = .031$$

With this low *P*-value, H_0 is rejected (in particular, $.031 \leq .05$). At the .05 significance level, we have evidence that the true median TN load at this location exceeds 60 kg/day.

The preceding test can also be reframed by defining a new parameter. Let $p = P(X_i > 60)$, the probability that a random TN observation will exceed 60. If H_0 is true, then 60 is the population median and $p = .5$. However, if H_a is true, then 60 kg/day is *less* than the actual median $\tilde{\mu}$, and so *more than half* of the X distribution exceeds 60. That is, H_a is equivalent to the assertion that $p > .5$. To test the modified hypotheses

$$H_0: p = .5$$
$$H_a: p > .5$$

we can use either of the one-proportion procedures presented in Section 9.3. The *P*-value for the exact binomial test from the end of that section is identical to the calculation above; alternatively, since n is fairly large, the one-proportion z test would also be appropriate. ∎

The hypothesis test illustrated in Example 14.2 is called a (one-sample) **sign test**. Why a "sign" test? One way to think about the binomial count is to look at the quantities $(X_1 - \tilde{\mu}_0), \ldots, (X_n - \tilde{\mu}_0)$: each has a positive sign (i.e., is > 0) when $X_i > \tilde{\mu}_0$ and a negative sign when $X_i < \tilde{\mu}_0$. In Example 14.2, the data was equivalent to 36 positive signs and 21 negative signs, and the test statistic value was the number of positive signs.

The one-sample sign test is often applied to paired data. Consider a study with two settings, A and B (e.g., A = after physical therapy and B = before). If we let X_i and Y_i denote the ith individual's response (e.g., range of motion) in settings A and B, respectively, we know from previous discussions to examine the within-subject differences $D_i = X_i - Y_i$. A "positive sign," meaning $D_i > 0$, indicates that the ith subject got a higher response under setting A than with setting B. We can test for a treatment effect in favor of setting A—for example, that physical therapy increases range of motion— by examining the hypotheses

$$H_0: \tilde{\mu}_D = 0$$
$$H_a: \tilde{\mu}_D > 0$$

Equivalently, if we define $p = P(X_i > Y_i) = P(D_i > 0)$, then we may test $H_0: p = .5$ versus $H_a: p > .5$ as at the end of Example 14.2.

Example 14.3 Do technological improvements "slow down" as a product spends more time on the market? The 2017 MIT paper "Exploring the Relationship Between Technological Improvement and Innovation Diffusion: An Empirical Test" attempts to answer this question by quantifying the improvement rate (%/year) of 18 items, from washing machines to laptops, during both the early stage and late stage of each item's market presence. The data is summarized below.

Early stage	Late stage	Difference	Early stage	Late stage	Difference
12.96	12.65	−0.31	15.49	18.80	+3.31
5.50	3.52	−1.98	34.30	25.73	−8.57
5.50	5.00	−0.50	32.15	32.80	+0.65
3.90	3.10	−0.80	16.62	15.84	−0.78
3.52	2.93	−0.59	32.37	36.33	+3.96
3.90	3.10	−0.80	24.25	27.15	+2.90
10.53	14.41	+3.88	3.87	3.92	+0.05
38.18	31.79	−6.39	180.84	84.51	−96.33
31.79	36.33	+4.54	26.80	47.52	+20.72

With D = difference in improvement rate (late stage minus early stage), the authors tested the hypotheses $H_0: \tilde{\mu}_D = 0$ versus $H_a: \tilde{\mu}_D < 0$; the alternative hypothesis aligns with the prevailing theory of a late-stage slowdown. Eight of the 18 differences are positive, and the lower-tailed P-value is the chance of observing eight or fewer positive differences if the true median difference is zero (so positive and negative are each equally likely):

$$P\text{-value} = P(K \le 8 \text{ when } K \sim \text{Bin}(18, .5)) = B(8; 18, .50) = .408$$

The very large P-value, consistent with the close split (8 vs. 10) between negative and positive differences, suggest that H_0 should not be rejected. The data does not lend credence to the slowdown theory of technological improvement. ∎

Exercises: Section 14.1 (1–10)

1. Example 1.4 presented data on the starting salaries of $n = 38$ civil engineering graduates. Use this data to construct a 95% confidence interval for the population lower quartile (25th percentile) of civil engineering starting salaries.

2. The following Houston, TX house prices ($1000's) were extracted from zillow.com in August 2019:

162	165	167	188	189	194	200	233	236	247
248	257	258	286	290	307	330	345	377	389
459	460	513	569	1399					

   Treating these 25 houses as a random sample of all available homes in Houston, calculate a 90% upper confidence bound for the true median home price in that city.

3. Based on a random sample of 40 observations from *any* continuous population, construct a confidence interval formula for the population median that has confidence level (at least) 95%.

4. Let $\eta_{.3}$ denote the 30th percentile of a population. Find the smallest sample size n for which $P(Y_1 \le \eta_{.3} \le Y_n)$ is at least .99.

In other words, determine the smallest sample size for which the span of the sample, min to max, is a 99% CI for $\eta_{.3}$.

5. Refer back to Exercise 2. The median home value in the state of Texas in August 2019 was \$197,000. Use the data in Exercise 2 to test the hypothesis that the true median home price in Houston exceeds this value, at the .05 significance level.

6. The following data on grip strength (N) for 42 individuals was read from a graph in the article "Investigation of Grip Force, Normal Force, Contact Area, Hand Size, and Handle Size for Cylindrical Handles" (*Human Factors* 2008: 734–744):

16	18	20	26	33	41	54	56	66
68	87	91	95	98	106	109	111	118
127	131	135	145	147	149	151	168	172
183	189	190	200	210	220	229	230	
233	238	244	259	294	329	403		

   a. What does the data suggest about the population distribution of grip strength? Why might the median be a more appropriate measure of "typical" grip strength than the mean?
   b. Test the hypothesis that the population median grip strength is less than 170 N at the .05 significance level.

7. Child development specialists widely believe that diverse recreational activities can improve the social and emotional conduct of children. The article "Influence of Physical Activity on the Social and Emotional Behavior of Children Aged 2–5 Years" (*Cuban J. of Gen. Integr. Med.* 2016) reported a study of 25 young children diagnosed with social and emotional behavior problems. The children participated in a physical activity regimen for one year, and each child was measured for negative social behavior indicators (tantrums, crying, etc.) both before and after the regimen. Lower scores indicate improvement; the children's changes in score (post minus pre) are summarized below.

Score change < 0	Score change = 0	Score change > 0
17	1	7

   Use the sign test to determine whether the data indicates a statistically significant improvement in scores at the $\alpha = .05$ level. [*Hint:* Delete the one 0 observation and work with the other 24 differences; this is a common way to address "ties" in pre- and post-intervention scores.]

8. Consider the following scene: an actor watches a man in a gorilla suit hide behind haystack *A*. The actor leaves the area, and the "gorilla" moves to behind haystack *B*, after which the actor re-enters. A video of this scene was shown to 22 (real) apes, and eye-tracking software was used to track which haystack they stared at more after the actor re-entered. "False belief" theory says that the viewer will look at haystack *A* more, matching the *actor's* mistaken belief that the gorilla is still there, despite the *viewer* knowing better. In the study, 17 of the apes spent more time looking at haystack *A* ("Great Apes Anticipate That Other Individuals Will Act According to False Beliefs," *Science*, 7 October 2016).

   a. Use a sign test to determine whether the true median looking-time difference $(A - B)$ supports the false belief theory in primates.
   b. Since the data is paired (time looking at each of two haystacks for the 22 apes), the paired *t* procedure of Chapter 10 might also be applicable. What information would that test require, and what assumptions must be met?

9. The article "Hitting a High Note on Math Tests: Remembered Success Influences Test Preferences" (*J. of Exptl. Psych.* 2016: 17–38) reported a study in which 130 participants were administered two math tests (in random order): a shorter, more difficult exam and a longer, easier one. Participants were then asked to estimate how much time they had spent on each exam. Let $\tilde{\mu}_D$ denote

the true median difference in time estimates (short test minus long test). Test the hypotheses H_0: $\tilde{\mu}_D = 0$ versus H_a: $\tilde{\mu}_D > 0$, with H_a supporting the psychological contention that people perceive easier tasks to be quicker, using a sign test based on the fact that 109 participants gave positive differences and 21 gave negative differences. (In fact, test-takers required 3.2 min longer, on average, to complete the lengthier exam.)

10. Consider the following data on resting energy expenditure (REE, in calories per day) for eight subjects both while on an intermittent fasting regimen and while on a standard diet ("Intermittent Fasting Does Not Affect Whole-Body Glucose, Lipid, or Protein Metabolism," *Amer. J. of Clinical Nutr.* 2009: 1244–1251):

	Subject				
Diet	1	2	3	4	5
I.F.	1753.7	1604.4	1576.5	1279.7	1754.2
Standard	1755.0	1691.1	1697.1	1477.7	1785.2

	Subject		
Diet	6	7	8
I.F.	1695.5	1700.1	1717.0
Standard	1669.7	1901.3	1735.3

Let $\tilde{\mu}_D$ = true median difference in REE (IF minus standard diet). Test the hypotheses H_0: $\tilde{\mu}_D = 0$ versus H_a: $\tilde{\mu}_D < 0$ at the .05 significance level using the sign test.

14.2 One-Sample Rank-Based Inference

The previous section introduced the *sign test* for assessing the plausibility of H_0: $\tilde{\mu} = \tilde{\mu}_0$, where $\tilde{\mu}$ denotes a population median. The basis of the test was to consider the quantities $X_1 - \tilde{\mu}_0, \ldots, X_n - \tilde{\mu}_0$ and count how many of those differences are positive. Thus the original sample is reduced to a collection of n "signs" (+ or −). If H_0 is true, there should be roughly equal numbers of +'s and −'s, and the test statistic measures the degree of discrepancy from that 50–50 balance. The sign test is applicable to any continuous population distribution.

Here, we consider a test procedure that is more powerful than the sign test but requires an additional distributional assumption. Suppose a research chemist replicated a particular experiment a total of 10 times and obtained the following values of reaction temperature (°C), ordered from smallest to largest:

$$-.76 \quad -.19 \quad -.05 \quad .57 \quad 1.30 \quad 2.02 \quad 2.17 \quad 2.46 \quad 2.68 \quad 3.02$$

The distribution of reaction temperature is of course continuous. Suppose the investigator is willing to assume that this distribution is *symmetric*, in which case the two halves of the distribution on either side of $\tilde{\mu}$ are mirror images of each other. (Provided that the mean μ exists for this symmetric distribution, $\tilde{\mu} = \mu$ and they are both the point of symmetry.) The assumption of symmetry may at first seem quite bold, but remember that we have frequently assumed a normal distribution for inference procedures. Since a normal distribution is symmetric, the assumption of symmetry without any additional distributional specification is actually a weaker assumption than normality.

Let's now consider testing the specific null hypothesis that $\tilde{\mu} = 0$. Symmetry implies that a temperature of any particular magnitude, say 1.50, is no more likely to be positive (+1.50) than to be negative (−1.50). A glance at the data above casts doubt on this hypothesis; for example, the sample median is 1.66, which is far larger in magnitude than any of the three negative observations.

Figure 14.1 shows graphs of two symmetric pdfs, one for which H_0 is true and the other for which the median of the distribution considerably exceeds 0. In the first case we expect the magnitudes of the negative observations in the sample to be comparable to those of the positive sample observations. However, in the second case observations of large absolute magnitude will tend to be positive rather than negative.
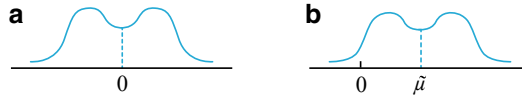
Figure 14.1 Distributions for which (a) $\tilde{\mu} = 0$; (b) $\tilde{\mu} \gg 0$

A Rank-Based Test Statistic

For the sample of ten reaction temperatures, let's for the moment disregard the signs of the observations and *rank* their magnitudes (i.e., absolute values) from 1 to 10, with the smallest getting rank 1, the second smallest rank 2, and so on. Then apply the original sign of each observation to the corresponding rank, so some **signed ranks** will be negative (e.g., −3) whereas others will be positive (e.g., +8).

Absolute value	.05	.19	.57	.76	1.30	2.02	2.17	2.46	2.68	3.02
Rank	1	2	3	4	5	6	7	8	9	10
Signed rank	−1	−2	3	−4	5	6	7	8	9	10

The test statistic for the procedure developed in this section will be $S_+ =$ *the sum of the positively signed ranks*. For the given data, the observed value of S_+ is

$$s_+ = \text{sum of the positive ranks} = 3 + 5 + 6 + 7 + 8 + 9 + 10 = 48$$

When the median of the distribution is much greater than 0, most of the observations with large absolute magnitudes should be positive, resulting in positively signed ranks and a large value of s_+. On the other hand, if the median is 0, magnitudes of positively signed observations should be intermingled with those of negatively signed observations, in which case s_+ will not be very large. (As noted before, this characterization depends on the underlying distribution being symmetric.) Thus we should reject $H_0 \colon \tilde{\mu} = 0$ in favor of $H_a \colon \tilde{\mu} > 0$ when s_+ is "quite large"—the rejection region should have the form $s_+ \geq c$.

The critical value c should be chosen so that the test has a desired significance level (type I error probability), such as .05 or .01. This necessitates finding the distribution of the test statistic S_+ when the null hypothesis is true. Let's consider $n = 5$, in which case there are $2^5 = 32$ ways of applying signs to the five ranks 1, 2, 3, 4, and 5 (each rank could have a − sign or a + sign). The key is that when H_0 is true, *any collection of five signed ranks has the same chance as does any other collection.* That is, the smallest observation in absolute magnitude is equally likely to be positive or negative, the same is true of the second smallest observation in absolute magnitude, and so on. Thus the collection −1, 2, 3, −4, 5 of signed ranks is just as likely as the collection 1, 2, 3, 4, −5, and just as likely as any one of the other 30 possibilities.

Table 14.1 lists the 32 possible signed-rank sequences when $n = 5$ along with the value s_+ for each sequence. This immediately gives the "null distribution" of S_+. For example, Table 14.1 shows that three of the 32 possible sequences have $s_+ = 8$, so $P(S_+ = 8 \text{ when } H_0 \text{ is true}) = 3/32$. This null distribution appears in Table 14.2 and Figure 14.2. Notice that it is symmetric about 7.5; more generally, $S_+ = 8$ is symmetrically distributed over the possible values $0, 1, 2, \ldots, n(n+1)/2$ when H_0 is true. This symmetry will be important in relating the rejection region of lower-tailed and two-tailed tests to that of an upper-tailed test.

Table 14.1 Possible signed-rank sequences for $n = 5$

Sequence					s_+	Sequence					s_+
−1	−2	−3	−4	−5	0	−1	+2	−3	−4	+5	7
+1	−2	−3	−4	−5	1	−1	−2	+3	−4	+5	8
−1	+2	−3	−4	−5	2	+1	+2	−3	−4	+5	8
−1	−2	+3	−4	−5	3	+1	−2	+3	−4	+5	9
+1	+2	−3	−4	−5	3	−1	+2	+3	−4	+5	10
+1	−2	+3	−4	−5	4	+1	+2	+3	−4	+5	11
−1	−2	−3	+4	−5	4	−1	+2	+3	+4	−5	9
+1	−2	−3	+4	−5	5	+1	+2	+3	+4	−5	10
−1	+2	−3	+4	−5	6	−1	−2	−3	+4	+5	9
−1	−2	+3	+4	−5	7	+1	−2	−3	+4	+5	10
+1	+2	−3	+4	−5	7	−1	+2	−3	+4	+5	11
+1	−2	+3	+4	−5	8	−1	−2	+3	+4	+5	12
−1	+2	+3	−4	−5	5	+1	+2	−3	+4	+5	12
+1	+2	+3	−4	−5	6	+1	−2	+3	+4	+5	13
−1	−2	−3	−4	+5	5	−1	+2	+3	+4	+5	14
+1	−2	−3	−4	+5	6	+1	+2	+3	+4	+5	15

Table 14.2 Null distribution of S_+ when $n = 5$

s_+	0	1	2	3	4	5	6	7
$p(s_+)$	1/32	1/32	1/32	2/32	2/32	3/32	3/32	3/32
s_+	8	9	10	11	12	13	14	15
$p(s_+)$	3/32	3/32	3/32	2/32	2/32	1/32	1/32	1/32
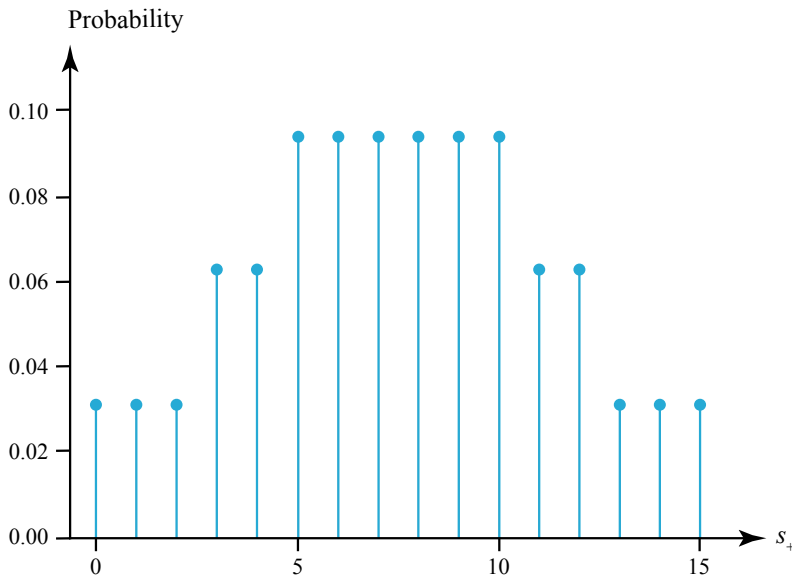
Figure 14.2 Null distribution of S_+ when $n = 5$

For $n = 10$ there are $2^{10} = 1024$ possible signed-rank sequences, so a listing would involve much effort. Each sequence, though, would have probability 1/1024 when H_0 is true, from which the distribution of S_+ when H_0 is true can be obtained.

We are now in a position to determine a rejection region for testing H_0: $\tilde{\mu} = 0$ versus H_a: $\tilde{\mu} > 0$ that has a suitably small significance level α. For the case $n = 5$, consider the rejection region $R = \{s_+ : s_+ \geq 13\} = \{13, 14, 15\}$. Then

$$\alpha = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$
$$= P(S_+ = 13, 14, \text{ or } 15 \text{ when } H_0 \text{ is true})$$
$$= 1/32 + 1/32 + 1/32 = 3/32$$
$$= .094$$

so that $R = \{13, 14, 15\}$ specifies a test with approximate level .1. For the rejection region $\{14, 15\}$, $\alpha = 2/32 = .063$. For the sample $x_1 = .58$, $x_2 = 2.50$, $x_3 = -.21$, $x_4 = 1.23$, $x_5 = .97$, the signed-rank sequence is $-1, +2, +3, +4, +5$, so $s_+ = 14$ and at level .063 (or anything higher) H_0 would be rejected.

The Wilcoxon Signed-Rank Test

Because the underlying distribution is assumed symmetric, $\mu = \tilde{\mu}$, so we will state the hypotheses of interest in terms of μ rather than $\tilde{\mu}$.[1] When the hypothesized value of μ is μ_0, the absolute differences $|x_1 - \mu_0|, \ldots, |x_n - \mu_0|$ must be ranked from smallest to largest.

WILCOXON SIGNED-RANK TEST

Null hypothesis: H_0: $\mu = \mu_0$

Test statistic value: s_+ = the sum of the ranks associated with positive $(x_i - \mu_0)$'s

Alternative Hypothesis	Rejection Region for Level α Test
H_a: $\mu > \mu_0$	$s_+ \geq c_1$
H_a: $\mu < \mu_0$	$s_+ \leq n(n+1)/2 - c_1$
H_a: $\mu \neq \mu_0$	either $s_+ \geq c$ or $s_+ \leq n(n+1)/2 - c$

where the critical values c_1 and c obtained from Appendix Table A.11 satisfy $P(S_+ \geq c_1) \approx \alpha$ and $P(S_+ \geq c) \approx \alpha/2$ when H_0 is true.

Example 14.4 A producer of breakfast cereals wants to verify that a filler machine is operating correctly. The machine is supposed to fill one-pound boxes with 460 g, on average. This is a little above the 453.6 g needed for one pound. When the contents are weighed, it is found that 15 boxes yield the following measurements:

454.4	470.8	447.5	453.2	462.6	445.0	455.9	458.2
461.6	457.3	452.0	464.3	459.2	453.5	465.8	

Does the data provide convincing statistical evidence that the true mean weight differs from 460 g? Let's use the seven-step hypothesis testing procedure outlined earlier in the book.

1. Parameter: μ = the true average weight of all such cereal boxes
2. Hypotheses: H_0: $\mu = 460$ versus H_a: $\mu \neq 460$
3. It is believed that deviations of any magnitude from 460 g are just as likely to be positive as negative (in accord with the symmetry assumption), but the distribution may not be normal. Therefore, the Wilcoxon signed-rank test will be used to see if the filler machine is calibrated correctly.

---

[1] If the tails of the distribution are "too heavy," as is the case with the Cauchy distribution, then μ will not exist. In such cases, the Wilcoxon test will still be valid for tests concerning $\tilde{\mu}$.

4. The test statistic will be s_+ = the sum of the ranks associated with positive $(x_i - 460)$'s
5. From Appendix Table A.11, $P(S_+ \geq 95) = P(S_+ \leq 25) = .024$ when H_0 is true, so the two-tailed test with approximate level .05 rejects H_0 when either $s_+ \geq 95$ or $s_+ \leq 25$ [the exact α is $2(.024) = .048$].
6. Subtracting 460 from each measurement gives

−5.6	10.8	−12.5	−6.8	2.6	−15.0	−4.1	−1.8
1.6	−2.7	−8.0	4.3	−.8	−6.5	5.8	

The ranks are obtained by ordering these from smallest to largest without regard to sign.

Absolute magnitude	.8	1.6	1.8	2.6	2.7	4.1	4.3	5.6	5.8	6.5	6.8	8.0	10.8	12.5	15.0
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Sign	−	+	−	+	−	−	+	−	+	−	−	−	+	−	−

Thus $s_+ = 2 + 4 + 7 + 9 + 13 = 35$.

7. Since $P(S_+ \leq 30)$ is not in the rejection region, it cannot be concluded at level .05 that μ differs from 460. Even at level .094 (approximately .1), H_0 cannot be rejected, since $s_+ P(S_+ \leq 30) = P(S_+ \geq 90) = .047$ implies that s_+ values between 30 and 90 are not significant at that level. The P-value for this test thus exceeds .1. ∎

Although a theoretical implication of the continuity of the underlying distribution is that ties will not occur, in practice they often do because of the discreteness of measuring instruments. If there are several data values with the same absolute magnitude, then they are typically assigned the average of the ranks they would receive if they differed very slightly from one another. For example, if in Example 14.4 $x_8 = 458.2$ were instead 458.4, then two different values of $(x_i - 460)$ would have absolute magnitude 1.6. The ranks to be averaged would be 2 and 3, so each would be assigned rank 2.5.

Large-Sample Distribution of S_+

Figure 14.2 displays the null distribution of S_+ for $n = 5$, a symmetric distribution centered at 7.5. It is straightforward to show (see Exercise 18) that when H_0 is true,

$$E(S_+) = \frac{n(n+1)}{4} \qquad V(S_+) = \frac{n(n+1)(2n+1)}{24}$$

Moreover, when n is not small (say, $n > 20$), Lyapunov's central limit theorem (Chapter 6, Exercise 68) implies that S_+ has an approximately normal distribution. Appendix Table A.11 only presents critical values for the Wilcoxon signed-rank test for $n \leq 20$; beyond that, the test may be performed using the "large-sample" test statistic

$$Z = \frac{S_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

which has approximately a standard normal distribution when H_0 is true.

The Wilcoxon Test for Paired Data

When the data consisted of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ and the differences $D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n$ were normally distributed, in Chapter 10 we used a paired t test for hypotheses about the expected difference μ_D. If normality is not assumed, hypotheses about μ_D can be tested by using the Wilcoxon signed-rank test on the D_i's provided that the distribution of the differences is continuous

and symmetric. The null hypothesis is H_0: $\mu_D = \Delta_0$ for some null value Δ_0 (most frequently 0), and the test statistic S_+ is the sum of the ranks associated with the positive $(D_i - \Delta_0)$'s.

Example 14.5 Poor sleep including insomnia is common among veterans, particularly those with PTSD. The article "Cognitive Behavioral Therapy for Insomnia and Imagery Rehearsal in Combat Veterans with Comorbid Posttraumatic Stress: A Case Series" (*Mil. Behav. Health* 2016: 58–64) reports on a pilot study involving 11 combat veterans diagnosed with both insomnia and PTSD. Each participant attended eight weekly individual therapy sessions with lessons including sleep education, relaxation training, and nightmare re-scripting. Total nightly sleep time (min) was recorded for each veteran both before the 8-week intervention and after. In the accompanying table, differences represent (sleep time after therapy) minus (sleep time before therapy).

Subject	1	2	3	4	5	6	7	8	9	10	11
Before	255	261	257	275	191	528	298	247	314	340	315
After	330	323	312	308	251	559	261	296	387	386	387
Difference	75	62	55	33	60	31	−37	49	73	46	72
Signed rank	11	8	6	2	7	1	−3	5	10	4	9

The relevant hypotheses are H_0: $\mu_D = 0$ versus H_a: $\mu_D > 0$. Appendix Table A.11 shows that for a test with significance level approximately .01, the null hypothesis should be rejected if $s_+ \geq 59$. The test statistic value is $11 + 8 + \cdots + 9 = 63$, the sum of every rank except 3, which falls in the rejection region. We therefore reject H_0 at significance level .01 in favor of the conclusion that the therapy regimen increases sleep time, on average, for this population. Figure 14.3 shows R output for this test, including the test statistic value (as V) and also the corresponding *P*-value, which is $P(S_+ \geq 63 \text{ when } H_0 \text{ is true})$.

```
            wilcoxon signed rank test

data:  After and Before
V = 63, p-value = 0.002441
alternative hypothesis: true location shift is greater than 0
```

Figure 14.3 R output for Example 14.5   ∎

Efficiency of the Sign Test and Signed-Rank Test

When the underlying distribution being sampled is normal, any one of three procedures—the *t* test, the signed-rank test, or the sign test—can be used to test a hypothesis about μ (the point of symmetry). The *t* test is the best test in this case because among all level α tests it is the one having the greatest power (smallest type II error probabilities). On the other hand, neither the *t* test nor the signed-rank test should be applied to data from a clearly skewed distribution, for two reasons. First, lack of normality (resp., symmetry) violates the requirements for the validity of the *t* test (resp., signed-rank test). Second, both of these latter test procedures concern the mean μ, and for a heavily skewed population the mean is arguably of less interest than the median $\tilde{\mu}$.

Test procedure	Population assumption	Parameter of interest	Power (assuming normality)
Sign test	Continuous	$\tilde{\mu}$	Least powerful
Signed-rank test	Symmetric	$\mu = \tilde{\mu}$	
One-sample *t* test	Normal	$\mu = \tilde{\mu}$	Most powerful

Let us now specifically compare Wilcoxon's signed-rank test to the t test. Two questions will be addressed:

1. When the underlying distribution is normal, the "home ground" of the t test, how much is lost by using the signed-rank test?
2. When the underlying distribution is not normal, how much improvement can be achieved by using the signed-rank test?

Unfortunately, there are no simple answers to the two questions. The difficulty is that power for the Wilcoxon test is very difficult to determine for every possible distribution, and the same can be said for the t test when the distribution is not normal. Even if power were easily obtained, any measure of efficiency would clearly depend on which underlying distribution was assumed.

A number of different efficiency measures have been proposed by statisticians; one that many statisticians regard as credible is called **asymptotic relative efficiency** (ARE). The ARE of one test with respect to another is essentially the limiting ratio of sample sizes necessary to obtain identical error probabilities for the two tests. Thus if the ARE of one test with respect to a second equals .5, then when sample sizes are large, twice as large a sample size will be required of the first test to perform as well as the second test. Although the ARE does not characterize test performance for small sample sizes, the following results can be shown to hold:

1. When the underlying distribution is normal, the ARE of the Wilcoxon test with respect to the t test is approximately .95.
2. For any distribution, the ARE will be at least .86 and for many distributions will be much greater than 1.

We can summarize these results by saying that, in large-sample situations, the Wilcoxon test is never very much less efficient than the t test and may be much more efficient if the underlying distribution is far from normal. Although the issue is far from resolved in the case of sample sizes obtained in most practical problems, studies have shown that the Wilcoxon test performs reasonably and is thus a viable alternative to the t test. In contrast, the sign test has ARE less than .64 with respect to the t test when the underlying distribution is normal. (But, again, the sign test is arguably the only appropriate test for heavily skewed populations.)

The Wilcoxon Signed-Rank Interval
In Section 9.6, we discussed the "duality principle" that links hypothesis tests and confidence intervals. Suppose we have a level α test procedure for testing $H_0\colon \theta = \theta_0$ versus $H_a\colon \theta \neq \theta_0$ based on sample data $x_1, \ldots, x_n$. If we let A denote the set of all θ_0 values for which H_0 is *not* rejected, then A is a $100(1 - \alpha)\%$ CI for θ.[2]

The two-tailed Wilcoxon signed-rank test rejects H_0 if s_+ is either $\geq c$ or $\leq n(n + 1)/2 - c$, where c is obtained from Appendix Table A.11 once the desired significance level α is specified. For fixed $x_1, \ldots, x_n$, the $100(1 - \alpha)\%$ **signed-rank interval** will consist of all μ_0 for which $H_0\colon \mu = \mu_0$ is not rejected at level α. To identify this interval, it is convenient to express the test statistic S_+ in another form.

---

[2]There are pathological examples in which the set A is not an interval of θ values, but instead the complement of an interval or something even stranger. To be more precise, we should really replace the notion of a CI with that of a *confidence set*. In the cases of interest here, the set A does turn out to be an interval.

PROPOSITION S_+ = the number of pairwise averages $(X_i + X_j)/2$ with $i \leq j$
that are $\geq \mu_0$
(These pairwise averages are known as **Walsh averages**.)

That is, if we average each x_j in the list with each x_i to its left, including $(x_j + x_j)/2 = x_j$, and count the number of these averages that are $\geq \mu_0$, s_+ results. In moving from left to right in the list of sample values, we are simply averaging every pair of observations in the sample—again, including $(x_j + x_j)/2$—exactly once, so the order in which the observations are listed before averaging is not important. The equivalence of the two methods for computing s_+ is not difficult to verify. The number of pairwise averages is $\binom{n}{2} + n = n(n+1)/2$. If either too many or too few of these pairwise averages are $\geq \mu_0$, H_0 is rejected.

Example 14.6 The following observations are values of cerebral metabolic rate for rhesus monkeys: $x_1 = 4.51$, $x_2 = 4.59$, $x_3 = 4.90$, $x_4 = 4.93$, $x_5 = 6.80$, $x_6 = 5.08$, $x_7 = 5.67$. The 28 pairwise averages are, in increasing order,

4.51	4.55	4.59	4.705	4.72	4.745	4.76	4.795	4.835	4.90
4.915	4.93	4.99	5.005	5.08	5.09	5.13	5.285	5.30	5.375
5.655	5.67	5.695	5.85	5.865	5.94	6.235	6.80		

The first few and the last few of these are pictured on a measurement axis in Figure 14.4.
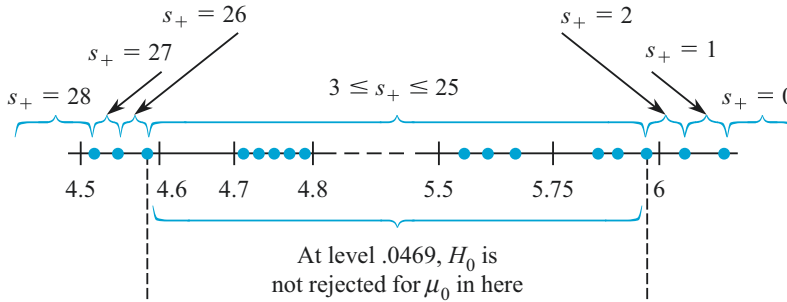
Figure 14.4 Plot of the data for Example 14.6

Because of the discreteness of the distribution of S_+, $\alpha = .05$ cannot be obtained exactly. The rejection region {0, 1, 2, 26, 27, 28} has $\alpha = .046$, which is as close as possible to .05, so the level is approximately .05. Thus if the number of pairwise averages $\geq \mu_0$ is between 3 and 25, inclusive, H_0 is not rejected. As displayed in Figure 14.4, the approximate 95% CI for μ is (4.59, 5.94); the endpoints are the 3rd-lowest and 3rd-highest (3rd and 26th ordered) Walsh averages. ∎

In general, once the pairwise averages are ordered from smallest to largest, the endpoints of the Wilcoxon interval are two of the "extreme" averages. To express this precisely, let the smallest pairwise average be denoted by $\bar{x}_{(1)}$, the next smallest by $\bar{x}_{(2)}, \ldots,$ and the largest by $\bar{x}_{(n(n+1)/2)}$.

PROPOSITION   If the level α Wilcoxon signed-rank test for $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$ is to reject H_0 if either $s_+ \geq c$ or $s_+ \leq n(n+1)/2 - c$, then a $100(1 - \alpha)\%$ CI for μ is

$$\left(\bar{x}_{(n(n+1)/2 - c + 1)}, \bar{x}_{(c)} \right) \tag{14.2}$$

In words, the interval extends from the kth smallest pairwise average to the kth largest average, where $k = n(n+1)/2 - c + 1$. Appendix Table A.12 gives the values of c that correspond to the usual confidence levels for $n = 5, 6, \ldots, 25$. In R, the `wilcox.test` function applied to a vector x containing the sample data will return this signed-rank interval if the user includes the option `conf.int = T`.

Example 14.7 (Example 14.6 continued) For $n = 7$, an 89.1% interval (approximately 90%) is obtained by using $c = 24$, since the rejection region {0, 1, 2, 3, 4, 24, 25, 26, 27, 28} has $\alpha = .109$. The interval is $(\bar{x}_{(28-24+1)}, \bar{x}_{(24)}) = (\bar{x}_{(5)}, \bar{x}_{(24)}) = (4.72, 5.85)$, which extends from the fifth smallest to the fifth largest pairwise average. ∎

The derivation of the signed-rank interval depended on having a single sample from a continuous symmetric distribution with mean (median) μ. When the data is paired, the interval constructed from the Walsh averages of the differences $d_1, d_2, \ldots, d_n$ is a CI for the mean (median) difference μ_D.

For $n > 20$, the large-sample approximation to the Wilcoxon test based on standardizing S_+ gives an approximation to c in (14.2). The result for a $100(1 - \alpha)\%$ interval is

$$c \approx \frac{n(n+1)}{4} + z_{\alpha/2} \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

The efficiency of the Wilcoxon interval relative to the t interval is roughly the same as that for the Wilcoxon test relative to the t test. In particular, for large samples when the underlying population is normal, the Wilcoxon interval will tend to be slightly longer than the t interval, but if the population is quite nonnormal (e.g., symmetric but with heavy tails), then the Wilcoxon interval will tend to be much shorter than the t interval.

Exercises: Section 14.2 (11–24)

11. Reconsider the situation described in Exercise 34(a) of Section 9.2, and use the Wilcoxon test with $\alpha = .05$ to test the specified hypotheses.

12. Use the Wilcoxon test to analyze the data given in Example 9.12.

13. The following pH measurements at a proposed water intake site appear in the 2011 report "Sacramento River Water Quality Assessment for the Davis-Woodland Water Supply Project":

    7.20   7.24   7.31   7.38   7.45   7.60   7.86

    Use the Wilcoxon signed-rank test to determine whether the true mean pH level at this site exceeds 7.3 with significance level .05.

14. A random sample of 15 automobile mechanics certified to work on a certain type of car was selected, and the time (in minutes) necessary for each one to diagnose a particular problem was determined, resulting in the following data:

    30.6   30.1   15.6   26.7   27.1   25.4   35.0   30.8
    31.9   53.2   12.5   23.2   8.8   24.9   30.2

    Use the Wilcoxon test at significance level .10 to decide whether the data suggests that true average diagnostic time is less than 30 min.

15. Both a gravimetric and a spectrophotometric method are under consideration for determining phosphate content of a

particular material. Twelve samples of the material are obtained, each is split in half, and a determination is made on each half using one of the two methods, resulting in the following data:

Sample	1	2	3	4
Grav.	54.7	58.5	66.8	46.1
Spec.	55.0	55.7	62.9	45.5
Sample	5	6	7	8
Grav.	52.3	74.3	92.5	40.2
Spec.	51.1	75.4	89.6	38.4
Sample	9	10	11	12
Grav.	87.3	74.8	63.2	68.5
Spec.	86.8	72.5	62.3	66.0

Use the Wilcoxon test to decide whether one technique gives on average a different value than the other technique for this type of material.

16. Fifty-three participants performed a series of tests in which a small "zap" was delivered to one compass point, selected at random, on a joystick in their hand. In one setting, subjects were told to move the joystick in the same direction of the zap; in another setting, they were told to move the joystick in the direction opposite to the zap. A series of trials was performed under each setting, and the number of correct moves under both settings was recorded. ("An Experimental Setup to Test Dual-Joystick Directional Responses to Vibrotactile Stimuli," *IEEE Trans. on Haptics* 2018.)

a. The authors performed a Wilcoxon signed-rank test on the paired differences (number correct in same direction minus number correct in opposite direction, which is discrete but won't greatly impact the analysis). The resulting test statistic value was $s_+ = 695$. Test H_0: $\mu_D = 0$ versus H_a: $\mu_D \neq 0$ at the .10 significance level using the large-sample version of the test.

b. The same article also explored whether participants would make more correct moves if the "zaps" were instead delivered by a glove they wore while grasping the joystick. Again for $n = 53$ subjects, the difference (number correct

with joystick minus number correct with glove) was computed, and the signed-rank test statistic value was $s_+ = 136$. Perform a two-sided large-sample test, and interpret your findings.

17. The article "Punishers Benefit from Third-Party Punishment in Fish" (*Science*, 8 Jan 2010: 171) describes an experiment meant to simulate behavior of cleaner fish, so named because they eat parasites off of "client" fish but will sometimes take a bite of the client's mucus instead. (Cleaner fish prefer the mucus over the parasites.) Eight female cleaner fish were provided bits of prawn (preferred food) and fish-flake (less preferred), then a male cleaner fish chased them away. One minute later, the process was repeated.

a. The following data on the amount of prawn eaten by each female in the two rounds is consistent with information in the article. Use Wilcoxon's signed-rank test to determine whether female fish eat less of their preferred food, on average, after having been chased by a male.

Female	1	2	3	4	5	6	7	8
1st trial	.207	.215	.103	.182	.282	.228	.152	.293
2nd trial	.164	.033	.092	.003	.115	.250	.056	.247

b. The researchers recorded the same information on the male cleaner fish (the chasers), resulting in a signed-rank test statistic value of $s_+ = 28$. Does this provide evidence that males *increase* their average preferred food consumption the second time around?

18. The signed-rank statistic can be represented as $S_+ = 1 \cdot U_1 + 2 \cdot U_2 + \cdots + n \cdot U_n$ where $U_i = 1$ if the sign of the $(x_i - \mu_0)$ with the ith largest absolute magnitude is positive (in which case i is included in S_+) and $U_i = 0$ if this value is negative ($i = 1, 2, 3, \ldots, n$). Furthermore, when H_0 is true, the U_i's are independent Bernoulli rvs with $p = .5$.

a. Use this representation to obtain the mean and variance of S_+ when H_0 is true. [*Hint*: The sum of the first n positive integers is $n(n+1)/2$, and the sum

of the squares of the first n positive integers is $n(n+1)(2n+1)/6$.]

b. A particular type of steel beam has been designed to have a compressive strength (lb/in^2) of at least 50,000. An experimenter obtained a random sample of 25 beams and determined the strength of each one, resulting in the following data (expressed as deviations from 50,000):

−10	−27	36	−55	73	−77	−81
90	−95	−99	113	−127	−129	136
−150	−155	−159	165	−178	−183	−192
−199	−212	−217	−229			

Carry out a test using a significance level of approximately .01 to see if there is strong evidence that the design condition has been violated.

19. Reconsider the calorie-burning data in Exercise 10 from Section 14.1.

   a. Utilize the Wilcoxon signed-rank procedure to test H_0: $\mu_D = 0$ versus H_a: $\mu_D < 0$ for the population of REE differences (IF minus standard diet). What assumption is required here that was not necessary for the sign test?

   b. Now apply the paired t test to the hypotheses in part (a). What extra assumptions are required?

   c. Compare the results of the three tests (sign, signed-rank, and paired t) and discuss what you find.

20. Suppose that observations $X_1, X_2, \ldots, X_n$ are made on a process at times 1, 2, …, n. On the basis of this data, we wish to test
H_0: the X_i's constitute an independent and identically distributed sequence
versus
H_a: X_{i+1} tends to be larger than X_i for $i = 1$, …, n (an increasing trend)

Suppose the X_i's are ranked from 1 to n. Then when H_a is true, larger ranks tend to occur later in the sequence, whereas if H_0 is true, large and small ranks tend to be mixed together. Let R_i be the rank of X_i and consider the test statistic $D = \sum_{i=1}^{n} (R_i - i)^2$. Then small values of D give support to H_a (e.g., the smallest value is 0 for $R_1 = 1, R_2 = 2, \ldots, R_n = n$), so H_0 should be rejected in favor of H_a if $d \le c$. When H_0 is true, any sequence of ranks has probability $1/n!$. Use this to find c for which the test has a level as close to .10 as possible in the case $n = 4$. [Hint: List the 4! rank sequences, compute d for each one, and then obtain the null distribution of D. See the Lehmann book in the bibliography for more information.]

21. Obtain the 99% signed-rank interval for true average pH using the data in Exercise 13.

22. Obtain a 95% signed-rank interval for true average diagnostic time using the data in Exercise 14. [Hint: Try to compute only those pairwise averages having relatively small or large values, rather than all 120 averages.]

23. Obtain a CI for μ_D of Exercise 17 using the data given there; your confidence level should be roughly 95%.

24. The following observations are copper contents (%) for a sample of Bidri artifacts (a type of ancient Indian metal handicraft) at the Victoria and Albert Museum in London ("Enigmas of Bidri," *Surface Engr.* 2005: 333–339): 2.4, 2.7, 5.3, and 10.1. What confidence levels are achievable for this sample size using the signed-rank interval? Select an appropriate confidence level and compute the interval.

14.3 Two-Sample Rank-Based Inference

When at least one of the sample sizes in a two-sample problem is small, the t test requires the assumption of normality (at least approximately). There are situations, though, in which an investigator would want to use a test that is valid even if the underlying distributions are quite nonnormal. We now describe such a test, called the **Wilcoxon rank-sum test**. An alternative name for the procedure is the **Mann–Whitney test**, although the Mann–Whitney test statistic is sometimes

expressed slightly differently from that of the Wilcoxon test. The Wilcoxon test procedure is "distribution-free" because it will have the desired level of significance for a very large collection of underlying distributions rather than just the normal distribution.

ASSUMPTIONS $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are two independent random samples from continuous distributions with means μ_1 and μ_2, respectively. The X and Y distributions have the same shape and spread, the only possible difference between the two being in the values of μ_1 and μ_2.

When $H_0: \mu_1 - \mu_2 = \Delta_0$ is true, the X distribution is shifted by the amount Δ_0 to the right of the Y distribution; i.e., $f_X(x) = f_Y(x - \Delta_0)$. When H_0 is false, the shift is by an amount other than Δ_0; note, though, that we still assume the two distributions only differ by a shift in means. This assumption can be difficult to verify in practice, but the Wilcoxon rank-sum test is nonetheless a popular approach to comparisons based on small samples.

A Rank-Based Test Statistic
Let's first test $H_0: \mu_1 - \mu_2 = 0$; then the X's and Y's are identically distributed when H_0 is true. Consider the case $n_1 = 3$, $n_2 = 4$. Denote the observations by x_1, x_2, and x_3 (the first sample) and y_1, y_2, y_3, and y_4 (the second sample). If μ_1 is actually much larger than μ_2, then most of the observed x's will be larger than the observed y's. However, if H_0 is true, then the values from the two samples should be intermingled. The test statistic will quantify how much intermingling there is in the two samples.

To begin, pool the x's and y's into a single combined sample of size $m + n = 7$ and rank these observations from smallest to largest, with the smallest receiving rank 1 and the largest, rank 7. If most of the large ranks (or most of the small ranks) were associated with x observations, we would begin to doubt H_0. This suggests the test statistic

W = the sum of the ranks in the combined sample associated with X observations   (14.3)

For the values of m and n under consideration, the smallest possible value of W is $1 + 2 + 3 = 6$ (if all three x's are smaller than all four y's), and the largest possible value is $5 + 6 + 7 = 18$ (if all three x's are larger than all four y's).

As an example, suppose $x_1 = -3.10$, $x_2 = 1.67$, $x_3 = 2.01$, $y_1 = 5.27$, $y_2 = 1.89$, $y_3 = 3.86$, and $y_4 = .19$. Then the pooled ordered sample is $-3.10, .19, 1.67, 1.89, 2.01, 3.86$, and 5.27. The X ranks for this sample are 1 (for -3.10), 3 (for 1.67), and 5 (for 2.01), giving $w = 1 + 3 + 5 = 9$.

The test procedure based on the statistic (14.3) requires knowledge of the null distribution of W. When H_0 is true, all seven observations come from the same population. This means that under H_0, any possible triple of ranks associated with the three x's—such as (1, 4, 5), (3, 5, 6), or (5, 6, 7)— has the same probability as any other possible rank triple. Since there are $\binom{7}{3} = 35$ possible rank triples, under H_0 each rank triple has probability $1/35$. From a list of all 35 rank triples and the W value associated with each, the null distribution of W can immediately be determined. For example, there are four rank triples for which $W = 11$—(1, 3, 7), (1, 4, 6), (2, 3, 6), and (2, 4, 5)—so $P(W = 11) = 4/35$. The complete sampling distribution appears in Table 14.3 and Figure 14.5.

Table 14.3 Probability distribution of W when H_0 is true ($n_1 = 3$, $n_2 = 4$)

w	6	7	8	9	10	11	12	13	14	15	16	17	18
$p(w)$	1/35	1/35	2/35	3/35	4/35	4/35	5/35	4/35	4/35	3/35	2/35	1/35	1/35
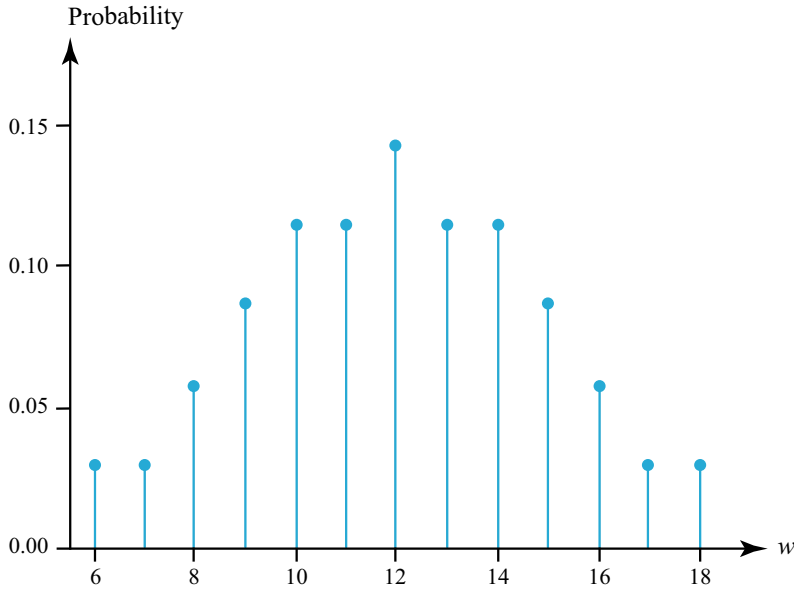
Figure 14.5 Sampling distribution of W when H_0 is true ($n_1 = 3$, $n_2 = 4$)

Suppose we wished to test H_0: $\mu_1 - \mu_2 = 0$ against H_a: $\mu_1 - \mu_2 < 0$ based on the example data given previously, for which the observed value of W was 9. Then the P-value associated with the test is the chance of observing a W-value of 9 *or lower*, assuming H_0 is true. Using Table 14.3,

$$P\text{-value} = P(W \leq 9 \text{ when } H_0 \text{ is true}) = P(W = 6, 7, 8, 9) = \frac{7}{35} = .2$$

We would thus not reject H_0 at any reasonable significance level.

Constructing the null sampling distribution of W manually can be tedious, since there are generally $\binom{m+n}{n}$ possible arrangements of ranks to consider. Software will provide w and the associated P-value quickly, though various packages perform the P-value calculation slightly differently.

The Wilcoxon Rank-Sum Test

The null hypothesis H_0: $\mu_1 - \mu_2 = \Delta_0$ is handled by subtracting Δ_0 from each X_i and using the $(X_i - \Delta_0)$'s as the X_i's were previously used. The smallest possible value of the statistic W is $1 + 2 + \cdots + m = m(m+1)/2$, which occurs when the $(X_i - \Delta_0)$'s are all to the left of the Y sample. The largest possible value of W occurs when the $(X_i - \Delta_0)$'s lie entirely to the right of the Y's; in this case, $W = (n+1) + \cdots + (m+n) =$ (sum of first $m+n$ integers)$-$ (sum of first n integers), which gives $m(m+2n+1)/2$. As with the special case $m = 3$, $n = 4$, the null distribution of W is symmetric about the value that is halfway between the smallest and largest values; this middle value is $m(m+n+1)/2$. Because of this symmetry, probabilities involving lower-tail critical values can be obtained from corresponding upper-tail values.

WILCOXON RANK-SUM TEST

Null hypothesis: H_0: $\mu_1 - \mu_2 = \Delta_0$

Test statistic value: $w = \sum_{i=1}^{m} r_i$,

where r_i = rank of $(x_i - \Delta_0)$ in the combined sample
of $m + n$ $(x - \Delta_0)$'s and y's

Alternative Hypothesis	Rejection Region for Level α Test
H_a: $\mu_1 - \mu_2 > \Delta_0$	$w \geq c_1$
H_a: $\mu_1 - \mu_2 < \Delta_0$	$w \leq m(m+n+1) - c_1$
H_a: $\mu_1 - \mu_2 \neq \Delta_0$	either $w \geq c$ or $w \leq m(m+n+1) - c$

where $P(W \geq c_1$ when H_0 is true$) \approx \alpha$, $P(W \geq c$ when H_0 is true$) \approx \alpha/2$

Because W has a discrete probability distribution, there will not usually exist a critical value corresponding exactly to one of the usual significance levels. Appendix Table A.13 gives upper-tail critical values for probabilities closest to .05, .025, .01, and .005, from which level .05 or .01 one- and two-tailed tests can be obtained. The table gives information only for $3 \leq m \leq n \leq 8$. To use the table, *the X and Y samples should be labeled so that $m \leq n$.* Ties are handled as suggested for the signed-rank test in the previous section.

Example 14.8   The urinary fluoride concentration (parts per million) was measured both for a sample of livestock grazing in an area previously exposed to fluoride pollution and for a similar sample grazing in an unpolluted region:

Polluted	21.3 [11]	18.7 [7]	23.0 [12]	17.1 [3]	16.8 [2]	20.9 [10]	19.7 [8]
Unpolluted	14.2 [1]	18.3 [5]	17.2 [4]	18.4 [6]	20.0 [9]		

The values in brackets indicate the rank of each observation in the combined sample of 12 values. Does the data indicate strongly that the true average fluoride concentration for livestock grazing in the polluted region is larger than for the unpolluted region? Let's use the Wilcoxon rank-sum test at level $\alpha = .01$.

1. The sample sizes here are 7 and 5. To obtain $m \leq n$, label the unpolluted observations as the x's ($x_1 = 14.2, \ldots, x_5 = 20.0$) and the polluted observations as the y's. Thus the parameters are

   μ_1 = the true average fluoride concentration *without* pollution
   μ_2 = the true average concentration *with* pollution

2. The hypotheses are

   H_0: $\mu_1 - \mu_2 = 0$
   H_a: $\mu_1 - \mu_2 < 0$ (pollution is associated with an increase in concentration)

3. In order to perform the Wilcoxon rank-sum test, we will assume that the fluoride concentration distributions for these two livestock populations have the same shape and spread, but possibly differ in mean.

4. The test statistic value is $w = \sum_{i=1}^{5} r_i$, where r_i = the rank of x_i among all 12 observations.

5. From Appendix Table A.13 with $m = 5$ and $n = 7$, $P(W \geq 47$ when H_0 is true$) \approx .01$. The critical value for the lower-tailed test is therefore $m(m + n + 1) - 47 = 5(13) - 47 = 18$; H_0 will now be rejected if $w \leq 18$.

6. The computed W is $w = r_1 + r_2 + \cdots + r_5 = 1 + 5 + 4 + 6 + 9 = 25$.
7. Since 25 is not ≤ 18, H_0 is not rejected at (approximately) level .01. The data does not provide convincing statistical evidence at the .01 significance level that average fluoride concentration is higher among livestock grazing in the polluted region.   ∎

Alternative Versions of the Rank-Sum Test

Appendix Table A.13 allows us to perform the Wilcoxon rank-sum test provided that m and n are both ≤ 8. For larger sample sizes, a central limit theorem for nonindependent variables can be used to show that W has an approximately normal distribution. (The genesis of a bell-shaped curve can even be seen in Figure 14.5 where $m = 3$ and $n = 4$.) When H_0 is true, the mean and variance of W (see Exercise 32) are

$$E(W) = \frac{m(m+n+1)}{2} \qquad V(W) = \frac{mn(m+n+1)}{12}$$

These suggest that when $m > 8$ and $n > 8$ the rank-sum test may be performed using the test statistic

$$Z = \frac{W - m(m+n+1)/2}{\sqrt{mn(m+n+1)/12}}$$

which has approximately a standard normal distribution when H_0 is true.

Some statistical software packages, including R, use an alternative formulation called the *Mann–Whitney U test*. Consider all possible pairs (X_i, Y_j), of which there are mn. Define a test statistic U by

$$U = \text{the number of} \left(X_i, Y_j\right) \text{ pairs for which } X_i - Y_j > \Delta_0 \qquad (14.4)$$

It can be shown (Exercise 34) that the test statistics U and W are related by $U = W - m(m+1)/2$. The mean and variance of U can thus be obtained from the corresponding expressions for W, and the normal approximation to W applies equally to U.

Finally, when using the normal approximation, some slightly tedious algebra can be used to rearrange the standardized version of W so that it looks similar to the two-sample z test statistic:

$$Z = \frac{W - E(W)}{\sqrt{V(W)}} = \cdots = \frac{\bar{R}_1 - \bar{R}_2}{\sqrt{\dfrac{\sigma^2}{m} + \dfrac{\sigma^2}{n}}},$$

where $\bar{R}_1$ and $\bar{R}_2$ denote the average ranks for the two samples, and $\sigma^2 = (m+n)(m+n+1)/12$.

Efficiency of the Wilcoxon Rank-Sum Test

When the distributions being sampled are both normal with $\sigma_1 = \sigma_2$ and therefore have the same shapes and spreads, either the *pooled t* test or the Wilcoxon test can be used. (The two-sample t test assumes normality but not equal standard deviations, so assumptions underlying its use are more restrictive in one sense and less in another than those for Wilcoxon's test.) In this situation, the pooled t test is best among all possible tests in the sense of maximizing power for any fixed α. However, an investigator can never be absolutely certain that underlying assumptions are satisfied. It is therefore relevant to ask (1) how much is lost by using Wilcoxon's test rather than the pooled t test when the distributions are normal with equal variances and (2) how W compares to T in nonnormal situations.

The notion of asymptotic relative efficiency was discussed in the previous section in connection with the one-sample t test and Wilcoxon signed-rank test. The results for the two-sample tests are the same as those for the one-sample tests. When normality and equal variances both hold, the rank-sum

test is approximately 95% as efficient as the pooled t test in large samples. That is, the t test will give the same error probabilities as the Wilcoxon test using slightly smaller sample sizes. On the other hand, the Wilcoxon test will always be at least 86% as efficient as the pooled t test and may be much more efficient if the underlying distributions are very nonnormal, especially with heavy tails. The comparison of the Wilcoxon test with the two-sample (unpooled) t test is less clear-cut. The t test is not known to be the best test in any sense, so it seems safe to conclude that as long as the population distributions have similar shapes and spreads, the behavior of the Wilcoxon test should compare quite favorably to the two-sample t test.

Lastly, we note that power calculations for the Wilcoxon test are quite difficult. This is because the distribution of W when H_0 is false depends not only on $\mu_1 - \mu_2$ but also on the shape of the two distributions. For most underlying distributions, the nonnull distribution of W is virtually intractable. This is why statisticians developed asymptotic relative efficiency as a means of comparing tests. With the capabilities of modern-day computer software, another approach to power calculations is to carry out a simulation experiment.

The Wilcoxon Rank-Sum Interval

Similar to the signed-rank interval of Section 14.2, a CI for $\mu_1 - \mu_2$ based on the Wilcoxon rank-sum test is obtained by determining, for fixed x_i's and y_j's, the set of all Δ_0 values for which $H_0: \mu_1 - \mu_2 = \Delta_0$ is not rejected. This is easiest to do if we use the Mann–Whitney U statistic (14.4), according to which H_0 should be rejected if the number of $(x_i - y_j)$'s $\geq \Delta_0$ is either too small or too large.

This, in turn, suggests that we compute $x_i - y_j$ for each i and j and order these mn differences from smallest to largest. Then if the null value Δ_0 is neither smaller than most of the differences nor larger than most, $H_0: \mu_1 - \mu_2 = \Delta_0$ is not rejected. Varying Δ_0 now shows that a CI for $\mu_1 - \mu_2$ will have as its lower endpoint one of the ordered $(x_i - y_j)$'s, and similarly for the upper endpoint.

PROPOSITION   Let $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ be the observed values in two independent samples from continuous distributions that differ only in location (and not in shape or spread). With $d_{ij} = x_i - y_j$ and the ordered differences denoted by $d_{ij(1)}, d_{ij(2)}, \ldots, d_{ij(mn)}$, the general form of a $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$\left(d_{ij(mn-c+1)}, d_{ij(c)} \right) \tag{14.5}$$

where c is the critical value for the two-tailed level α Wilcoxon rank-sum test.

Notice that the form of the Wilcoxon rank-sum interval (14.5) is very similar to the Wilcoxon signed-rank interval (14.2); (14.2) uses pairwise averages from a single sample, whereas (14.5) uses pairwise differences from two samples. Appendix Table A.14 gives values of c for selected values of m and n. In R, the `wilcox.test` function applied to vectors `x` and `y` containing the sample data will return this signed-rank interval if the user includes the option `conf.int = T`.

Example 14.9 The article "Some Mechanical Properties of Impregnated Bark Board" (*Forest Products J.*) reports the following data on maximum crushing strength (psi) for a sample of epoxy-impregnated bark board and for a sample of bark board impregnated with another polymer:

Epoxy (*x*'s)	10,860	11,120	11,340	12,130	14,380	13,070
Other (*y*'s)	4590	4850	6510	5640	6390	

Let's obtain a 95% CI for the true average difference in crushing strength between the epoxy-impregnated board and the other type of board.

From Appendix Table A.14, since the smaller sample size is 5 and the larger sample size is 6, $c = 26$ for a confidence level of approximately 95%, and $mn - c + 1 = (5)(6) - 26 + 1 = 5$. All 30 d_{ij}'s appear in Table 14.4. The five smallest d_{ij}'s are $d_{ij(1)} = 4350$, 4470, 4610, 4730, and $d_{ij(5)} = 4830$; and the five largest d_{ij}'s are (in descending order) 9790, 9530, 8740, 8480, and 8220. Thus the CI is $(d_{ij(5)}, d_{ij(26)}) = (4830, 8220)$.

Table 14.4 Differences (d_{ij}) for the rank-sum interval in Example 14.9

		y_j				
		4590	4850	5640	6390	6510
	10,860	6270	6010	5220	4470	4350
	11,120	6530	6270	5480	4730	4610
x_i	11,340	6750	6490	5700	4950	4830
	12,130	7540	7280	6490	5740	5620
	13,070	8480	8220	7430	6680	6560
	14,380	9790	9530	8740	7990	7870

∎

When m and n are both large, the aforementioned normal approximation can be used to derive a large-sample approximation for the value c in interval (14.5). The result is

$$c \approx \frac{mn}{2} + z_{\alpha/2}\sqrt{\frac{mn(m+n+1)}{12}}$$

As with the signed-rank interval, the rank-sum interval (14.5) is quite efficient with respect to the t interval; in large samples, (14.5) will tend to be only a bit longer than the t interval when the underlying populations are normal and may be considerably shorter than the t interval if the underlying populations have heavier tails than do normal populations. And once again, the actual confidence level for the t interval may be quite different from the nominal level in the presence of substantial nonnormality.

Exercises: Section 14.3 (25–36)

25. In an experiment to compare the bond strength of two different adhesives, each adhesive was used in five bondings of two surfaces, and the force necessary to separate the surfaces was determined for each bonding. For adhesive 1, the resulting values were 229, 286, 245, 299, and 250, whereas the adhesive 2 observations were 213, 179, 163, 247, and 225. Let μ_i denote the true average bond strength of adhesive type i. Use the Wilcoxon rank-sum test at level .05 to test $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 > \mu_2$.

26. The accompanying data shows the alcohol content (percent) for random samples of 7 German beers and 8 domestic beers. Does the data suggest that German beers have a different average alcohol content than those brewed in the USA? Use the Wilcoxon rank-sum test at $\alpha = .05$.

German	5.00	4.90	3.80	4.82	4.80	5.44	6.60	
Domestic	4.85	5.04	4.20	4.10	4.50	4.70	4.30	5.50

27. A modification has been made to one assembly line for a particular automobile chassis. Because the modification involves extra cost, it will be implemented throughout all lines only if sample data strongly indicates that the modification has decreased true average assembly time by more than 1 h. Assuming that the assembly time distributions differ only with respect to location if at all, use the Wilcoxon rank-sum test at level .05 on the accompanying

data (also in hours) to test the appropriate hypotheses.

Original process	8.6	5.1	4.5	5.4	6.3	6.6	5.7	8.5
Modified process	5.5	4.0	3.8	6.0	5.8	4.9	7.0	5.7

28. Can video games improve balance among the elderly? The article "The Effect of Virtual Reality Gaming on Dynamic Balance in Older Adults" (*Age and Ageing* 2012: 549–552) reported an experiment in which 34 senior citizens were randomly assigned to one of two groups: (1) 16 who engaged in a six-week exercise regimen using the Wii Fit Balance Board (WBB) and (2) 18 who were told not to vary their daily physical activity during that interval. The accompanying data on improvement in 8-foot-up-and-go time (sec), a standard test of agility and balance, is consistent with information in the article. Test whether the true average improvement is greater using the WBB than under control conditions at the .05 significance level.

WBB	−1.9	−0.8	0.1	0.5	0.6	0.7	0.8	0.9
	1.1	1.2	1.5	2.0	2.1	2.7	3.2	3.7
Control	−2.6	−2.2	−2.1	−1.8	−1.4	−1.1	−0.7	−0.6
	−0.3	−0.1	0.0	0.3	0.4	1.0	1.3	2.3
	2.4	4.5						

29. Reconsider the situation described in Exercise 110 of Chapter 10 and the following Minitab output (the Greek letter eta is used to denote a median).

```
Mann-Whitney Confidence Interval and
Test
good         N = 8        Median = 0.540
poor         N = 8        Median = 2.400
Point estimate for ETA1 − ETA2 is −1.155
95.9% CI for ETA1 − ETA2 is
(−3.160 − 0.409) W = 41.0
Test of ETA1 = ETA2 versus ETA1 < ETA2 is
significant at 0.0027
```

   a. Verify that Minitab's test statistic value is correct.

   b. Carry out an appropriate test of hypotheses using a significance level of .01.

30. The article "Opioid Use and Storage Patterns by Patients after Hospital Discharge following Surgery" (*PLoS ONE* 2016) reported a study of 30 women who just had Caesarian sections. The women were classified into two groups: 14 with a high need for pain medicine post-surgery and 16 with a low need. The total oral morphine equivalent prescribed at discharge was determined for each woman, and the resulting Wilcoxon rank-sum test statistic was $W = 249.5$. (The .5 comes from ties in the data, but that won't affect the P-value much.) Test the hypotheses $H_0: \mu_1 = \mu_2$ versus $H_a: \mu_1 \neq \mu_2$ at the .05 significance level. Does it appear that physicians prescribe opioids in proportion to patients' pain-control needs?

31. The article "Mutational Landscape Determines Sensitivity to PD-1 Blockade in Non-Small Cell Lung Cancer" (*Science*, 3 April 2015) described a study of 16 cancer patients taking the drug Keytruda. For each patient, the number of nonsynonymous mutations per tumor was determined; higher numbers indicate better drug effectiveness. The data is separated into patients that showed a durable clinical benefit (partial or stable response lasting >6 months) and those with no durable benefit. Use the methods of this section to determine whether patients experiencing durable clinical benefit tend to have a higher average number of nonsynonymous mutations than those with no durable benefit (use $\alpha = .05$).

Durable benefit	170	228	300	302	315	490	774	
No durable benefit	11	28	46	115	148	161	180	300	625

32. The Wilcoxon rank-sum statistic can be represented as $W = R_1 + R_2 + \cdots + R_m$, where R_i is the rank of $X_i - \Delta_0$ among all $m + n$ such differences. When H_0 is true, each R_i is equally likely to be one of the first $m + n$ positive integers; that is, R_i has a discrete uniform distribution on the values 1, 2, 3, …, $m + n$.

a. Determine the mean value of each R_i when H_0 is true and then show that the mean value of W is $m(m + n + 1)/2$. [*Hint*: The sum of the first k positive integers is $k(k + 1)/2$.]

b. The variance of each R_i is easily determined. However, the R_i's are not independent random variables because, for example, if $m = n = 10$ and we are told that $R_1 = 5$, then R_2 must be one of the *other* 19 integers between 1 and 20. However, if a and b are any two distinct positive integers between 1 and $m + n$ inclusive, it follows that $P(R_i = a$ and $R_j = b) = 1/[(m+n)(m+n-1)]$ since two integers are being sampled without replacement from among 1, 2, ..., $m + n$. Use this fact to show that $\text{Cov}(R_i, R_j) = -(m+n+1)/12$, and then show that the variance of W is $mn(m+n+1)/12$.

33. The article "Controlled Clinical Trial of Canine Therapy Versus Usual Care to Reduce Patient Anxiety in the Emergency Department" (*PLoS ONE* 2019) reported on an experiment in which 80 adult hospital patients were randomly assigned to either 15 min with a certified therapy dog ($m = 40$) or usual care ($n = 40$). Each patient's change in self-reported pain, depression, and anxiety (pre-treatment minus post treatment) was recorded. The researchers employed a rank-sum test to compare the two treatment groups on each of these three outcomes; the resulting test statistic values appear below.

Change in:	Pain	Depression	Anxiety
$w =$	1475	1316	1171

Test the hypotheses H_0: $\mu_1 - \mu_2 = 0$ against H_a: $\mu_1 - \mu_2 < 0$ (1 = dog therapy treatment, 2 = control) for each of the three response variables at the .01 significance level. What can be said about the chance of committing at least one type I error among the three tests?

34. Refer to Exercise 32. Sort the ranks of the X_i's, so that $R_1 < R_2 < \cdots < R_m$.

a. In terms of the R's, how many of the Y_j's are less than the smallest X_i? Less than the second-smallest X_i?

b. When $\Delta_0 = 0$, the Mann–Whitney test statistic is $U =$ the number of (X_i, Y_j) pairs for which $X_i > Y_j$. Use part (a) to express U as a sum, then show this sum is equal to $W - m(m+1)/2$.

c. Use the mean and variance of W to determine $E(U)$ and $V(U)$.

35. Obtain the 90% rank-sum CI for $\mu_1 - \mu_2$ using the data in Exercise 25.

36. Obtain a 95% CI for $\mu_1 - \mu_2$ using the data in Exercise 27. Is your interval consistent with the result of the hypothesis test in that exercise?

14.4   Nonparametric ANOVA

The analysis of variance (ANOVA) procedures in Chapter 11 for comparing I population or treatment means assumed that every population/treatment distribution is normal with the same standard deviation, so that the only potential difference is their means $\mu_1, \ldots, \mu_I$. Here we present methods for testing equality of the μ_i's that apply to a broader class of population distributions.

The Kruskal–Wallis Test

The Kruskal–Wallis test extends the Wilcoxon rank-sum test of the previous section to the case of three or more populations or treatments. The I population/treatment distributions under consideration are assumed to be continuous, have the same shape and spread, but possibly different means. More formally, with f_i denoting the pdf of the ith distribution, we assume that $f_1(x - \mu_1) = f_2(x - \mu_2) = \cdots = f_I(x - \mu_I)$, so that the distributions differ by (at most) a shift. Following the notation of Chapter 11,

let J_i = the ith sample size, $n = \sum J_i$ = the total number of observations in the data set, and X_{ij} = the jth observation in the ith sample ($j = 1, \ldots, J_i; i = 1, \ldots, I$). As in the rank-sum test, we replace each observation X_{ij} with its rank, R_{ij}, among all n observations. So, the smallest observation across all samples receives rank 1, the next-smallest rank 2, and so on through n.

Example 14.10 Diabetes and its associated health issues among children are of ever-increasing concern worldwide. The article "Clinical and Metabolic Characteristics among Mexican Children with Different Types of Diabetes Mellitus" (*PLoS ONE*, Dec. 16, 2016) reported on an in-depth study of children with one of four types of diabetes: (1) type 1 autoimmune, (2) type 2, (3) type 1 idiopathic (the most common type in children), and (4) what the researchers called "type 1.5." (The last category describes children who exhibit characteristics consistent with both type 1 and type 2; the authors note that the American Diabetes Association does not recognize such a category.)

To illustrate the Kruskal–Wallis method, presented here is a subset of the triglyceride measurements (mmol/L) on these children, along with their associated ranks in brackets.

Diabetes group	Triglyceride level (mmol/L)				
1	1.06 [5]	1.94 [11]	1.07 [6]		
2	1.30 [9]	2.08 [12]	2.15 [13]		
3	1.08 [7]	0.45 [1]	0.85 [2]	1.13 [8]	2.28 [14]	0.87 [3]
4	1.39 [10]	0.89 [4]			

In this example, $I = 4$ (four populations), $J_1 = J_2 = 3$, $J_3 = 6$, $J_4 = 2$, and $n = 14$. The first observation is $x_{11} = 1.06$ with associated rank $r_{11} = 5$. ∎

When $H_0: \mu_1 = \cdots = \mu_I$ is true, the I population/treatment distributions are identical, and so the X_{ij}'s form a random sample from a single population distribution. It follows that each R_{ij} is uniformly distributed on the integers 1, 2, …, n, so that $E(R_{ij}) = (n+1)/2$ for every i and j when H_0 is true. If we let $\bar{R}_{i\cdot}$ denote the mean of the ranks in the ith sample, then

$$E(\bar{R}_{i\cdot}) = \frac{1}{J_i} \sum_{i=1}^{J_i} E(R_{ij}) = \frac{n+1}{2}$$

Moreover, regardless of whether H_0 is true, the "grand mean" of all n ranks is $(n + 1)/2$, the average of the first n positive integers.

Similar to the treatment sum of squares SSTr from one-way ANOVA, the Kruskal–Wallis test statistic, denoted by H, quantifies "between-groups" variability by measuring how much the $\bar{R}_{i\cdot}$'s differ from the grand mean:

$$H = \frac{12}{n(n+1)} \text{SSTr} = \frac{12}{n(n+1)} \sum_{i=1}^{I} \sum_{j=1}^{J_i} (\bar{R}_{i\cdot} - \bar{R}_{\cdot\cdot})^2 = \frac{12}{n(n+1)} \sum_{i=1}^{I} J_i \left(\bar{R}_{i\cdot} - \frac{n+1}{2} \right)^2 \qquad (14.6)$$

When the null hypothesis is true, each $\bar{R}_{i\cdot}$ will be close to its expected value (the grand mean), whereas when H_0 is false certain samples should have an overabundance of high ranks and others too many low ranks, resulting in a larger value of H. Even for small J_i's, the exact null distribution of H in Expression (14.6) is unwieldy. Thankfully, when the sample sizes are not *too* small, the approximate sampling distribution of H under the null hypothesis is known.

KRUSKAL–WALLIS TEST   Null hypothesis: H_0: $\mu_1 = \cdots = \mu_I$
Alternative hypothesis: not all μ_i's are equal

Test statistic value: $h = \dfrac{12}{n(n+1)} \sum\limits_{i=1}^{I} J_i \left(\bar{r}_{i\cdot} - \dfrac{n+1}{2} \right)^2$,

where $\bar{r}_{i\cdot}$ denotes the average rank within the ith sample.

When H_0 is true, H has approximately a chi-squared distribution with $I - 1$ df. This approximation is reasonable provided that all $J_i \geq 5$.

It should not be surprising that H has an approximate χ^2_{I-1} distribution. By the Central Limit Theorem, the averages $\bar{R}_{i\cdot}$ should be approximately normal, and so H is similar to the sum of squares of I standardized normal rvs. However, as in the distribution of sample variance S^2, these rvs are *not* independent—the sum of all ranks is fixed—and that one constraint costs one degree of freedom.

Example 14.11 (Example 14.10 continued) Though the sample sizes in our illustrative example are a bit too small to meet the requirements of the Kruskal–Wallis test, we proceed with the rest of the test procedure on this reduced data set. The rank averages are $\bar{r}_{1\cdot} = (5 + 11 + 6)/3 = 7.33$, $\bar{r}_{2\cdot} = 11.33$, $\bar{r}_{3\cdot} = 5.83$, and $\bar{r}_{4\cdot} = 7$. The grand mean of all 14 ranks is $(n + 1)/2 = 7.5$, and the test statistic value is

$$h = \frac{12}{14(14+1)} \sum_{i=1}^{4} J_i (\bar{r}_{i\cdot} - 7.5)^2 = \frac{12}{210} [3(7.33 - 7.5)^2 + \cdots + 2(7 - 7.5)^2] = 3.50$$

Comparing this to the critical value $\chi^2_{.05, 4-1} = 7.815$, we would not reject H_0 at the .05 significance level. Equivalently, the P-value is $P(H \geq 3.50$ when $H \sim \chi^2_3) = .321$, again indicating no reason to reject H_0.

Details of the Kruskal–Wallis test for the full sample appear below. The small test statistic of 5.78 and relatively large P-value of .123 indicate that the mean triglyceride levels for these four populations of diabetic children are not statistically significantly different.

Diabetes group	J_i	$\bar{r}_{i\cdot}$	
1	25	64.0	$h = 5.78$
2	31	80.8	P-value = .123
3	63	62.0	
4	17	76.6	
	$n = 136$	$\bar{r}_{\cdot\cdot} = 68.5$	

∎

Expression (14.6) is sometimes written in other forms. For example, with W_i denoting the *sum* of the ranks for the ith sample (analogous to the Wilcoxon statistic W), it can be shown that

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{1}{J_i} \left(W_i - J_i \cdot \frac{n+1}{2} \right)^2 = \frac{12}{n(n+1)} \sum_{i=1}^{I} \frac{W_i^2}{J_i} - 3(n+1) \qquad (14.7)$$

The quantity $J_i(n+1)/2$ in the middle expression of (14.7) is the expected rank sum for the ith sample when H_0 is true. The far-right expression in (14.7) is computationally quicker than (14.6). Alternatively, some software packages report the Kruskal–Wallis statistic in the form $H = \sum Z_i^2$, where

$$Z_i = \frac{\bar{R}_{i\cdot} - (n+1)/2}{\sigma/\sqrt{J_i}}$$

and $\sigma^2 = n(n+1)/12$.

In Chapter 11, we emphasized the need for *two* measures of variability, SSTr and SSE, with the latter measuring the variability within each sample. Why is SSE not required here? The fundamental ANOVA identity SST = SSTr + SSE still applies to the R_{ij}'s, but because the ranks are just a re-arrangement of the integers 1 through n, the total sum of squares SST depends only on n and not on the raw data (Exercise 41). Thus, once we know n and SSTr, the other two sums of squares are completely determined.

Nonparametric ANOVA for a Randomized Block Design

The Kruskal–Wallis test is applicable to data resulting from a completely randomized design (independent random samples from I population or treatment distributions). Suppose instead that we have data from a randomized block experiment and wish to test the null hypothesis that the true population/treatment means are equal (i.e., "no treatment effect"). To test $H_0: \mu_1 = \cdots = \mu_I$ in this situation, the observations *within each block* are ranked from 1 to I, and then the average rank $\bar{r}_{i\cdot}$ is computed for each of the I treatments.

Example 14.12 The article "Modeling Cycle Times in Production Planning Models for Wafer Fabrication" (*IEEE Trans. on Semiconductor Manuf.* 2016: 153–167) reports on a study to compare three different linear programming models used in the simulation of factory processes: allocated cleaning function (ACF), fractional lead time (FLT), and simple rounding down (SRD). Simulations were run under five different demand representations, and the profit from the (simulated) manufacture of a particular product was determined.

In this study, there are $I = 3$ treatments being compared using $J = 5$ blocks. The profit data presented in the article, along with the rank of each observation within its block and the rank average for each treatment, appears below.

LP model	Demand representation (block)					$\bar{r}_{i\cdot}$
	1	2	3	4	5	
ACF	\$44,379 [1]	\$69,465 [3]	\$18,317 [2]	\$69,981 [3]	\$32,354 [3]	2.4
FLT	\$47,825 [3]	\$43,354 [1]	\$17,512 [1]	\$48,707 [2]	\$30,993 [2]	1.8
SRD	\$47,446 [2]	\$53,393 [2]	\$27,554 [3]	\$45,435 [1]	\$25,662 [1]	1.8

∎

Within each block, the average of the ranks 1, …, I is simply $(I + 1)/2$, and hence this is also the grand mean. If the null hypothesis of "no treatment effect" is true, then all $I!$ possible arrangements of the ranks within each block are equally likely, so each rank R_{ij} is uniformly distributed on $\{1, …, I\}$ and has expected value $(I + 1)/2$. The nonparametric method for analyzing this type of data, known as the **Friedman test** (developed by the Nobel Prize-winning economist Milton Friedman) relies on the following statistic:

$$Fr = \frac{12}{I(I+1)}\text{SSA} = \frac{12}{I(I+1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{R}_{i\cdot} - \bar{R}_{\cdot\cdot})^2 = \frac{12J}{I(I+1)}\sum_{i=1}^{I}\left(\bar{R}_{i\cdot} - \frac{I+1}{2}\right)^2 \qquad (14.8)$$

As with the Kruskal–Wallis test, the Friedman test rejects H_0 when the computed value of the test statistic is too large (an upper-tailed test). For even small-to-moderate values of J (the number of blocks), the test statistic Fr in (14.8) has approximately a chi-squared distribution with $I - 1$ df.

Example 14.13 (Example 14.12 continued) Applying Expression (14.8) to the profit data, the observed value of the test statistic is

$$fr = \frac{12(5)}{3(3+1)}\left[(2.4-2)^2 + (1.8-2)^2 + (1.8-2)^2\right] = 1.2$$

Since this is far less than the critical value $\chi^2_{.10,3-1} = 4.605$, the null hypothesis of equal mean profit for all three linear programming models is certainly not rejected. ∎

The Friedman test is used frequently to analyze "expert ranking" data (see, for example, Exercise 46). If each of J individuals ranks I items on some criterion, then the data is naturally of the type for which the Friedman test was devised. Each ranker acts as a block, and the test seeks out significant differences in the mean rank received by each of the I items.

As with the Kruskal–Wallis test, the total sum of squares for the Friedman test is fixed (in fact, it is a simple function of I and J; see Exercise 47(b)). In randomized block ANOVA in Chapter 11, the block sum of squares SSB gave an indication of whether blocking accounted for a significant amount of the total variation in the response values. In the data of Example 14.12, the blocking variable of demand representation clearly has an impact—for instance, the profits in the $j = 3$ block (middle column) are far lower than in other blocks. Unfortunately, SSB for Friedman's test is identically 0 (Exercise 47(a)), and so the effectiveness of blocking remains unquantified.

Exercises: Section 14.4 (37–48)

37. The article cited in Example 14.10 also reported the fasting C-peptide levels (FCP, nmol/L) for the children in the study.

Diabetes group	J_i	$\bar{r}_i$
1	26	72.8
2	32	79.2
3	65	56.4
4	17	104.6

(Sample sizes are different here than in Example 14.10 due to missing data in the latter.) Use a Kruskal–Wallis test (as the article's authors did) to determine whether true average FCP levels differ across these four populations of diabetic children.

38. The article "Analyses of Phenotypic Differentiations among South Georgian Diving Petrel Populations Reveal an Undescribed and Highly Endangered Species from New Zealand" (*PLoS ONE*, June 27, 2018) reports the possible discovery of a new species of *P. georgicus*, distinct from the birds of the same name found in the South Atlantic and South Indian Oceans. The table below summarizes information on the bill length (mm) of birds sampled for the study; bill length distributions are skewed, so a nonparametric method is appropriate.

Bird origin	J_i	$\bar{r}_{i\cdot}$
S. Atlantic	22	121.0
S. Indian	38	109.3
New Zealand	126	83.9

Test at the .01 significance level to see whether the mean bird length differs across these three geographic groups of *P. georgicus*. [The researchers performed over a dozen similar tests and found many features in which the New Zealand petrels are starkly different from the others.]

39. The following data on the fracture load (kN) of Plexiglas at three different loading point locations appeared in the article "Evaluating Fracture Behavior of Brittle Polymeric Materials Using an IASCB Specimen" (*J. of Engr. Manuf.* 2013: 133–140).

Distance	Fracture load			
31.2 mm	4.78	4.41	4.91	5.06
36.0 mm	3.47	3.85	3.77	3.63
42.0 mm	2.62	2.99	3.39	2.86

Use a rank-based method to determine whether loading point distance affects true mean fracture load, at the $\alpha = .01$ level.

40. Dental composites used to fill cavities will decay over time. The article "In vitro Aging Behavior of Dental Composites Considering the Influence of Filler Content, Storage Media and Incubation Time" (*PLoS ONE*, April 9, 2018) reported on a study to measure the hardness (MPa) of a particular type of resin after 14 days stored in artificial saliva, lactic acid, citric acid, or 40% ethanol.

Saliva	542.18	508.31	473.44	514.33	488.41
Lactic	478.99	501.15	488.97	463.68	471.14
Citric	427.97	388.59	378.01	341.61	395.12
Ethanol	482.96	451.48	436.69	424.42	465.64
Saliva	477.46	501.71	513.65	471.46	421.90
Lactic	568.14	494.15	494.99	483.89	520.33
Citric	433.59	353.03	344.90	387.09	501.81
Ethanol	387.59	322.55	277.84	367.36	385.75

Use a Kruskal–Wallis test with significance level .05 to determine whether true average hardness differs by liquid medium.

41. Let SST denote the total sum of squares for the Kruskal–Wallis test: $SST = \sum\sum (R_{ij} - \bar{R}_{..})^2$. Verify that $SST = n(n^2 - 1)/12$. [*Hint:* The R_{ij}'s are a re-arrangement of the integers 1 through

n. Use the formulas for the sum and sum of squares of the first *n* positive integers.]

42. Show that the two formulas for the Kruskal–Wallis test statistic in Expression (14.7) are identical and both equal the original formula for *H*.

43. Many people suffer back or neck pain due to bulging discs in the lumbar or cervical spine, but the thoracic spine (the section in-between) is less well-studied. The article "Kinematic analysis of the space available for cord and disc bulging of the thoracic spine using kinematic magnetic resonance imaging (kMRI)" (*The Spine J.* 2018: 1122–1127) describes a study using kMRI to measure disc bulge (mm) in neutral, flexion, and extension positions.

a. Suppose measurements were taken on just 6 subjects. The following bulge measurements at the T11–T12 disc (bottom of the thoracic spine) are consistent with information in the article:

Position	Subject					
	1	2	3	4	5	6
Neutral	1.28	0.88	0.69	1.52	0.83	2.58
Flexion	1.29	0.76	0.43	2.11	1.07	2.18
Extension	1.51	1.12	0.23	1.54	0.20	1.67

Convert these measurements into within-block ranks, and use the Friedman test to determine if the true average disc bulge at T11–T12 varies by position.

b. The study actually involved 105 subjects, each serving as her/his own block. The sum of the ranks for the three positions were neutral = 219, flexion = 222, extension = 189. Use these to perform the Friedman test, and report your conclusion at the .05 significance level.

c. Similar measurements were also taken on all 105 subjects at the T4–T5 disc (top of the thoracic spine); rank sums consistent with the article are 207, 221, and 202. Repeat the test of part (b) for the T4–T5 disc.

44. Is that Yelp review real or fake? The article "A Framework for Fake Review Detection in Online Consumer Electronics Retailers" (*Information Processing and Management* 2019: 1234–1244) tested five different classification algorithms on a large corpus of Yelp reviews from New York, Los Angeles, Miami, and San Francisco whose authenticity (real or fake) was known. Since review styles differ greatly by city, the researchers used city as a blocking variable. The table below shows the F_1 score, a standard measure of classification accuracy, for each algorithm-city pairing. (F_1 scores range from 0 to 1, with higher values implying better accuracy.)

Algorithm	NYC	LA	Miami	SF
Logistic regression	.79	.73	.78	.77
Decision tree	.81	.74	.81	.81
Random forest	.82	.78	.80	.80
Gaussian Naïve Bayes	.72	.69	.71	.69
AdaBoost	.83	.79	.82	.82

Test the null hypothesis that the five algorithms are equally accurate in classifying real and fake Yelp reviews at the .10 significance level.

45. Image segmentation is a key tool in computer vision (i.e., helping computers "see" the meaning in pictures). The article "Efficient Quantum Inspired Meta-Heuristics for Multi-Level True Colour Image Thresholding" (*Applied Soft Computing* 2017: 472–513) reported a study to compare 10 image segmentation algorithms—six conventional, four inspired by quantum computing. Each algorithm was applied to 10 different images, from an elephant to Mono Lake to the Mona Lisa; the images serve as blocks in this study. Kapur's method, an entropy measure for image segmentation tools, was applied to each (algorithm, image) pair; lower numbers are better. The article reports the following rank averages for the 10 algorithms.

GA	SA	PSO
8.30	9.10	8.90
QIGAMLTCI	QISAMLTCI	QIPSOMLTCI
2.15	3.65	1.85
DE	BSA	CoDE
6.60	5.90	6.20
QIDEMLTCI		
2.35		

Does the data indicate that the 10 algorithms are not equally effective at minimizing Kapur's entropy measure? Test at the .01 significance level. What do the rank averages suggest about quantum-inspired versus conventional image segmentation methods?

46. *Sustainability* in corporate culture is typically described as having three dimensions: economic, environmental, and social. The article "Development of Indicators for the Social Dimension of Sustainability in a U.S. Business Context" (*J. of Cleaner Production* 2019: 687–697) reported on the development of a survey instrument for the least-studied of these "three pillars," social sustainability. The researchers had 26 experts take the final version of the survey. In each survey section, participants were asked to rank a set of possible metrics from most important to least important.

a. The four metrics listed below were categorized as "public actualization needs." Use the mean ranks to test at the .05 significance level whether experts systematically prioritize some of these metrics over others with respect to social sustainability.

Metric	Mean rank
Ratio of public contributions (e.g., donations) to market capitalization	1.80
% of public that says company is making the world a better place	2.50
% of employees that contribute service for the public good	2.85
Ratio of minority management to minority workforce	2.85

b. Repeat part (a) using the following survey metrics for "public safety and security needs."

Metric	Mean rank
% of employees receiving human rights policy/procedure training	1.69
% of investment agreements that include human rights clauses	2.04
% of company sales that support human/environmental health/safety	2.27

47. a. In Chapter 11, the block sum of squares was defined by $SSB = I \sum_j (\bar{X}_{\cdot j} - \bar{X}_{\cdot \cdot})^2$. Replacing X's with R's in this expression, explain why $SSB = 0$ for the Friedman test.

    b. The total sum of squares for ranks R_{ij} is $SST = \sum \sum (R_{ij} - \bar{R}_{\cdot \cdot})^2$. Determine SST for the Friedman test. [*Hint:* Your answer should depend only on I and J.]

48. In the context of a randomized block experiment, let W_i denote the sample rank sum associated with the ith population/ treatment. Show that the Friedman's test statistic can be re-expressed as

$$Fr = \frac{12}{I(I+1)J} \sum_{i=1}^{I} \left(W_i - \frac{J(I+1)}{2} \right)^2$$

$$= \frac{12}{I(I+1)J} \sum_{i=1}^{I} W_i^2 - 3(I+1)J$$

Supplementary Exercises: (49–58)

49. In a study described in the article "Hyaluronan Impairs Vascular Function and Drug Delivery in a Mouse Model of Pancreatic Cancer" (*Gut* 2013: 112–120), hyaluronan was depleted from mice using either PEGPH20 or an equal dose of a standard treatment. The vessel patency (%) for each mouse was recorded.

PEGPH20	62	68	70	76
Standard	24	29	35	41

Use a rank-sum test (as did the article's authors) to determine if PEGPH20 yields higher vessel patency than the standard treatment at the .05 level. Would you also reject the null hypothesis at the .01 level? Comment on these results in light of the fact that every PEGPH20 measurement is higher than every standard-treatment measurement.

50. The article "Long Telomeres are Associated with Clonality in Wild Populations of ... *C. tenuispina*" (*Heredity* 2015: 437–443) reported the following telomere measurements for (1) a normal arm and (2) a regenerating arm of 12 Mediterranean starfish.

Normal	11.246	11.493	11.136	11.120	10.928	11.556
Regen.	11.142	11.047	11.004	11.506	11.067	10.875

Normal	11.313	11.164	10.878	12.680	11.937	11.172
Regen.	11.484	11.517	10.756	10.973	11.078	11.182

It is theorized that such measurements should be smaller for regenerating arms, because the above values are inversely related to telomere length and longer telomeres are associated with younger tissue. Use a nonparametric test to see if the data supports this theory at the .05 significance level.

51. Physicians use a variety of quantitative sensory testing (QST) tools to assess pain in patients, but there is concern about the consistency of such tools. The article "Test-Retest Reliability of [QST] in Knee Osteoarthritis and Healthy Participants" (*Osteoarthritis and Cartilage* 2011: 655–658) describes a study in which participants' responses to various stimuli were measured and then re-measured one week later. For example, pressure was applied to each subject's knee, and the level (kPa) at which the patient first experienced pain was recorded.

    a. Pressure pain measurements were taken twice on each of 50 patients with osteoarthritis in the examined knee. The Wilcoxon signed-rank test statistic value computed from this paired data was

$s_+ = 616$. Use a two-sided, large-sample test to assess the reliability of the sensory test at the .10 significance level.

b. The same measurements were made of 50 *healthy* patients, and the resulting test statistic value was $s_+ = 814$. Carry out the test indicated in part (a) using this information. Does the pain pressure test appear reliable for the population of healthy patients?

52. Adding mileage information to roadside amenity signs ("Motel 6 in 1.2 miles") can be helpful but might also increase accidents as drivers strain to read the detailed information at a distance. The article "Evaluation of Adding Distance Information to Freeway-Specific Service (Logo) Signs" (*Transp. Engr.* 2011: 782–788) provides the following information on number of crashes per year before and after mileage information was added to signs at six locations in Virginia.

Location	1	2	3	4	5	6
Before	15	26	66	115	62	64
After	16	24	42	80	78	73

a. Use a one-sample sign test to determine whether more accidents occur after mileage information is added to roadside amenity signs. Be sure to state the hypotheses, and indicate what assumptions are required.

b. Use a signed-rank test to determine whether more accidents tend to occur after mileage information is added to roadside amenity signs. What are the hypotheses now, and what additional assumptions are required?

53. The accompanying observations on axial stiffness index resulted from a study of metal-plate connected trusses in which five different plate lengths—4 in., 6 in., 8 in., 10 in., and 12 in.—were used ("Modeling Joints Made with Light-Gauge Metal Connector Plates," *Forest Products J.* 1979: 39–44).

$i = 1$ (4 in.):	309.2	309.7	311.0	316.8
	326.5	349.8	409.5	
$i = 2$ (6 in.):	331.0	347.2	348.9	361.0
	381.7	402.1	404.5	
$i = 3$ (8 in.):	351.0	357.1	366.2	367.3
	382.0	392.4	409.9	
$i = 4$ (10 in.):	346.7	362.6	384.2	410.6
	433.1	452.9	461.4	
$i = 5$ (12 in.):	407.4	410.7	419.9	441.2
	441.8	465.8	473.4	

Use the Kruskal–Wallis test to decide at significance level .01 whether the true average axial stiffness index depends somehow on plate length.

54. The article "Production of Gaseous Nitrogen in Human Steady-State Conditions" (*J. Appl. Physiol.* 1972: 155–159) reports the following observations on the amount of nitrogen expired (in liters) under four dietary regimens: (1) fasting, (2) 23% protein, (3) 32% protein, and (4) 67% protein. Use the Kruskal–Wallis test at level .05 to test equality of the corresponding μ_i's.

1	4.079	4.859	3.540	5.047	3.298
	4.679	2.870	4.648	3.847	
2	4.368	5.668	3.752	5.848	3.802
	4.844	3.578	5.393	4.374	
3	4.169	5.709	4.416	5.666	4.123
	5.059	4.403	4.496	4.688	
4	4.928	5.608	4.940	5.291	4.674
	5.038	4.905	5.208	4.806	

55. The article "Physiological Effects During Hypnotically Requested Emotions" (*Psychosomatic Med.* 1963: 334–343) reports the following data (x_{ij}) on skin potential in millivolts when the emotions of fear, happiness, depression, and calmness were requested from each of eight subjects.

	Blocks (subjects)			
	1	2	3	4
Fear	23.1	57.6	10.5	23.6
Happiness	22.7	53.2	9.7	19.6
Depression	22.5	53.7	10.8	21.1
Calmness	22.6	53.1	8.3	21.6
	5	6	7	8
Fear	11.9	54.6	21.0	20.3
Happiness	13.8	47.1	13.6	23.6
Depression	13.7	39.2	13.7	16.3
Calmness	13.3	37.0	14.8	14.8

Use Friedman's test to decide whether emotion has an effect on skin potential.

56. In an experiment to study the way in which different anesthetics affect plasma epinephrine concentration, ten dogs were selected, and concentration was measured while they were under the influence of the anesthetics isoflurane, halothane, and cyclopropane ("Sympathoadrenal and Hemodynamic Effects of Isoflurane, Halothane, and Cyclopropane in Dogs," *Anesthesiology* 1974: 465–470). Test at level .05 to see whether there is an anesthetic effect on concentration.

	Dog				
	1	2	3	4	5
Isoflurane	.28	.51	1.00	.39	.29
Halothane	.30	.39	.63	.38	.21
Cyclopropane	1.07	1.35	.69	.28	1.24
	6	7	8	9	10
Isoflurane	.36	.32	.69	.17	.33
Halothane	.88	.39	.51	.32	.42
Cyclopropane	1.53	.49	.56	1.02	.30

57. Suppose we wish to test

H_0: the X and Y distributions are identical versus

H_a: the X distribution is less spread out than the Y distribution

The accompanying figure pictures X and Y distributions for which H_a is true. The Wilcoxon rank-sum test is not appropriate in this situation because when H_a is true as pictured, the Y's will tend to be at the extreme ends of the combined sample (resulting in small and large Y ranks), so the sum of X ranks will result in a W value that is neither large nor small.

"Ranks" :	1	3	5	$\cdots$	6	4	2
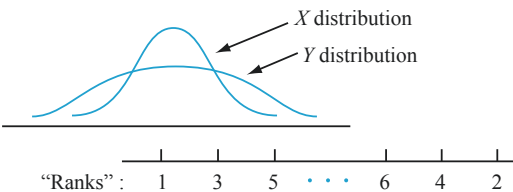
Consider modifying the procedure for assigning ranks as follows: After the combined sample of $m + n$ observations is ordered, the smallest observation is given rank 1, the largest observation is given rank 2, the second smallest is given rank 3, the second largest is given rank 4, and so on. Then if H_a is true as pictured, the X values will tend to be in the middle of the sample and thus receive large ranks. Let W' denote the sum of the X ranks and consider rejecting H_0 in favor of H_a when $w' \geq c$. When H_0 is true, every possible set of X ranks has the same probability, so W' *has the same distribution as does W when H_0 is true*. Thus c can be chosen from Appendix Table A.13 to yield a level α test. The accompanying data refers to medial muscle thickness for arterioles from the lungs of children who died from sudden infant death syndrome (x's) and a control group of children (y's). Carry out the test of H_0 versus H_a at level .05.

SIDS	4.0	4.4	4.8	4.9	
Control	3.7	4.1	4.3	5.1	5.6

[*Note:* Consult the Lehmann book in the bibliography for more information on this test, called the *Siegel–Tukey test*.]

58. The ranking procedure described in the previous exercise is somewhat asymmetric, because the smallest observation receives rank 1 whereas the largest receives rank 2, and so on. Suppose both the smallest and the largest receive rank 1, the second smallest and second largest receive rank 2, and so on, and let W'' be the sum of the X ranks. The null distribution of W'' is not identical to the null distribution of W, so different tables are needed. Consider the case $m = 3$, $n = 4$. List all 35 possible orderings of the three X values among the seven observations (e.g., 1, 3, 7 or 4, 5, 6), assign ranks in the manner described, compute the value of W'' for each possibility, and then tabulate the null distribution of W''. For the test that rejects if $w'' \geq c$, what value of c prescribes approximately a level .10 test? [*Note:* This is the *Ansari–Bradley test*; for additional information, see the book by Hollander and Wolfe in the bibliography.]

Introduction to Bayesian Estimation

15

Introduction

In this final chapter, we briefly introduce the *Bayesian* approach to parameter estimation. The standard *frequentist* view of inference is that the parameter of interest, θ, has a fixed but unknown value. Bayesians, however, regard θ as a random variable having a *prior* probability distribution that incorporates whatever is known about its value. Then to learn more about θ, a sample from the *conditional* distribution $f(x|\theta)$ is obtained, and Bayes' theorem is used to produce the *posterior* distribution of θ given the data $x_1, \ldots, x_n$. All Bayesian methods are based on this posterior distribution.

15.1 Prior and Posterior Distributions

Throughout this book, we have regarded parameters such as μ, σ, p, and λ as having an unknown but single, fixed value. This is often referred to as the *classical* or *frequentist* approach to statistical inference. However, there is a different paradigm, called *subjective* or *Bayesian inference*, in which an unknown parameter is assigned a distribution of possible values, analogous to a probability distribution. This distribution reflects all available information—past experience, intuition, common sense —about the value of the parameter prior to observing the data. For this reason, it is called the **prior distribution** of the parameter.

DEFINITION   A **prior distribution** for a parameter θ, denoted $\pi(\theta)$, is a probability distribution on the set of possible values for θ. In particular, if the possible values of the parameter θ form an interval I, then $\pi(\theta)$ is a pdf that must satisfy

$$\int_I \pi(\theta)\, d\theta = 1$$

Similarly, if θ is potentially any value in a discrete set D, then $\pi(\theta)$ is a pmf that must satisfy

$$\sum_{\theta \in D} \pi(\theta) = 1$$

Example 15.1 Consider the parameter μ = the mean GPA of all students at your university. Since GPAs are always between 0.0 and 4.0, μ must also lie in this interval. But common sense tells you that μ is almost certainly not below 2.0, or very few people would graduate, and it would be likewise surprising to find μ above 3.5. This "prior belief" can be expressed mathematically as a prior distribution for μ on the interval $I = [0, 4]$. If our best guess a priori is that $\mu \approx 2.5$, then our prior distribution $\pi(\mu)$ should be centered around 2.5. The variability of the prior distribution we select should reflect how sure we feel about our initial information.

If we feel very sure that μ is near 2.5, then we should select a prior distribution for μ that has less variation around that value. On the other hand, if we are less certain, this can be reflected by a prior distribution with much greater variability. Figure 15.1 illustrates these two cases; both of the pdfs depicted are beta distributions with $A = 0$ and $B = 4$.
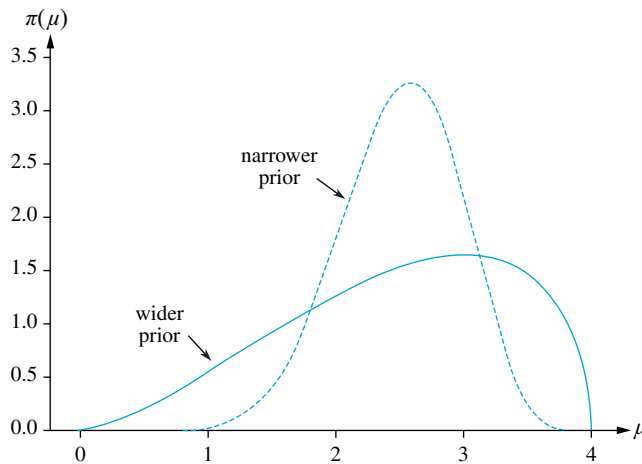
Figure 15.1 Two prior distributions for a parameter: a more diffuse prior (less certainty) and a more concentrated prior (more certainty)   ∎

The Posterior Distribution of a Parameter

The key to Bayesian inference is having a mathematically rigorous way to combine the sample data with prior belief. Suppose we observe values $x_1, \ldots, x_n$ from a distribution depending on the unknown parameter θ for which we have selected some prior distribution. Then a Bayesian statistician wants to "update" her or his belief about the distribution of θ, taking into account both prior belief and the observed x_i's. Recall from Chapter 2 that Bayes' theorem was used to obtain posterior probabilities of partitioning events $A_1, \ldots, A_k$ conditional on the occurrence of some other event B. The following definition relies on the analogous result for random variables.

DEFINITION   Suppose $X_1, \ldots, X_n$ have joint pdf $f(x_1, \ldots, x_n; \theta)$ and the unknown parameter θ has been assigned a continuous prior distribution $\pi(\theta)$. Then the **posterior distribution** of θ, given the observations $X_1 = x_1, \ldots, X_n = x_n$, is

$$\pi(\theta|x_1,\ldots,x_n) = \frac{\pi(\theta)f(x_1,\ldots,x_n;\theta)}{\int_{-\infty}^{\infty}\pi(\theta)f(x_1,\ldots,x_n;\theta)\,d\theta} \qquad (15.1)$$

The integral in the denominator of (15.1) insures that the posterior distribution is a valid probability density with respect to θ.

If $X_1, \ldots, X_n$ are discrete, the joint pdf is replaced by their joint pmf.

Notice that constructing the posterior distribution of a parameter requires a specific probability model $f(x_1, \ldots, x_n; \theta)$ for the observed data. In Example 15.1, it would not be enough to simply observe the GPAs of a random sample of n students; one must specify the underlying distribution, with mean μ, from which those GPAs are drawn.

Example 15.2 Emissions of subatomic particles from a radiation source are often modeled as a Poisson process. This implies that the time between successive emissions follows an exponential distribution. In practice, the parameter λ of this distribution is typically unknown. If researchers believe a priori that the average time between emissions is about half a second, so $\lambda \approx 2$, a prior distribution with a mean around 2 might be selected for λ. One example is the following gamma distribution, which has mean (and variance) of 2:

$$\pi(\lambda) = \lambda e^{-\lambda} \quad \lambda > 0$$

Notice that the gamma distribution has support equal to $0, \infty$, which is also the set of possible values for the unknown parameter λ.

The times $X_1, \ldots, X_5$ between five particle emissions will be recorded; it is these variables that have an exponential distribution with the unknown parameter λ (equivalently, mean $1/\lambda$). Because the X_i's are also independent, their joint pdf is

$$f(x_1,\ldots,x_5;\lambda) = f(x_1;\lambda)\cdot\cdots\cdot f(x_5;\lambda) = \lambda e^{-\lambda x_1}\cdot\cdots\cdot\lambda e^{-\lambda x_5} = \lambda^5 e^{-\lambda\sum x_i}$$

Applying (15.1) with these two components, the posterior distribution of λ given the observed data is

$$\pi(\lambda|x_1,\ldots,x_5) = \frac{\pi(\lambda)f(x_1,\ldots,x_5;\lambda)}{\int_{-\infty}^{\infty}\pi(\lambda)f(x_1,\ldots,x_5;\lambda)\,d\lambda} = \frac{\lambda e^{-\lambda}\cdot\lambda^5 e^{-\lambda\sum x_i}}{\int_0^{\infty}\lambda e^{-\lambda}\cdot\lambda^5 e^{-\lambda\sum x_i}d\lambda} = \frac{\lambda^6 e^{-\lambda(1+\sum x_i)}}{\int_0^{\infty}\lambda^6 e^{-\lambda(1+\sum x_i)}\,d\lambda}$$

Suppose the five observed interemission times are $x_1 = 0.66$, $x_2 = 0.48$, $x_3 = 0.44$, $x_4 = 0.71$, $x_5 = 0.56$. The sum of these five times is $\sum x_i = 2.85$, and so the posterior distribution simplifies to

$$\pi(\lambda|0.66,\ldots,0.56) = \frac{\lambda^6 e^{-3.85\lambda}}{\int_0^{\infty}\lambda^6 e^{-3.85\lambda}\,d\lambda} = \frac{3.85^7}{6!}\lambda^6 e^{-3.85\lambda} \quad \lambda > 0$$

The integral in the denominator was evaluated using the gamma integral formula (4.5) from Chapter 4; as noted previously, the purpose of this integral is to guarantee that the posterior distribution of λ is a valid probability density. As a function of λ, we recognize this as a gamma distribution with parameters $\alpha = 7$ and $\beta = 1/3.85$. The prior and posterior density curves of λ appear in Figure 15.2.
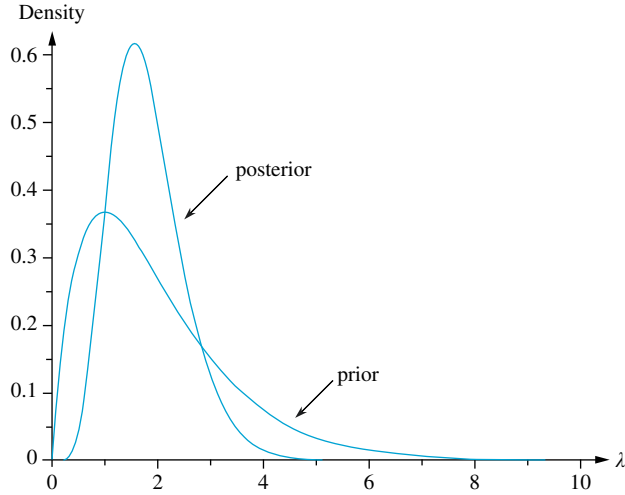
Figure 15.2   Prior and posterior distributions of λ for Example 15.2   ∎

Example 15.3   A 2010 National Science Foundation study found that 488 out of 939 surveyed adults incorrectly believe that antibiotics kill viruses (they only kill bacteria). Let θ denote the proportion of *all* U.S. adults that hold this mistaken view. Imagine that an NSF researcher, in advance of administering the survey, believed (hoped?) the value of θ was roughly 1 in 3, but he was very uncertain about this belief. Since any proportion must lie between 0 and 1, the standard beta family of distributions from Section 4.5 provides a natural source of priors for θ. One such beta distribution, with an expected value of 1/3, is the Beta(2, 4) model whose pdf is

$$\pi(\theta) = 20\theta(1-\theta)^3 \quad 0<\theta<1$$

The data mentioned at the beginning of the example can be considered either a random sample of size 939 from the Bernoulli distribution or, equivalently, a single observation from the binomial distribution with $n = 939$. Let Y = the number of U.S. adults in a random sample of 939 that believe antibiotics kill viruses. Then $Y \sim \text{Bin}(939, \theta)$, and the pmf of Y is $p(x; \theta) = \binom{939}{y}\theta^y(1-\theta)^{939-y}$. Substituting the observed value $y = 488$, (15.1) gives the posterior distribution of θ as

$$\pi(\theta|Y=488) = \frac{\pi(\theta)p(488;\theta)}{\int_0^1 \pi(\theta)p(488;\theta)\,d\theta} = \frac{20\theta(1-\theta)^3 \cdot \binom{939}{488}\theta^{488}(1-\theta)^{451}}{\int_0^1 20\theta(1-\theta)^3 \cdot \binom{939}{488}\theta^{488}(1-\theta)^{451}\,d\theta}$$

$$= \frac{\theta^{489}(1-\theta)^{454}}{\int_0^1 \theta^{489}(1-\theta)^{454}\,d\theta} = c \cdot \theta^{489}(1-\theta)^{454} \quad 0<\theta<1$$

Recall that the constant c, which equals the reciprocal of the integral in the denominator, serves to insure that the posterior distribution $\pi(\theta|Y=488)$ integrates to 1. Rather than evaluating the integral, we can simply recognize the expression $\theta^{489}(1-\theta)^{454}$ as a standard beta distribution, specifically with parameters $\alpha = 490$ and $\beta = 455$, that's just missing the constant of integration in front.

It follows that the posterior distribution of θ given $Y = 488$ must be Beta(490, 455); if we require c, it can be determined directly from the beta pdf.

This trick comes in handy quite often in Bayesian statistics: if we can recognize a posterior distribution as being proportional to a particular probability distribution, then it must necessarily be that distribution.

The prior and posterior density curves for θ are displayed in Figure 15.3. While the prior distribution is centered around 1/3 and exhibits a great deal of uncertainty (variability), the posterior distribution of θ is centered much closer to the sample proportion of incorrect answers, $488/939 \approx .52$, with considerably less uncertainty.
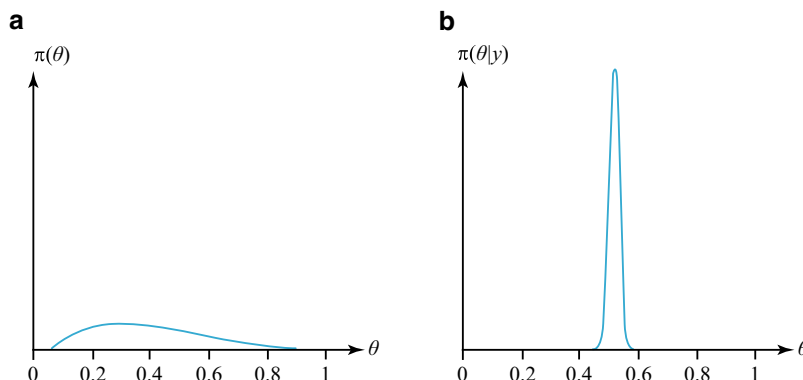
Figure 15.3   Density curves for the parameter θ in Example 15.3: (**a**) prior Beta(2, 4), (**b**) posterior Beta(490, 455)   ∎

Conjugate Priors
In the examples of this section, prior distributions were chosen partially by matching the mean of a distribution to someone's a priori "best guess" about the value of the parameter. We also mentioned at the beginning of the section that the variance of the prior distribution often reflects the strength of that belief. In practice, there is a third consideration for choosing a prior distribution: the ability to apply (15.1) in a simple fashion. Ideally, we would like to choose a prior distribution from a family (gamma, beta, etc.) such that the posterior distribution is from that same family. When this happens we say that the prior distribution is **conjugate** to the data distribution.

In Example 15.2, the prior $\pi(\lambda)$ is the Gamma(2, 1) pdf; we determined, using (15.1), that the posterior distribution was Gamma(7, 1/3.85). It can be shown in general (Exercise 6) that any gamma distribution is conjugate to an exponential data distribution. Similarly, the prior and posterior distributions of θ in Example 15.3 were Beta(2, 4) and Beta(490, 455), respectively. The following proposition generalizes the result of Example 15.3.

PROPOSITION   Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with unknown parameter value p. (Equivalently, let $Y = \sum X_i$ be a single observation from a Bin(n, p) distribution). If p is assigned a beta prior distribution with parameters α_0 and β_0, then the posterior distribution of p given the x_i's is the beta distribution with parameters $\alpha = \alpha_0 + y$ and $\beta = \beta_0 + n - y$.

Let that is, the beta distribution is a conjugate prior to the Bernoulli (or binomial) data model.

Proof The joint Bernoulli pmf of $X_1, \ldots, X_n$ is

$$p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} = p^{\sum x_i}(1-p)^{n-\sum x_i} = p^y(1-p)^{n-y}$$

The prior distribution assigned to p is $\pi(p) \propto p^{\alpha_0-1}(1-p)^{\beta_0-1}$. Apply (15.1):

$$\begin{aligned}
\pi(p|x_1, \ldots, x_n) &\propto p^{\alpha_0-1}(1-p)^{\beta_0-1} \cdot p^{\sum x_i}(1-p)^{n-\sum x_i} \\
&= p^{\alpha_0-1+\sum x_i}(1-p)^{\beta_0-1+n-\sum x_i} = p^{\alpha_0+y-1}(1-p)^{\beta_0+n-y-1}
\end{aligned}$$

As a function of p, we recognize this last expression as proportional to the beta pdf with parameters $\alpha = \alpha_0 + y$ and $\beta = \beta_0 + n - y$. ∎

The values α_0 and β_0 in the foregoing proposition are called **hyperparameters**; they are the parameters of the distribution *assigned to the original parameter*, p. In Example 15.2, the prior distribution $\pi(\lambda) = \lambda e^{-\lambda}$ is the Gamma(2, 1) pdf, so the hyperparameters of that distribution are $\alpha_0 = 2$ and $\beta_0 = 1$.

Conjugate priors have been determined for several of the named data distributions, including binomial, Poisson, gamma, and normal. For two-parameter families such as gamma and normal, it is sometimes reasonable to assume one parameter has a known value and then assign a prior distribution to the other. In some instances, a *joint* prior distribution for the two parameters can be found such that the posterior distribution is tractable, but these are less common.

PROPOSITION   Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma)$ distribution with σ known. If μ is assigned a normal prior distribution with hyperparameters μ_0 and σ_0, then the posterior distribution of μ is also normal, with posterior hyperparameters

$$\sigma_1^2 = \frac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}} \qquad \mu_1 = \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \cdot \sigma_1^2$$

That is, the normal distribution is a conjugate prior for μ in the normal data model when σ is assumed known.

Proof To determine the posterior distribution of μ, apply (15.1):

$$\begin{aligned}
\pi(\mu|x_1, \ldots, x_n) &\propto \pi(\mu)f(x_1, \ldots, x_n; \mu) = \frac{1}{\sqrt{2\pi}\sigma_0}e^{-(\mu-\mu_0)^2/2\sigma_0^2} \times \frac{1}{\sqrt{2\pi}\sigma}e^{-(x_1-\mu)^2/2\sigma^2} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}\sigma}e^{-(x_n-\mu)^2/2\sigma^2} \\
&\propto e^{-(1/2)[(x_1-\mu)^2/\sigma^2 + \cdots + (x_n-\mu)^2/\sigma^2 + (\mu-\mu_0)^2/\sigma_0^2]}
\end{aligned}$$

The trick here is to complete the square in the exponent, which yields

$$\pi(\mu|x_1,\ldots,x_n) \propto e^{-(\mu-\mu_1)^2/2\sigma_1^2 + C},$$

where C does not involve μ and

$$\sigma_1^2 = \frac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}} \qquad \mu_1 = \frac{\dfrac{\sum x_i}{\sigma^2} + \dfrac{\mu_0}{\sigma_0^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}} = \frac{\dfrac{n\bar{x}}{\sigma^2} + \dfrac{\mu_0}{\sigma_0^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}}$$

With respect to μ, the last expression for $\pi(\mu|x_1,\ldots,x_n)$ is proportional to the normal pdf with parameters μ_1 and σ_1. ∎

To make sense of these messy parameter expressions we define the **precision**, denoted by τ, as the reciprocal of the variance (because a lower variance implies a more precise measurement), and the weights then are the corresponding precisions. If we let $\tau = 1/\sigma^2$, $\tau_0 = 1/\sigma_0^2$, and $\tau_{\bar{X}} = 1/\sigma_{\bar{X}}^2 = n/\sigma^2$, then the posterior hyperparameters in the previous proposition can be restated as

$$\mu_1 = \frac{\tau_{\bar{X}} \cdot \bar{x} + \tau_0 \cdot \mu_0}{\tau_{\bar{X}} + \tau_0} \quad \text{and} \quad \tau_1 = \tau_{\bar{X}} + \tau_0$$

The posterior mean μ_1 is a weighted average of the prior mean μ_0 and the data mean $\bar{x}$, and the posterior precision is the sum of the prior precision plus the precision of the sample mean.

Exercises: Section 15.1 (1–10)

1. A certain type of novelty coin is manufactured so that 80% of the coins are fair while the rest have a .75 chance of landing heads. Let θ denote the probability of heads for a novelty coin randomly selected from this population.

   a. Express the given information as a prior distribution for the parameter θ.
   b. Five tosses of the randomly selected coin result in the sequence HHHTH. Use this data to determine the posterior distribution of θ.

2. Three assembly lines for the same product have different nonconformance rates: $p = .1$ for Line A, $p = .15$ for Line B, and $p = .2$ for Line C. One of the three lines will be selected at random (but you don't know which). Let X = the number of items inspected from the selected line until a nonconforming one is found.

   a. What is the distribution of X, as a function of the unknown p?

   b. Express the given information as a prior distribution for the parameter p. [*Hint:* There are three possible values for p. What should be their a priori likelihoods?]
   c. It is determined that the 8th item coming off the randomly selected line is the first nonconforming one. Use this information to determine the posterior distribution of p.

3. The number of customers arriving during a one-hour period at an ice cream shop is modeled by a Poisson distribution with unknown parameter μ. Based on past experience, the owner believes that the average number of customers in one hour is about 15.

   a. Assign a prior to μ from the gamma family of distributions, such that the mean of the prior is 15 and the standard deviation is 5 (reflecting moderate uncertainty).

b. The number of customers in ten randomly selected one-hour intervals is recorded:

16   9   11   13   17   17   8   15   14   16

Determine the posterior distribution of μ.

4. At children's party, kids toss ping-pong balls into the swimming pool hoping to land inside a plastic ring (picture a small hula hoop). Let p denote the probability of successfully tossing a ball into the ring, which we will assign a Beta(1, 3) prior distribution. The variable X = number of tosses required to get 5 balls in the ring (at which point the child wins a prize) will be observed for a sample of children.

a. What is a reasonable distribution for the rv X? What are its parameters?

b. The number of tosses required by eight kids was

12   8   14   12   12   6   8   27

Determine the posterior distribution of p.

5. Consider a random sample X_1, X_2, …, X_n from the Poisson distribution with mean μ. If the prior distribution for μ is a gamma distribution with hyperparameters α_0 and β_0, show that the posterior distribution is also gamma distributed. What are its hyperparameters?

6. Suppose you have a random sample X_1, X_2, …, X_n from the exponential distribution with parameter λ. If a gamma distribution with hyperparameters α_0 and β_0 is assigned as the prior distribution for λ, show that the posterior distribution is also gamma distributed. What are its hyperparameters?

7. Let $X_1, \ldots, X_n$ be a random sample from a negative binomial distribution with r known and p unknown. Assume a Beta (α_0, β_0) prior distribution for p. Show that the posterior distribution of p is also a beta distribution, and identify the updated hyperparameters.

8. Consider a random sample X_1, X_2, …, X_n from the normal distribution with mean 0 and precision τ (use τ as a parameter instead of $\sigma^2 = 1/\tau$). Assume a gamma-distributed prior for τ and show that the posterior distribution of τ is also gamma. What are its parameters?

9. Wind speeds in a certain area are modeled using a lognormal distribution, with unknown first parameter μ and known second parameter $\sigma = 1$. Suppose μ is assigned a normal prior distribution with mean μ_0 and precision τ_0. Based on observing a random sample of wind speeds $x_1, \ldots, x_n$ in this area, determine the posterior distribution of μ.

10. Wait times for Uber rides as people exit a certain sports arena are uniformly distributed on the interval $[0, \theta]$ with θ unknown. Suppose the following *Pareto* prior distribution is assigned to θ:

$$\pi(\theta) = \frac{24{,}000}{\theta^4} \quad \theta \geq 20$$

Based on observing the wait times $x_1, \ldots, x_n$ of n Uber customers, determine the posterior distribution of θ. [*Hint:* Some care must be taken to address the boundaries $\theta \geq 20$ and $x_i \leq \theta$.]

15.2 Bayesian Point and Interval Estimation

The previous section introduced the paradigm of Bayesian inference, wherein parameters are not just regarded as unknown but as having a distribution of possible values prior to observing any data. Such prior distributions are, by definition, valid probability distributions with respect to the parameter θ. The key to Bayesian inference is Equation (15.1), which applies Bayes' theorem for random variables to θ and the sample $X_1, \ldots, X_n$. The result is an update to our belief about θ, called the posterior distribution.

From a Bayesian perspective, the posterior distribution of θ represents the most complete expression of what can be inferred from the sample data. But the posterior distribution can give rise to

point and interval estimates for the parameter θ, although the interpretation of the latter differs from that of our earlier confidence intervals.

Bayesian Point Estimators

Although specifying a single-value estimate of an unknown parameter conflicts somewhat with the Bayesian philosophy, there are occasions when such an estimate is desired. The most common Bayesian point estimator of a parameter θ is the mean of its posterior distribution:

$$\hat{\theta} = E(\theta|X_1, \ldots, X_n) \tag{15.2}$$

Hereafter we shall refer to Expression (15.2) as the **Bayes estimator** of θ. A Bayes estimate is obtained by plugging in the observed values of the X_i's, resulting in the numerical value $\hat{\theta} = E(\theta|x_1, \ldots, x_n)$.

Example 15.4 (Example 15.2 continued) The posterior distribution of the exponential parameter λ given the five observed interemission times was determined to be gamma with parameters $\alpha = 7$ and $\beta = 1/3.85$. Since the mean of a gamma distribution is $\alpha\beta$, the Bayes estimate of λ here is

$$\hat{\lambda} = E(\lambda|0.66, \ldots, 0.56) = \alpha\beta = 7(1/3.85) = 1.82$$

This isn't too different from the researchers' prior belief that $\lambda \approx 2$.

   If we retrace the steps that led to this posterior distribution, we find more generally that for data model $X_1, \ldots, X_n \sim$ exponential(λ) and prior $\lambda \sim$ gamma(α_0, β_0), the posterior distribution of λ is the gamma pdf with $\alpha = \alpha_0 + n$ and $1/\beta = 1/\beta_0 + \sum X_i$. Therefore, the Bayes estimator for λ in this scenario is

$$\hat{\lambda} = E(\lambda|X_1, \ldots, X_n) = \alpha\beta = \frac{\alpha_0 + n}{1/\beta_0 + \sum X_i} \qquad\blacksquare$$

   Although the mean of the posterior distribution is commonly used as a point estimate for a parameter in Bayesian inference, that is not the only available choice. Some practitioners prefer to use the *mode* of the posterior distribution rather than the mean; this choice is called the **maximum a posteriori (MAP)** estimate of θ. For small samples, the Bayes estimate (i.e., mean) and MAP estimate can differ considerably. Typically, though, when n is large these estimates for the parameter will be reasonably close. This makes sense intuitively, since as n increases any sensible estimator should converge to the true, single value of the parameter (i.e., be consistent).

Example 15.5 (Example 15.3 continued) The posterior distribution of the parameter $\theta =$ the proportion of all U.S. adults that incorrectly believe antibiotics kill viruses was determined to have a Beta(490, 455) distribution. Since the mean of a Beta(α, β) distribution is $\alpha/(\alpha + \beta)$, a point estimate of θ is

$$\hat{\theta} = E(\theta|y = 488) = \frac{490}{490 + 455} = \frac{490}{945} = .5185$$

   It can be shown that the mode of a beta distribution occurs at $(\alpha - 1)/(\alpha + \beta - 2)$ provided that $\alpha > 1$ and $\beta > 1$. Hence the MAP estimate of θ here is $(490 - 1)/(490 + 455 - 2) = 489/943 = .5186$. Notice these are both quite close to the frequentist estimate $y/n = 488/939 = .5197$. $\qquad\blacksquare$

Properties of Bayes Estimators

In most cases, the contribution of the observed values $x_1, \ldots, x_n$ in shaping the posterior distribution of a parameter θ increases as the sample size n increases. Equivalently, the choice of prior distribution is less impactful for large samples, because the data "dominates" that original choice. It can be shown that under very general conditions, as $n \to \infty$

(1) the mean of the posterior distribution will converge to the true value of θ, and
(2) the variance of the posterior distribution of θ converges to zero.

The second property manifests itself in our two previous examples: the variability of the posterior distribution of λ based on $n = 5$ observations was still rather substantial, while the posterior distribution of θ based on a sample of size $n = 939$ was quite concentrated. In the language of Chapter 7, these two properties imply that Bayes estimators are generally *consistent*.

Since traditional estimators such as $\widehat{P}$ and $\overline{X}$ converge to the true values of corresponding parameters (e.g., p or μ) by the Law of Large Numbers, it follows that Bayesian and frequentist estimates will typically be quite close when n is large. This is true both for the point estimates and the interval estimates (Bayesian intervals will be introduced shortly). But when n is small—a common occurrence in Bayesian methodology—parameter estimates based on the two methods can differ drastically. This is especially true if the researcher's prior belief is very far from what's actually true (e.g., believing a proportion is around 1/3 when it's really greater than .5).

Example 15.6 Consider a Bernoulli(p) random sample $X_1, \ldots, X_n$ or, equivalently, a single binomial observation $Y = \sum X_i$. If we assign a Beta(α_0, β_0) prior to p, a proposition from the previous section establishes that the posterior distribution of p given that $Y = y$ is Beta($\alpha_0 + y, \ \beta_0 + n - y$). Hence the Bayes estimator of p is

$$\hat{p} = E(p | X_1, \ldots, X_n) = E(p | Y) = \frac{(\alpha_0 + Y)}{(\alpha_0 + Y) + (\beta_0 + n - Y)} = \frac{\alpha_0 + Y}{\alpha_0 + \beta_0 + n}$$

One way to think about the prior distribution here is that it "seeds" the sample with α_0 successes and β_0 failures before data is obtained. The quantities $\alpha_0 + Y$ and $\beta_0 + n - Y$ then represent the number of successes and failures after sampling, and the Bayes estimator $\hat{p}$ is the sample proportion of successes from this perspective.

With a little algebra, we can re-express the Bayes estimator as

$$\hat{p} = \frac{\alpha_0}{\alpha_0 + \beta_0 + n} + \frac{Y}{\alpha_0 + \beta_0 + n} = \frac{\alpha_0 + \beta_0}{\alpha_0 + \beta_0 + n} \cdot \frac{\alpha_0}{\alpha_0 + \beta_0} + \frac{n}{\alpha_0 + \beta_0 + n} \cdot \frac{Y}{n} \qquad (15.3)$$

Expression (15.3) represents the Bayesian estimator $\hat{p}$ as a weighted average of the prior expectation of p, $\alpha_0 / (\alpha_0 + \beta_0)$, and the sample proportion of successes $Y/n = \sum X_i / n$.

By the Law of Large Numbers, the sample proportion of successes Y/n converges to the true value of the Bernoulli parameter, which we will denote by p^*. Taking the limit of (15.3) as $n \to \infty$ yields

$$\hat{p} \to 0 \cdot \frac{\alpha_0}{\alpha_0 + \beta_0} + 1 \cdot p^* = p^*,$$

so that the Bayes estimator is indeed consistent.   ∎

Bayesian Interval Estimation

The Bayes estimate $\hat{\theta}$ provides a single-value "best guess" for the true value of the parameter θ based on its posterior distribution. An interval $[a, b]$ having posterior probability .95 gives a 95% **credible interval**, the Bayesian analogue of a 95% confidence interval (but the interpretation is different). Typically one selects the middle 95% of the posterior distribution; i.e., the endpoints of a 95% credible interval are ordinarily the .025 and .975 quantiles of the posterior distribution.

Example 15.7 (Example 15.4 continued) Given the observed values of $X_1, \ldots, X_5$, we previously found that the emission rate parameter λ has a Gamma(7, 1/3.85) posterior distribution. A 95% credible interval for λ requires determining the .025 and .975 quantiles of the Gamma(7, 1/3.85) model. Using statistical software, $\eta_{.025} = 0.7310$ and $\eta_{.975} = 3.3921$, so the 95% credible interval for λ is (0.7310, 3.3921). Under the Bayesian interpretation, having observed the five aforementioned interemission times, there is a 95% posterior probability that λ is between 0.7310 and 3.3921 emissions per second. Taking reciprocals, the mean time between emissions (i.e., $1/\lambda$) is estimated to lie in the interval (1/3.3921, 1/0.7310) = (0.295, 1.368) seconds with posterior probability .95.   ∎

Example 15.8 (Example 15.5 continued) The posterior distribution of the parameter θ = the proportion of all U.S. adults that incorrectly believe antibiotics kill viruses was a Beta(490, 455) distribution. The .025 and .975 quantiles of this beta distribution are $\eta_{.025} = .4866$ and $\eta_{.975} = .5503$. So, after observing the results of the NSF survey, there is a 95% posterior probability that θ is between .4866 and .5503.

For comparison, the one-proportion z interval based on $y/n = 488/939 = .5197$ is

$$.5197 \pm 1.96\sqrt{\frac{.5197(1 - .5197)}{939}} = (.4877, .5517)$$

Due to the large sample size, the two intervals are quite similar.   ∎

It must be emphasized that, even if the confidence interval is nearly the same as the credible interval for a parameter, they have different interpretations. To interpret the Bayesian credible interval, we say that there is a 95% *probability* that the parameter θ is in the interval. However, for the frequentist confidence interval such a probability statement does not make sense: as we discussed in Section 8.1, neither the parameter θ nor the endpoints of the interval are considered random under the frequentist view. (Instead, the confidence level is the long-run capture frequency if the formula is used repeatedly on different samples.)

Example 15.9 Consider the IQ scores of 18 first-grade boys, from the private speech data introduced in Exercise 81 from Chapter 1:

113 108 140 113 115 146 136 107 108 119 132 127 118 108 103 103 122 111

IQ scores are generally found to be normally distributed, and because IQs have a standard deviation of 15 nationwide, we can assume $\sigma = 15$ is known and valid here. Let's perform a Bayesian analysis on the mean IQ μ of all first-grade boys at the school.

For the normal prior distribution it is reasonable to use a mean of $\mu_0 = 110$, a ballpark figure for previous years in this school. It is harder to prescribe a standard deviation for the prior, but we will use $\sigma_0 = 7.5$. (This is the standard deviation for the average of four independent observations if the individual standard deviation is 15. As a result, the effect on the posterior mean will turn out to be the same as if there were four additional observations with average 110.)

The last proposition from Section 15.1 states that the posterior distribution of μ is also normal and specifies the posterior hyperparameters. Numerically, we have

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2} = \frac{1}{15^2/18} + \frac{1}{7.5^2} = .09778 = \frac{1}{10.227} = \frac{1}{3.198^2}$$

$$\mu_1 = \frac{\dfrac{n\bar{x}}{\sigma^2} + \dfrac{\mu_0}{\sigma_0^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_0^2}} = \frac{\dfrac{18(118.28)}{15^2} + \dfrac{110}{7.5^2}}{\dfrac{18}{15^2} + \dfrac{1}{7.5^2}} = 116.77$$

The posterior distribution is normal with mean $\mu_1 = 116.77$ and standard deviation $\sigma_1 = 3.198$. The mean μ_1 is a weighted average of $\bar{x} = 118.28$ and $\mu_0 = 110$, so μ_1 is necessarily between them. As n becomes large the weight given to μ_0 declines, and μ_1 will be closer to $\bar{x}$.

The 95% credible interval for μ is the middle 95% of the $N(116.77, 3.198)$ distribution, which works out to be $(110.502, 123.038)$. For comparison, the 95% confidence interval using $\bar{x} = 118.28$ and $\sigma = 15$ is $\bar{x} \pm 1.96\sigma/\sqrt{n} = (111.35, 125.21)$. Notice that the one-sample z interval must be wider: because the precisions add to give the posterior precision, the posterior precision is greater than the prior precision and it is greater than the data precision. Therefore, it is guaranteed that the posterior standard deviation σ_1 will be less than both σ_0 and $\sigma/\sqrt{n}$.

Both the credible interval and the confidence interval exclude 110, so we can be pretty sure that μ exceeds 110. Another way of looking at this is to calculate the posterior probability of μ being less than or equal to 110 (the Bayesian approach to hypothesis testing). Using $\mu_1 = 116.77$ and $\sigma_1 = 3.198$, we obtain the probability .0171, supporting the claim that μ exceeds 110.

What should be done if there are no prior observations and there are no strong opinions about the prior mean μ_0? In this case the prior standard deviation σ_0 can be taken as some number much larger than σ, such as $\sigma_0 = 1000$ in our example. The result is that the prior will have essentially no effect, and the posterior distribution will be based on the data: $\mu_1 \approx \bar{x} = 118.28$ and $\sigma_1 \approx \sigma = 15$. The 95% credible interval will be virtually the same as the 95% confidence interval based on the 18 observations, $(111.35, 125.21)$, but of course the interpretation is different.   ∎

Exercises: Section 15.2 (11–20)

11. Refer back to Exercise 3.

    a. Calculate the Bayes estimate of the Poisson mean parameter μ.
    b. Calculate and interpret a 95% credible interval for μ.

12. Refer back to Exercise 4.

    a. Calculate the Bayes estimate of the probability p.
    b. Calculate and interpret a 95% credible interval for p.

13. *Laplace's rule of succession* says that if all n Bernoulli trials have been successes, then the probability of a success on the next trial is $(n + 1)/(n + 2)$. For the derivation, Laplace used a Beta(1, 1) prior for the parameter p.

    a. Show that, if a Beta(1, 1) prior is assigned to p and there are n successes in n trials, then the posterior mean of p is $(n + 1)/(n + 2)$.
    b. Explain (a) in terms of total successes and failures; that is, explain the result in terms of two prior trials plus n later trials.
    c. Laplace applied his rule of succession to compute the probability that the sun will rise tomorrow using 5000 years, or $n = 1,826,214$ days of history in which the sun rose every day. Is Laplace's method equivalent to including two prior days when the sun rose once and failed to rise once? Criticize the answer in terms of total successes and failures.

14. In a study of 70 restaurant bills, 40 of the 70 were paid using cash. Let p denote the population proportion paying cash.

    a. Assuming a beta prior distribution for p with $\alpha_0 = 2$ and $\beta_0 = 2$, obtain the posterior distribution of p.
    b. Repeat (a) with α_0 and β_0 positive and close to 0.
    c. Calculate a 95% credible interval for p using (b). Is your interval compatible with $p = .5$?
    d. Calculate a 95% confidence interval for p using Equation (5.3), and compare with the result of (c).
    e. Compare the interpretations of the credible interval and the confidence interval.
    f. Based on the prior in (b), test the hypothesis $p \leq .5$ by using the posterior distribution to find $P(p \leq .5)$.

15. For the scenario of Example 15.9, assume the same normal prior distribution but suppose that the data set is just one observation $\bar{x} = 118.28$ with standard deviation $\sigma/\sqrt{n} = 15/\sqrt{18} = 3.5355$. Use Equation (15.1) to derive the posterior distribution, and compare your answer with the result of Example 15.9.

16. Here are the IQ scores for the 15 first-grade girls from the study mentioned in Example 15.9.

102	96	106	118	108	122	115	113
109	113	82	110	121	110	99	

    Assume that the data is a random sample from a normal distribution with mean μ and $\sigma = 15$, and assign to μ the same $N(110, 7.5)$ prior distribution used in Example 15.9.

    a. Determine the posterior distribution of μ.
    b. Calculate and interpret a 95% credible interval for μ.
    c. Add four observations with average 110 to the data and compute a 95% confidence one-sample z interval for μ using the 19 observations. Compare with the result of (b).
    d. Change the prior so the prior precision is very small but positive, and then re-compute (a) and (b).

    e. Calculate a 95% confidence one-sample z interval for μ using the 15 observations and compare with the credible interval of (d).

17. If α and β are large, then the beta distribution can be approximated by the normal distribution using the beta mean and variance given in Section 4.5. This is useful in case beta distribution software is unavailable. Use the approximation to compute the credible interval in Example 15.8.

18. Two political scientists wish to forecast the proportion of votes that a certain U.S. senator will earn in her upcoming reelection contest. The first political scientist assigns a Beta(3, 2) prior to p = the true support rate for the senator, while the second assigns a Beta(6, 6) prior.

    a. Determine the expectations of both prior distributions.
    b. Which political scientist appears to feel more sure about this prior belief? How can you tell?
    c. Determine both Bayes estimates in this scenario, assuming that y out of n randomly selected voters indicate they will vote to reelect the senator.
    d. For what survey size n are the two Bayes estimates guaranteed to be within .005 of each other, no matter the value of y?

19. Consider a random sample $X_1, \ldots, X_n$ from a Poisson distribution with unknown mean μ, and assign to μ a Gamma(α_0, β_0) prior distribution.

    a. What is the prior expectation of μ?
    b. Determine the Bayes estimator $\hat{\mu}$.
    c. Let μ^* denote the true value of μ. Show that $\hat{\mu}$ is a consistent estimator of μ^*. [*Hint:* Look back at Example 15.6.]

20. Consider a random sample $X_1, \ldots, X_n$ from an exponential distribution with parameter λ, and assign to λ a Gamma(α_0, β_0) prior distribution.

    a. What is the prior expectation of λ?
    b. Determine the Bayes estimator $\hat{\lambda}$.
    c. Let λ^* denote the true value of λ. Show that $\hat{\lambda}$ is a consistent estimator of λ^*.

Appendix

Table A.1 Cumulative binomial probabilities

$$B(x; n, p) = \sum_{y=0}^{x} b(y; n, p)$$

		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
								p								
a. *n* = 5																
x	0	.951	.774	.590	.328	.237	.168	.078	.031	.010	.002	.001	.000	.000	.000	.000
	1	.999	.977	.919	.737	.633	.528	.337	.188	.087	.031	.016	.007	.000	.000	.000
	2	1.000	.999	.991	.942	.896	.837	.683	.500	.317	.163	.104	.058	.009	.001	.000
	3	1.000	1.000	1.000	.993	.984	.969	.913	.812	.663	.472	.367	.263	.081	.023	.001
	4	1.000	1.000	1.000	1.000	.999	.998	.990	.969	.922	.832	.763	.672	.410	.226	.049
b. *n* = 10																
x	0	.904	.599	.349	.107	.056	.028	.006	.001	.000	.000	.000	.000	.000	.000	.000
	1	.996	.914	.736	.376	.244	.149	.046	.011	.002	.000	.000	.000	.000	.000	.000
	2	1.000	.988	.930	.678	.526	.383	.167	.055	.012	.002	.000	.000	.000	.000	.000
	3	1.000	.999	.987	.879	.776	.650	.382	.172	.055	.011	.004	.001	.000	.000	.000
	4	1.000	1.000	.998	.967	.922	.850	.633	.377	.166	.047	.020	.006	.000	.000	.000
	5	1.000	1.000	1.000	.994	.980	.953	.834	.623	.367	.150	.078	.033	.002	.000	.000
	6	1.000	1.000	1.000	.999	.996	.989	.945	.828	.618	.350	.224	.121	.013	.001	.000
	7	1.000	1.000	1.000	1.000	1.000	.998	.988	.945	.833	.617	.474	.322	.070	.012	.000
	8	1.000	1.000	1.000	1.000	1.000	1.000	.998	.989	.954	.851	.756	.624	.264	.086	.004
	9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.994	.972	.944	.893	.651	.401	.096
c. *n* = 15																
x	0	.860	.463	.206	.035	.013	.005	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.990	.829	.549	.167	.080	.035	.005	.000	.000	.000	.000	.000	.000	.000	.000
	2	1.000	.964	.816	.398	.236	.127	.027	.004	.000	.000	.000	.000	.000	.000	.000
	3	1.000	.995	.944	.648	.461	.297	.091	.018	.002	.000	.000	.000	.000	.000	.000
	4	1.000	.999	.987	.836	.686	.515	.217	.059	.009	.001	.000	.000	.000	.000	.000
	5	1.000	1.000	.998	.939	.852	.722	.402	.151	.034	.004	.001	.000	.000	.000	.000
	6	1.000	1.000	1.000	.982	.943	.869	.610	.304	.095	.015	.004	.001	.000	.000	.000
	7	1.000	1.000	1.000	.996	.983	.950	.787	.500	.213	.050	.017	.004	.000	.000	.000
	8	1.000	1.000	1.000	.999	.996	.985	.905	.696	.390	.131	.057	.018	.000	.000	.000
	9	1.000	1.000	1.000	1.000	.999	.996	.966	.849	.597	.278	.148	.061	.002	.000	.000
	10	1.000	1.000	1.000	1.000	1.000	.999	.991	.941	.783	.485	.314	.164	.013	.001	.000
	11	1.000	1.000	1.000	1.000	1.000	1.000	.998	.982	.909	.703	.539	.352	.056	.005	.000
	12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.996	.973	.873	.764	.602	.184	.036	.000
	13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.995	.965	.920	.833	.451	.171	.010
	14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.995	.987	.965	.794	.537	.140

(continued)

Table A.1 (continued)

$$B(x;n,p) = \sum_{y=0}^{x} b(y;n,p)$$

		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
								p								

d. *n* = 20

x		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
	0	.818	.358	.122	.012	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.983	.736	.392	.069	.024	.008	.001	.000	.000	.000	.000	.000	.000	.000	.000
	2	.999	.925	.677	.206	.091	.035	.004	.000	.000	.000	.000	.000	.000	.000	.000
	3	1.000	.984	.867	.411	.225	.107	.016	.001	.000	.000	.000	.000	.000	.000	.000
	4	1.000	.997	.957	.630	.415	.238	.051	.006	.000	.000	.000	.000	.000	.000	.000
	5	1.000	1.000	.989	.804	.617	.416	.126	.021	.002	.000	.000	.000	.000	.000	.000
	6	1.000	1.000	.998	.913	.786	.608	.250	.058	.006	.000	.000	.000	.000	.000	.000
	7	1.000	1.000	1.000	.968	.898	.772	.416	.132	.021	.001	.000	.000	.000	.000	.000
	8	1.000	1.000	1.000	.990	.959	.887	.596	.252	.057	.005	.001	.000	.000	.000	.000
	9	1.000	1.000	1.000	.997	.986	.952	.755	.412	.128	.017	.004	.001	.000	.000	.000
	10	1.000	1.000	1.000	.999	.996	.983	.872	.588	.245	.048	.014	.003	.000	.000	.000
	11	1.000	1.000	1.000	1.000	.999	.995	.943	.748	.404	.113	.041	.010	.000	.000	.000
	12	1.000	1.000	1.000	1.000	1.000	.999	.979	.868	.584	.228	.102	.032	.000	.000	.000
	13	1.000	1.000	1.000	1.000	1.000	1.000	.994	.942	.750	.392	.214	.087	.002	.000	.000
	14	1.000	1.000	1.000	1.000	1.000	1.000	.998	.979	.874	.584	.383	.196	.011	.000	.000
	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.994	.949	.762	.585	.370	.043	.003	.000
	16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.984	.893	.775	.589	.133	.016	.000
	17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.996	.965	.909	.794	.323	.075	.001
	18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.992	.976	.931	.608	.264	.017
	19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.997	.988	.878	.642	.182

e. *n* = 25

x		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
	0	.778	.277	.072	.004	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	1	.974	.642	.271	.027	.007	.002	.000	.000	.000	.000	.000	.000	.000	.000	.000
	2	.998	.873	.537	.098	.032	.009	.000	.000	.000	.000	.000	.000	.000	.000	.000
	3	1.000	.966	.764	.234	.096	.033	.002	.000	.000	.000	.000	.000	.000	.000	.000
	4	1.000	.993	.902	.421	.214	.090	.009	.000	.000	.000	.000	.000	.000	.000	.000
	5	1.000	.999	.967	.617	.378	.193	.029	.002	.000	.000	.000	.000	.000	.000	.000
	6	1.000	1.000	.991	.780	.561	.341	.074	.007	.000	.000	.000	.000	.000	.000	.000
	7	1.000	1.000	.998	.891	.727	.512	.154	.022	.001	.000	.000	.000	.000	.000	.000
	8	1.000	1.000	1.000	.953	.851	.677	.274	.054	.004	.000	.000	.000	.000	.000	.000
	9	1.000	1.000	1.000	.983	.929	.811	.425	.115	.013	.000	.000	.000	.000	.000	.000
	10	1.000	1.000	1.000	.994	.970	.902	.586	.212	.034	.002	.000	.000	.000	.000	.000
	11	1.000	1.000	1.000	.998	.980	.956	.732	.345	.078	.006	.001	.000	.000	.000	.000
	12	1.000	1.000	1.000	1.000	.997	.983	.846	.500	.154	.017	.003	.000	.000	.000	.000
	13	1.000	1.000	1.000	1.000	.999	.994	.922	.655	.268	.044	.020	.002	.000	.000	.000
	14	1.000	1.000	1.000	1.000	1.000	.998	.966	.788	.414	.098	.030	.006	.000	.000	.000
	15	1.000	1.000	1.000	1.000	1.000	1.000	.987	.885	.575	.189	.071	.017	.000	.000	.000
	16	1.000	1.000	1.000	1.000	1.000	1.000	.996	.946	.726	.323	.149	.047	.000	.000	.000
	17	1.000	1.000	1.000	1.000	1.000	1.000	.999	.978	.846	.488	.273	.109	.002	.000	.000
	18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.993	.926	.659	.439	.220	.009	.000	.000
	19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.971	.807	.622	.383	.033	.001	.000
	20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.991	.910	.786	.579	.098	.007	.000
	21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.967	.904	.766	.236	.034	.000
	22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.991	.968	.902	.463	.127	.002
	23	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.998	.993	.973	.729	.358	.026
	24	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	.999	.996	.928	.723	.222

Table A.2 Cumulative Poisson probabilities

$$P(x; \mu) = \sum_{y=0}^{x} \frac{e^{-\mu}\mu^y}{y!}$$

		.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
						μ					
x	0	.905	.819	.741	.670	.607	.549	.497	.449	.407	.368
	1	.995	.982	.963	.938	.910	.878	.844	.809	.772	.736
	2	1.000	.999	.996	.992	.986	.977	.966	.953	.937	.920
	3		1.000	1.000	.999	.998	.997	.994	.991	.987	.981
	4				1.000	1.000	1.000	.999	.999	.998	.996
	5							1.000	1.000	1.000	.999
	6										1.000

		2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	15.0	20.0
							μ						
x	0	.135	.050	.018	.007	.002	.001	.000	.000	.000	.000	.000
	1	.406	.199	.092	.040	.017	.007	.003	.001	.000	.000	.000
	2	.677	.423	.238	.125	.062	.030	.014	.006	.003	.000	.000
	3	.857	.647	.433	.265	.151	.082	.042	.021	.010	.000	.000
	4	.947	.815	.629	.440	.285	.173	.100	.055	.029	.001	.000
	5	.983	.916	.785	.616	.446	.301	.191	.116	.067	.003	.000
	6	.995	.966	.889	.762	.606	.450	.313	.207	.130	.008	.000
	7	.999	.988	.949	.867	.744	.599	.453	.324	.220	.018	.001
	8	1.000	.996	.979	.932	.847	.729	.593	.456	.333	.037	.002
	9		.999	.992	.968	.916	.830	.717	.587	.458	.070	.005
	10		1.000	.997	.986	.957	.901	.816	.706	.583	.118	.011
	11			.999	.995	.980	.947	.888	.803	.697	.185	.021
	12			1.000	.998	.991	.973	.936	.876	.792	.268	.039
	13				.999	.996	.987	.966	.926	.864	.363	.066
	14				1.000	.999	.994	.983	.959	.917	.466	.105
	15					.999	.998	.992	.978	.951	.568	.157
	16					1.000	.999	.996	.989	.973	.664	.221
	17						1.000	.998	.995	.986	.749	.297
	18							.999	.998	.993	.819	.381
	19							1.000	.999	.997	.875	.470
	20								1.000	.998	.917	.559
	21									.999	.947	.644
	22									1.000	.967	.721
	23										.981	.787
	24										.989	.843
	25										.994	.888
	26										.997	.922
	27										.998	.948
	28										.999	.966
	29										1.000	.978
	30											.987
	31											.992
	32											.995
	33											.997
	34											.999
	35											.999
	36											1.000

Table A.3 Standard normal curve areas

$$\Phi(z) = P(Z \le z)$$

Standard normal density curve

Shaded area = $\Phi(z)$
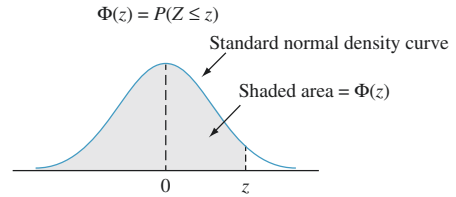
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3482
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

(continued)

Table A.3 (continued)

									$\Phi(z) = P(Z \leq z)$
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table A.4 The incomplete gamma function

$$G(x;\alpha) = \int_0^x \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y} dy$$

	α									
x	1	2	3	4	5	6	7	8	9	10
1	.632	.264	.080	.019	.004	.001	.000	.000	.000	.000
2	.865	.594	.323	.143	.053	.017	.005	.001	.000	.000
3	.950	.801	.577	.353	.185	.084	.034	.012	.004	.001
4	.982	.908	.762	.567	.371	.215	.111	.051	.021	.008
5	.993	.960	.875	.735	.560	.384	.238	.133	.068	.032
6	.998	.983	.938	.849	.715	.554	.394	.256	.153	.084
7	.999	.993	.970	.918	.827	.699	.550	.401	.271	.170
8	1.000	.997	.986	.958	.900	.809	.687	.547	.407	.283
9		.999	.994	.979	.945	.884	.793	.676	.544	.413
10		1.000	.997	.990	.971	.933	.870	.780	.667	.542
11			.999	.995	.985	.962	.921	.857	.768	.659
12			1.000	.998	.992	.980	.954	.911	.845	.758
13				.999	.996	.989	.974	.946	.900	.834
14				1.000	.998	.994	.986	.968	.938	.891
15					.999	.997	.992	.982	.963	.930

Table A.5 Critical values for chi-squared distributions

χ^2_v density curve

Shaded area $= \alpha$

$\chi^2_{\alpha,v}$
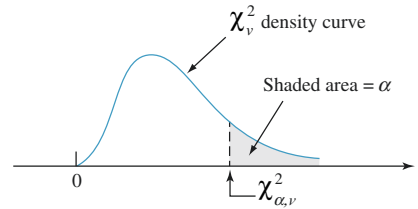
						α				
v	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.478	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.047	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	25.695	28.196	50.660	54.572	58.119	62.426	65.473
40	20.706	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766

For $v > 40$, $\chi^2_{\alpha,v} \approx v\left(1 - \dfrac{2}{9v} + z_\alpha \sqrt{\dfrac{2}{9v}}\right)^3$

Table A.6 Critical values for t distributions
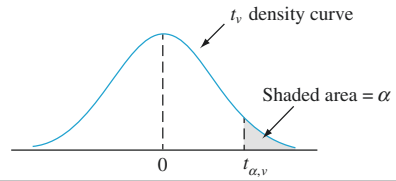
				α			
v	.10	.05	.025	.01	.005	.001	.0005
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
32	1.309	1.694	2.037	2.449	2.738	3.365	3.622
34	1.307	1.691	2.032	2.441	2.728	3.348	3.601
36	1.306	1.688	2.028	2.434	2.719	3.333	3.582
38	1.304	1.686	2.024	2.429	2.712	3.319	3.566
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	3.262	3.496
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Table A.7 *t* curve tail areas

t_v density curve — Area to the right of *t*
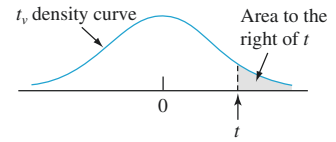
t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0.0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.1	.468	.465	.463	.463	.462.	.462	.462	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461
0.2	.437	.430	.427	.426	.425	.424	.424	.423	.423	.423	.423	.422	.422	.422	.422	.422	.422	.422
0.3	.407	.396	.392	.390	.388	.387	.386	.386	.386	.385	.385	.385	.384	.384	.384	.384	.384	.384
0.4	.379	.364	.358	.355	.353	.352	.351	.350	.349	.349	.348	.348	.348	.347	.347	.347	.347	.347
0.5	.352	.333	.326	.322	.319	.317	.316	.315	.315	.314	.313	.313	.313	.312	.312	.312	.312	.312
0.6	.328	.305	.295	.290	.287	.285	.284	.283	.282	.281	.280	.280	.279	.279	.279	.278	.278	.278
0.7	.306	.278	.267	.261	.258	.255	.253	.252	.251	.250	.249	.249	.248	.247	.247	.247	.247	.246
0.8	.285	.254	.241	.234	.230	.227	.225	.223	.222	.221	.220	.220	.219	.218	.218	.218	.217	.217
0.9	.267	.232	.217	.210	.205	.201	.199	.197	.196	.195	.194	.193	.192	.191	.191	.191	.190	.190
1.0	.250	.211	.196	.187	.182	.178	.175	.173	.172	.170	.169	.169	.168	.167	.167	.166	.166	.165
1.1	.235	.193	.176	.167	.162	.157	.154	.152	.150	.149	.147	.146	.146	.144	.144	.144	.143	.143
1.2	.221	.177	.158	.148	.142	.138	.135	.132	.130	.129	.128	.127	.126	.124	.124	.124	.123	.123
1.3	.209	.162	.142	.132	.125	.121	.117	.115	.113	.111	.110	.109	.108	.107	.107	.106	.105	.105
1.4	.197	.148	.128	.117	.110	.106	.102	.100	.098	.096	.095	.093	.092	.091	.091	.090	.090	.089
1.5	.187	.136	.115	.104	.097	.092	.089	.086	.084	.082	.081	.080	.079	.077	.077	.077	.076	.075
1.6	.178	.125	.104	.092	.085	.080	.077	.074	.072	.070	.069	.068	.067	.065	.065	.065	.064	.064
1.7	.169	.116	.094	.082	.075	.070	.065	.064	.062	.060	.059	.057	.056	.055	.055	.054	.054	.053
1.8	.161	.107	.085	.073	.066	.061	.057	.055	.053	.051	.050	.049	.048	.046	.046	.045	.045	.044
1.9	.154	.099	.077	.065	.058	.053	.050	.047	.045	.043	.042	.041	.040	.038	.038	.038	.037	.037
2.0	.148	.092	.070	.058	.051	.046	.043	.040	.038	.037	.035	.034	.033	.032	.032	.031	.031	.030
2.1	.141	.085	.063	.052	.045	.040	.037	.034	.033	.031	.030	.029	.028	.027	.027	.026	.025	.025
2.2	.136	.079	.058	.046	.040	.035	.032	.029	.028	.026	.025	.024	.023	.022	.022	.021	.021	.021
2.3	.131	.074	.052	.041	.035	.031	.027	.025	.023	.022	.021	.020	.019	.018	.018	.018	.017	.017
2.4	.126	.069	.048	.037	.031	.027	.024	.022	.020	.019	.018	.017	.016	.015	.015	.014	.014	.014
2.5	.121	.065	.044	.033	.027	.023	.020	.018	.017	.016	.015	.014	.013	.012	.012	.012	.011	.011
2.6	.117	.061	.040	.030	.024	.020	.018	.016	.014	.013	.012	.012	.011	.010	.010	.010	.009	.009
2.7	.113	.057	.037	.027	.021	.018	.015	.014	.012	.011	.010	.010	.009	.008	.008	.008	.008	.007
2.8	.109	.054	.034	.024	.019	.016	.013	.012	.010	.009	.009	.008	.008	.007	.007	.006	.006	.006
2.9	.106	.051	.031	.022	.017	.014	.011	.010	.009	.008	.007	.007	.006	.005	.005	.005	.005	.005
3.0	.102	.048	.029	.020	.015	.012	.010	.009	.007	.007	.006	.006	.005	.004	.004	.004	.004	.004
3.1	.099	.045	.027	.018	.013	.011	.009	.007	.006	.006	.005	.005	.004	.004	.004	.003	.003	.003
3.2	.096	.043	.025	.016	.012	.009	.008	.006	.005	.005	.004	.004	.003	.003	.003	.003	.003	.002
3.3	.094	.040	.023	.015	.011	.008	.007	.005	.005	.004	.004	.003	.003	.002	.002	.002	.002	.002
3.4	.091	.038	.021	.014	.010	.007	.006	.005	.004	.003	.003	.003	.002	.002	.002	.002	.002	.002
3.5	.089	.036	.020	.012	.009	.006	.005	.004	.003	.003	.002	.002	.002	.002	.002	.001	.001	.001
3.6	.086	.035	.018	.011	.008	.006	.004	.004	.003	.002	.002	.002	.002	.001	.001	.001	.001	.001
3.7	.084	.033	.017	.010	.007	.005	.004	.003	.002	.002	.002	.002	.001	.001	.001	.001	.001	.001
3.8	.082	.031	.016	.010	.006	.004	.003	.003	.002	.002	.001	.001	.001	.001	.001	.001	.001	.001
3.9	.080	.030	.015	.009	.006	.004	.003	.002	.002	.001	.001	.001	.001	.001	.001	.001	.001	.001
4.0	.078	.029	.014	.008	.005	.004	.003	.002	.002	.001	.001	.001	.001	.001	.001	.001	.000	.000

(continued)

Table A.7 (continued)

t	19	20	21	22	23	24	25	26	27	28	29	30	35	40	60	120	∞ (=z)
										v							
0.0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.1	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.460	.460	.460	.460	.460
0.2	.422	.422	.422	.422	.422	.422	.422	.422	.421	.421	.421	.421	.421	.421	.421	.421	.421
0.3	.384	.384	.384	.383	.383	.383	.383	.383	.383	.383	.383	.383	.383	.383	.383	.382	.382
0.4	.347	.347	.347	.347	.346	.346	.346	.346	.346	.346	.346	.346	.346	.346	.345	.345	.345
0.5	.311	.311	.311	.311	.311	.311	.311	.311	.311	.310	.310	.310	.310	.310	.309	.309	.309
0.6	.278	.278	.278	.277	.277	.277	.277	.277	.277	.277	.277	.277	.276	.276	.275	.275	.274
0.7	.246	.246	.246	.246	.245	.245	.245	.245	.245	.245	.245	.245	.244	.244	.243	.243	.242
0.8	.217	.217	.216	.216	.216	.216	.216	.215	.215	.215	.215	.215	.215	.214	.213	.213	.212
0.9	.190	.189	.189	.189	.189	.189	.188	.188	.188	.188	.188	.188	.187	.187	.186	.185	.184
1.0	.165	.165	.164	.164	.164	.164	.163	.163	.163	.163	.163	.163	.162	.162	.161	.160	.159
1.1	.143	.142	.142	.142	.141	.141	.141	.141	.141	.140	.140	.140	.139	.139	.138	.137	.136
1.2	.122	.122	.122	.121	.121	.121	.121	.120	.120	.120	.120	.120	.119	.119	.117	.116	.115
1.3	.105	.104	.104	.104	.103	.103	.103	.103	.102	.102	.102	.102	.101	.101	.099	.098	.097
1.4	.089	.089	.088	.088	.087	.087	.087	.087	.086	.086	.086	.086	.085	.085	.083	.082	.081
1.5	.075	.075	.074	.074	.074	.073	.073	.073	.073	.072	.072	.072	.071	.071	.069	.068	.067
1.6	.063	.063	.062	.062	.062	.061	.061	.061	.061	.060	.060	.060	.059	.059	.057	.056	.055
1.7	.053	.052	.052	.052	.051	.051	.051	.051	.050	.050	.050	.050	.049	.048	.047	.046	.045
1.8	.044	.043	.043	.043	.042	.042	.042	.042	.042	.041	.041	.041	.040	.040	.038	.037	.036
1.9	.036	.036	.036	.035	.035	.035	.035	.034	.034	.034	.034	.034	.033	.032	.031	.030	.029
2.0	.030	.030	.029	.029	.029	.028	.028	.028	.028	.028	.027	.027	.027	.026	.025	.024	.023
2.1	.025	.024	.024	.024	.023	.023	.023	.023	.023	.022	.022	.022	.022	.021	.020	.019	.018
2.2	.020	.020	.020	.019	.019	.019	.019	.018	.018	.018	.018	.018	.017	.017	.016	.015	.014
2.3	.016	.016	.016	.016	.015	.015	.015	.015	.015	.015	.014	.014	.014	.013	.012	.012	.011
2.4	.013	.013	.013	.013	.012	.012	.012	.012	.012	.012	.012	.011	.011	.011	.010	.009	.008
2.5	.011	.011	.010	.010	.010	.010	.010	.010	.009	.009	.009	.009	.009	.008	.008	.007	.006
2.6	.009	.009	.008	.008	.008	.008	.008	.008	.007	.007	.007	.007	.007	.007	.006	.005	.005
2.7	.007	.007	.007	.007	.006	.006	.006	.006	.006	.006	.006	.006	.005	.005	.004	.004	.003
2.8	.006	.006	.005	.005	.005	.005	.005	.005	.005	.005	.005	.004	.004	.004	.003	.003	.003
2.9	.005	.004	.004	.004	.004	.004	.004	.004	.004	.004	.004	.003	.003	.003	.003	.002	.002
3.0	.004	.004	.003	.003	.003	.003	.003	.003	.003	.003	.003	.003	.002	.002	.002	.002	.001
3.1	.003	.003	.003	.003	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002	.001	.001	.001
3.2	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.001	.001	.001	.001	.001
3.3	.002	.002	.002	.002	.002	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.000
3.4	.002	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.000	.000
3.5	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.000	.000	.000
3.6	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.000	.000	.000	.000	.000
3.7	.001	.001	.001	.001	.001	.001	.001	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
3.8	.001	.001	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
3.9	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
4.0	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Table A.8 Critical values for F distributions

		α	1	2	3	4	5	6	7	8	9
						v_1 = numerator df					
	1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
		.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
		.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
		.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
	2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
		.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
		.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
		.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
	3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
		.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
		.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
		.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
	4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
		.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
		.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
		.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
v_2 = denominator df	5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
		.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
		.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
		.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24
	6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
		.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
		.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
		.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69
	7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
		.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
		.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
		.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33
	8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
		.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
		.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
		.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77
	9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
		.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
		.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
		.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11
	10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
		.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
		.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
		.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96
	11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
		.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
		.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
		.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12
	12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
		.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
		.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
		.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48

(continued)

Table A.8 (continued)

10	12	15	20	25	30	40	50	60	120	1000
60.19	60.71	61.22	61.74	62.05	62.26	62.53	62.69	62.79	63.06	63.30
241.88	243.91	245.95	248.01	249.26	250.10	251.14	251.77	252.20	253.25	254.19
6055.8	6106.3	6157.3	6208.7	6239.8	6260.6	6286.8	6302.5	6313.0	6339.4	6362.7
605621	610668	615764	620908	624017	626099	628712	630285	631337	633972	636301
9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.49
19.40	19.41	19.43	19.45	19.46	19.46	19.47	19.48	19.48	19.49	19.49
99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.48	99.49	99.50
999.40	999.42	999.43	999.45	999.46	999.47	999.47	999.48	999.48	999.49	999.50
5.23	5.22	5.20	5.18	5.17	5.17	5.16	5.15	5.15	5.14	5.13
8.79	8.74	8.70	8.66	8.63	8.62	8.59	8.58	8.57	8.55	8.53
27.23	27.05	26.87	26.69	26.58	26.50	26.41	26.35	26.32	26.22	26.14
129.25	128.32	127.37	126.42	125.84	125.45	124.96	124.66	124.47	123.97	123.53
3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.76
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.70	5.69	5.66	5.63
14.55	14.37	14.20	14.02	13.91	13.84	13.75	13.69	13.65	13.56	13.47
48.05	47.41	46.76	46.10	45.70	45.43	45.09	44.88	44.75	44.40	44.09
3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.12	3.11
4.74	4.68	4.62	4.56	4.52	4.50	4.46	4.44	4.43	4.40	4.37
10.05	9.89	9.72	9.55	9.45	9.38	9.29	9.24	9.20	9.11	9.03
26.92	26.42	25.91	25.39	25.08	24.87	24.60	24.44	24.33	24.06	23.82
2.94	2.90	2.87	2.84	2.81	2.80	2.78	2.77	2.76	2.74	2.72
4.06	4.00	3.94	3.87	3.83	3.81	3.77	3.75	3.74	3.70	3.67
7.87	7.72	7.56	7.40	7.30	7.23	7.14	7.09	7.06	6.97	6.89
18.41	17.99	17.56	17.12	16.85	16.67	16.44	16.31	16.21	15.98	15.77
2.70	2.67	2.63	2.59	2.57	2.56	2.54	2.52	2.51	2.49	2.47
3.64	3.57	3.51	3.44	3.40	3.38	3.34	3.32	3.30	3.27	3.23
6.62	6.47	6.31	6.16	6.06	5.99	5.91	5.86	5.82	5.74	5.66
14.08	13.71	13.32	12.93	12.69	12.53	12.33	12.20	12.12	11.91	11.72
2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.30
3.35	3.28	3.22	3.15	3.11	3.08	3.04	3.02	3.01	2.97	2.93
5.81	5.67	5.52	5.36	5.26	5.20	5.12	5.07	5.03	4.95	4.87
11.54	11.19	10.84	10.48	10.26	10.11	9.92	9.80	9.73	9.53	9.36
2.42	2.38	2.34	2.30	2.27	2.25	2.23	2.22	2.21	2.18	2.16
3.14	3.07	3.01	2.94	2.89	2.86	2.83	2.80	2.79	2.75	2.71
5.26	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.48	4.40	4.32
9.89	9.57	9.24	8.90	8.69	8.55	8.37	8.26	8.19	8.00	7.84
2.32	2.28	2.24	2.20	2.17	2.16	2.13	2.12	2.11	2.08	2.06
2.98	2.91	2.85	2.77	2.73	2.70	2.66	2.64	2.62	2.58	2.54
4.85	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.08	4.00	3.92
8.75	8.45	8.13	7.80	7.60	7.47	7.30	7.19	7.12	6.94	6.78
2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.00	1.98
2.85	2.79	2.72	2.65	2.60	2.57	2.53	2.51	2.49	2.45	2.41
4.54	4.40	4.25	4.10	4.01	3.94	3.86	3.81	3.78	3.69	3.61
7.92	7.63	7.32	7.01	6.81	6.68	6.52	6.42	6.35	6.18	6.02
2.19	2.15	2.10	2.06	2.03	2.01	1.99	1.97	1.96	1.93	1.91
2.75	2.69	2.62	2.54	2.50	2.47	2.43	2.40	2.38	2.34	2.30
4.30	4.16	4.01	3.86	3.76	3.70	3.62	3.57	3.54	3.45	3.37
7.29	7.00	6.71	6.40	6.22	6.09	5.93	5.83	5.76	5.59	5.44

(continued)

Table A.8 (continued)

| | | | v_1 = numerator df | | | | | | | |
	α	1	2	3	4	5	6	7	8	9
	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
13	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98
	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
14	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58
	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
15	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26
	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
16	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98
v_2 =	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
denominator	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
df 17	.010	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75
	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
18	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
19	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
20	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
21	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11
	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
22	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99
	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
23	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89
	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
24	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80

(continued)

Table A.8 (continued)

					v_1 = numerator df					
10	12	15	20	25	30	40	50	60	120	1000
2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.85
2.67	2.60	2.53	2.46	2.41	2.38	2.34	2.31	2.30	2.25	2.21
4.10	3.96	3.82	3.66	3.57	3.51	3.43	3.38	3.34	3.25	3.18
6.80	6.52	6.23	5.93	5.75	5.63	5.47	5.37	5.30	5.14	4.99
2.10	2.05	2.01	1.96	1.93	1.91	1.89	1.87	1.86	1.83	1.80
2.60	2.53	2.46	2.39	2.34	2.31	2.27	2.24	2.22	2.18	2.14
3.94	3.80	3.66	3.51	3.41	3.35	3.27	3.22	3.18	3.09	3.02
6.40	6.13	5.85	5.56	5.38	5.25	5.10	5.00	4.94	4.77	4.62
2.06	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.82	1.79	1.76
2.54	2.48	2.40	2.33	2.28	2.25	2.20	2.18	2.16	2.11	2.07
3.80	3.67	3.52	3.37	3.28	3.21	3.13	3.08	3.05	2.96	2.88
6.08	5.81	5.54	5.25	5.07	4.95	4.80	4.70	4.64	4.47	4.33
2.03	1.99	1.94	1.89	1.86	1.84	1.81	1.79	1.78	1.75	1.72
2.49	2.42	2.35	2.28	2.23	2.19	2.15	2.12	2.11	2.06	2.02
3.69	3.55	3.41	3.26	3.16	3.10	3.02	2.97	2.93	2.84	2.76
5.81	5.55	5.27	4.99	4.82	4.70	4.54	4.45	4.39	4.23	4.08
2.00	1.96	1.91	1.86	1.83	1.81	1.78	1.76	1.75	1.72	1.69
2.45	2.38	2.31	2.23	2.18	2.15	2.10	2.08	2.06	2.01	1.97
3.59	3.46	3.31	3.16	3.07	3.00	2.92	2.87	2.83	2.75	2.66
5.58	5.32	5.05	4.78	4.60	4.48	4.33	4.24	4.18	4.02	3.87
1.98	1.93	1.89	1.84	1.80	1.78	1.75	1.74	1.72	1.69	1.66
2.41	2.34	2.27	2.19	2.14	2.11	2.06	2.04	2.02	1.97	1.92
3.51	3.37	3.23	3.08	2.98	2.92	2.84	2.78	2.75	2.66	2.58
5.39	5.13	4.87	4.59	4.42	4.30	4.15	4.06	4.00	3.84	3.69
1.96	1.91	1.86	1.81	1.78	1.76	1.73	1.71	1.70	1.67	1.64
2.38	2.31	2.23	2.16	2.11	2.07	2.03	2.00	1.98	1.93	1.88
3.43	3.30	3.15	3.00	2.91	2.84	2.76	2.71	2.67	2.58	2.50
5.22	4.97	4.70	4.43	4.26	4.14	3.99	3.90	3.84	3.68	3.53
1.94	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.68	1.64	1.61
2.35	2.28	2.20	2.12	2.07	2.04	1.99	1.97	1.95	1.90	1.85
3.37	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.61	2.52	2.43
5.08	4.82	4.56	4.29	4.12	4.00	3.86	3.77	3.70	3.54	3.40
1.92	1.87	1.83	1.78	1.74	1.72	1.69	1.67	1.66	1.62	1.59
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.94	1.92	1.87	1.82
3.31	3.17	3.03	2.88	2.79	2.72	2.64	2.58	2.55	2.46	2.37
4.95	4.70	4.44	4.17	4.00	3.88	3.74	3.64	3.58	3.42	3.28
1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.60	1.57
2.30	2.23	2.15	2.07	2.02	1.98	1.94	1.91	1.89	1.84	1.79
3.26	3.12	2.98	2.83	2.73	2.67	2.58	2.53	2.50	2.40	2.32
4.83	4.58	4.33	4.06	3.89	3.78	3.63	3.54	3.48	3.32	3.17
1.89	1.84	1.80	1.74	1.71	1.69	1.66	1.64	1.62	1.59	1.55
2.27	2.20	2.13	2.05	2.00	1.96	1.91	1.88	1.86	1.81	1.76
3.21	3.07	2.93	2.78	2.69	2.62	2.54	2.48	2.45	2.35	2.27
4.73	4.48	4.23	3.96	3.79	3.68	3.53	3.44	3.38	3.22	3.08
1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.57	1.54
2.25	2.18	2.11	2.03	1.97	1.94	1.89	1.86	1.84	1.79	1.74
3.17	3.03	2.89	2.74	2.64	2.58	2.49	2.44	2.40	2.31	2.22
4.64	4.39	4.14	3.87	3.71	3.59	3.45	3.36	3.29	3.14	2.99

(continued)

Table A.8 (continued)

						v_1 = numerator df				
		α	1	2	3	4	5	6	7	8	9
	25	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
		.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
		.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
		.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71
	26	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
		.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
		.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
		.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64
	27	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
		.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
		.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
		.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57
	28	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
		.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
		.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
		.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50
v_2 = denominator df	29	.100	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86
		.050	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
		.010	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09
		.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45
	30	.100	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
		.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
		.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
		.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
	40	.100	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79
		.050	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
		.010	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
		.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02
	50	.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76
		.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
		.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
		.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82
	60	.100	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
		.050	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
		.010	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
		.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
	100	.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69
		.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
		.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
		.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44
	200	.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66
		.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
		.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
		.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26
	1000	.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64
		.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
		.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
		.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13

(continued)

Table A.8 (continued)

				v_1 = numerator df						
10	12	15	20	25	30	40	50	60	120	1000
1.87	1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.59	1.56	1.52
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.82	1.77	1.72
3.13	2.99	2.85	2.70	2.60	2.54	2.45	2.40	2.36	2.27	2.18
4.56	4.31	4.06	3.79	3.63	3.52	3.37	3.28	3.22	3.06	2.91
1.86	1.81	1.76	1.71	1.67	1.65	1.61	1.59	1.58	1.54	1.51
2.22	2.15	2.07	1.99	1.94	1.90	1.85	1.82	1.80	1.75	1.70
3.09	2.96	2.81	2.66	2.57	2.50	2.42	2.36	2.33	2.23	2.14
4.48	4.24	3.99	3.72	3.56	3.44	3.30	3.21	3.15	2.99	2.84
1.85	1.80	1.75	1.70	1.66	1.64	1.60	1.58	1.57	1.53	1.50
2.20	2.13	2.06	1.97	1.92	1.88	1.84	1.81	1.79	1.73	1.68
3.06	2.93	2.78	2.63	2.54	2.47	2.38	2.33	2.29	2.20	2.11
4.41	4.17	3.92	3.66	3.49	3.38	3.23	3.14	3.08	2.92	2.78
1.84	1.79	1.74	1.69	1.65	1.63	1.59	1.57	1.56	1.52	1.48
2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.71	1.66
3.03	2.90	2.75	2.60	2.51	2.44	2.35	2.30	2.26	2.17	2.08
4.35	4.11	3.86	3.60	3.43	3.32	3.18	3.09	3.02	2.86	2.72
1.83	1.78	1.73	1.68	1.64	1.62	1.58	1.56	1.55	1.51	1.47
2.18	2.10	2.03	1.94	1.89	1.85	1.81	1.77	1.75	1.70	1.65
3.00	2.87	2.73	2.57	2.48	2.41	2.33	2.27	2.23	2.14	2.05
4.29	4.05	3.80	3.54	3.38	3.27	3.12	3.03	2.97	2.81	2.66
1.82	1.77	1.72	1.67	1.63	1.61	1.57	1.55	1.54	1.50	1.46
2.16	2.09	2.01	1.93	1.88	1.84	1.79	1.76	1.74	1.68	1.63
2.98	2.84	2.70	2.55	2.45	2.39	2.30	2.25	2.21	2.11	2.02
4.24	4.00	3.75	3.49	3.33	3.22	3.07	2.98	2.92	2.76	2.61
1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.42	1.38
2.08	2.00	1.92	1.84	1.78	1.74	1.69	1.66	1.64	1.58	1.52
2.80	2.66	2.52	2.37	2.27	2.20	2.11	2.06	2.02	1.92	1.82
3.87	3.64	3.40	3.14	2.98	2.87	2.73	2.64	2.57	2.41	2.25
1.73	1.68	1.63	1.57	1.53	1.50	1.46	1.44	1.42	1.38	1.33
2.03	1.95	1.87	1.78	1.73	1.69	1.63	1.60	1.58	1.51	1.45
2.70	2.56	2.42	2.27	2.17	2.10	2.01	1.95	1.91	1.80	1.70
3.67	3.44	3.20	2.95	2.79	2.68	2.53	2.44	2.38	2.21	2.05
1.71	1.66	1.60	1.54	1.50	1.48	1.44	1.41	1.40	1.35	1.30
1.99	1.92	1.84	1.75	1.69	1.65	1.59	1.56	1.53	1.47	1.40
2.63	2.50	2.35	2.20	2.10	2.03	1.94	1.88	1.84	1.73	1.62
3.54	3.32	3.08	2.83	2.67	2.55	2.41	2.32	2.25	2.08	1.92
1.66	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.34	1.28	1.22
1.93	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.45	1.38	1.30
2.50	2.37	2.22	2.07	1.97	1.89	1.80	1.74	1.69	1.57	1.45
3.30	3.07	2.84	2.59	2.43	2.32	2.17	2.08	2.01	1.83	1.64
1.63	1.58	1.52	1.46	1.41	1.38	1.34	1.31	1.29	1.23	1.16
1.88	1.80	1.72	1.62	1.56	1.52	1.46	1.41	1.39	1.30	1.21
2.41	2.27	2.13	1.97	1.87	1.79	1.69	1.63	1.58	1.45	1.30
3.12	2.90	2.67	2.42	2.26	2.15	2.00	1.90	1.83	1.64	1.43
1.61	1.55	1.49	1.43	1.38	1.35	1.30	1.27	1.25	1.18	1.08
1.84	1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.33	1.24	1.11
2.34	2.20	2.06	1.90	1.79	1.72	1.61	1.54	1.50	1.35	1.16
2.99	2.77	2.54	2.30	2.14	2.02	1.87	1.77	1.69	1.49	1.22

Table A.9 Critical values for studentized range distributions

v	α	2	3	4	5	6	7	8	9	10	11	12
						m						
5	.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32
	.01	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48	10.70
6	.05	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65	6.79
	.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.48
7	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43
	.01	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55	8.71
8	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18
	.01	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03	8.18
9	.05	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87	5.98
	.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	7.65	7.78
10	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83
	.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.49
11	.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71
	.01	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25
12	.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51	5.61
	.01	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06
13	.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53
	.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90
14	.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46
	.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77
15	.05	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40
	.01	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66
16	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35
	.01	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56
17	.05	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21	5.31
	.01	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48
18	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27
	.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41
19	.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23
	.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34
20	.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20
	.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.28
24	.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10
	.01	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11
30	.05	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00
	.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93
40	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82	4.90
	.01	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.69	5.76
60	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81
	.01	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60
120	.05	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71
	.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.37	5.44
∞	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62
	.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29

Table A.10 Chi-squared curve tail areas

Upper-tail area	$v = 1$	$v = 2$	$v = 3$	$v = 4$	$v = 5$
>.100	<2.70	<4.60	<6.25	<7.77	<9.23
.100	2.70	4.60	6.25	7.77	9.23
.095	2.78	4.70	6.36	7.90	9.37
.090	2.87	4.81	6.49	8.04	9.52
.085	2.96	4.93	6.62	8.18	9.67
.080	3.06	5.05	6.75	8.33	9.83
.075	3.17	5.18	6.90	8.49	10.00
.070	3.28	5.31	7.06	8.66	10.19
.065	3.40	5.46	7.22	8.84	10.38
.060	3.53	5.62	7.40	9.04	10.59
.055	3.68	5.80	7.60	9.25	10.82
.050	3.84	5.99	7.81	9.48	11.07
.045	4.01	6.20	8.04	9.74	11.34
.040	4.21	6.43	8.31	10.02	11.64
.035	4.44	6.70	8.60	10.34	11.98
.030	4.70	7.01	8.94	10.71	12.37
.025	5.02	7.37	9.34	11.14	12.83
.020	5.41	7.82	9.83	11.66	13.38
.015	5.91	8.39	10.46	12.33	14.09
.010	6.63	9.21	11.34	13.27	15.08
.005	7.87	10.59	12.83	14.86	16.74
.001	10.82	13.81	16.26	18.46	20.51
<.001	>10.82	>13.81	>16.26	>18.46	>20.51
Upper-tail area	$v = 6$	$v = 7$	$v = 8$	$v = 9$	$v = 10$
>.100	<10.64	<12.01	<13.36	<14.68	<15.98
.100	10.64	12.01	13.36	14.68	15.98
.095	10.79	12.17	13.52	14.85	16.16
.090	10.94	12.33	13.69	15.03	16.35
.085	11.11	12.50	13.87	15.22	16.54
.080	11.28	12.69	14.06	15.42	16.75
.075	11.46	12.88	14.26	15.63	16.97
.070	11.65	13.08	14.48	15.85	17.20
.065	11.86	13.30	14.71	16.09	17.44
.060	12.08	13.53	14.95	16.34	17.71
.055	12.33	13.79	15.22	16.62	17.99
.050	12.59	14.06	15.50	16.91	18.30
.045	12.87	14.36	15.82	17.24	18.64
.040	13.19	14.70	16.17	17.60	19.02
.035	13.55	15.07	16.56	18.01	19.44
.030	13.96	15.50	17.01	18.47	19.92
.025	14.44	16.01	17.53	19.02	20.48
.020	15.03	16.62	18.16	19.67	21.16
.015	15.77	17.39	18.97	20.51	22.02
.010	16.81	18.47	20.09	21.66	23.20
.005	18.54	20.27	21.95	23.58	25.18
.001	22.45	24.32	26.12	27.87	29.58
<.001	>22.45	>24.32	>26.12	>27.87	>29.58

(continued)

Table A.10 (continued)

Upper-tail area	v = 11	v = 12	v = 13	v = 14	v = 15
>.100	<17.27	<18.54	<19.81	<21.06	<22.30
.100	17.27	18.54	19.81	21.06	22.30
.095	17.45	18.74	20.00	21.26	22.51
.090	17.65	18.93	20.21	21.47	22.73
.085	17.85	19.14	20.42	21.69	22.95
.080	18.06	19.36	20.65	21.93	23.19
.075	18.29	19.60	20.89	22.17	23.45
.070	18.53	19.84	21.15	22.44	23.72
.065	18.78	20.11	21.42	22.71	24.00
.060	19.06	20.39	21.71	23.01	24.31
.055	19.35	20.69	22.02	23.33	24.63
.050	19.67	21.02	22.36	23.68	24.99
.045	20.02	21.38	22.73	24.06	25.38
.040	20.41	21.78	23.14	24.48	25.81
.035	20.84	22.23	23.60	24.95	26.29
.030	21.34	22.74	24.12	25.49	26.84
.025	21.92	23.33	24.73	26.11	27.48
.020	22.61	24.05	25.47	26.87	28.25
.015	23.50	24.96	26.40	27.82	29.23
.010	24.72	26.21	27.68	29.14	30.57
.005	26.75	28.29	29.81	31.31	32.80
.001	31.26	32.90	34.52	36.12	37.69
<.001	>31.26	>32.90	>34.52	>36.12	>37.69
Upper-tail area	v = 16	v = 17	v = 18	v = 19	v = 20
>.100	<23.54	<24.77	<25.98	<27.20	<28.41
.100	23.54	24.76	25.98	27.20	28.41
.095	23.75	24.98	26.21	27.43	28.64
.090	23.97	25.21	26.44	27.66	28.88
.085	24.21	25.45	26.68	27.91	29.14
.080	24.45	25.70	26.94	28.18	29.40
.075	24.71	25.97	27.21	28.45	29.69
.070	24.99	26.25	27.50	28.75	29.99
.065	25.28	26.55	27.81	29.06	30.30
.060	25.59	26.87	28.13	29.39	30.64
.055	25.93	27.21	28.48	29.75	31.01
.050	26.29	27.58	28.86	30.14	31.41
.045	26.69	27.99	29.28	30.56	31.84
.040	27.13	28.44	29.74	31.03	32.32
.035	27.62	28.94	30.25	31.56	32.85
.030	28.19	29.52	30.84	32.15	33.46
.025	28.84	30.19	31.52	32.85	34.16
.020	29.63	30.99	32.34	33.68	35.01
.015	30.62	32.01	33.38	34.74	36.09
.010	32.00	33.40	34.80	36.19	37.56
.005	34.26	35.71	37.15	38.58	39.99
.001	39.25	40.78	42.31	43.81	45.31
<.001	>39.25	>40.78	>47.31	>43.81	>45.31

Table A.11 Critical values for the Wilcoxon signed-ranked test

			$P_0(S_+ \geq c_1) = P(S_+ \geq c_1$ when H_0 is true)		
n	c_1	$P_0(S_+ \geq c_1)$	n	c_1	$P_0(S_+ \geq c_1)$
3	6	.125		78	.011
4	9	.125		79	.009
	10	.062		81	.005
5	13	.094	14	73	.108
	14	.062		74	.097
	15	.031		79	.052
6	17	.109		84	.025
	19	.047		89	.010
	20	.031		92	.005
	21	.016	15	83	.104
7	22	.109		84	.094
	24	.055		89	.053
	26	.023		90	.047
	28	.008		95	.024
8	28	.098		100	.011
	30	.055		101	.009
	32	.027		104	.005
	34	.012	16	93	.106
	35	.008		94	.096
	36	.004		100	.052
9	34	.102		106	.025
	37	.049		112	.011
	39	.027		113	.009
	42	.010		116	.005
	44	.004	17	104	.103
10	41	.097		105	.095
	44	.053		112	.049
	47	.024		118	.025
	50	.010		125	.010
	52	.005		129	.005
11	48	.103	18	116	.098
	52	.051		124	.049
	55	.027		131	.024
	59	.009		138	.010
	61	.005		143	.005
12	56	.102	19	128	.098
	60	.055		136	.052
	61	.046		137	.048
	64	.026		144	.025
	68	.010		152	.010
	71	.005		157	.005
13	64	.108	20	140	.101
	65	.095		150	.049
	69	.055		158	.024
	70	.047		167	.010
	74	.024		172	.005

Table A.12 Critical values for the Wilcoxon signed-rank interval

							$\left(\overline{x}_{(n(n+1)/2)-c+1}, \overline{x}_{(c)}\right)$	
n	Confidence level (%)	c	n	Confidence level (%)	c	n	Confidence level (%)	c
5	93.8	15	13	99.0	81	20	99.1	173
	87.5	14		95.2	74		95.2	158
6	96.9	21		90.6	70		90.3	150
	93.7	20	14	99.1	93	21	99.0	188
	90.6	19		95.1	84		95.0	172
7	98.4	28		89.6	79		89.7	163
	95.3	26	15	99.0	104	22	99.0	204
	89.1	24		95.2	95		95.0	187
8	99.2	36		90.5	90		90.2	178
	94.5	32	16	99.1	117	23	99.0	221
	89.1	30		94.9	106		95.2	203
9	99.2	44		89.5	100		90.2	193
	94.5	39	17	99.1	130	24	99.0	239
	90.2	37		94.9	118		95.1	219
10	99.0	52		90.2	112		89.9	208
	95.1	47	18	99.0	143	25	99.0	257
	89.5	44		95.2	131		95.2	236
11	99.0	61		90.1	124		89.9	224
	94.6	55	19	99.1	158			
	89.8	52		95.1	144			
12	99.1	71		90.4	137			
	94.8	64						
	90.8	61						

Table A.13 Critical values for the Wilcoxon rank-sum test

			$P_0(W \geq c) = P(W \geq c$ when H_0 is true)				
m	n	c	$P_0(W \geq c)$	m	n	c	$P_0(W \geq c)$
3	3	15	.05			40	.004
	4	17	.057		6	40	.041
		18	.029			41	.026
	5	20	.036			43	.009
		21	.018			44	.004
	6	22	.048		7	43	.053
		23	.024			45	.024
		24	.012			47	.009
	7	24	.058			48	.005
		26	.017		8	47	.047
		27	.008			49	.023
	8	27	.042			51	.009
		28	.024			52	.005
		29	.012	6	6	50	.047
		30	.006			52	.021
4	4	24	.057			54	.008
		25	.029			55	.004
		26	.014		7	54	.051
	5	27	.056			56	.026
		28	.032			58	.011
		29	.016			60	.004
		30	.008		8	58	.054
	6	30	.057			61	.021
		32	.019			63	.01
		33	.010			65	.004
		34	.005	7	7	66	.049
	7	33	.055			68	.027
		35	.021			71	.009
		36	.012			72	.006
		37	.006		8	71	.047
	8	36	.055			73	.027
		38	.024			76	.01
		40	.008			78	.005
		41	.004	8	8	84	.052
5	5	36	.048			87	.025
		37	.028			90	.01
		39	.008			92	.005

Table A.14 Critical values for the Wilcoxon rank-sum interval

$$\left(d_{ij(mn-c+1)}, d_{ij(c)}\right)$$

	Smaller sample size							
Larger sample size	5		6		7		8	
	Confidence level (%)	c	Confidence level (%)	c	Confidence level (%)	c	Confidence level (%)	c
5	99.2	25						
	94.4	22						
	90.5	21						
6	99.1	29	99.1	34				
	94.8	26	95.9	31				
	91.8	25	90.7	29				
7	99.0	33	99.2	39	98.9	44		
	95.2	30	94.9	35	94.7	40		
	89.4	28	89.9	33	90.3	38		
8	98.9	37	99.2	44	99.1	50	99.0	56
	95.5	34	95.7	40	94.6	45	95.0	51
	90.7	32	89.2	37	90.6	43	89.5	48
9	98.8	41	99.2	49	99.2	56	98.9	62
	95.8	38	95.0	44	94.5	50	95.4	57
	88.8	35	91.2	42	90.9	48	90.7	54
10	99.2	46	98.9	53	99.0	61	99.1	69
	94.5	41	94.4	48	94.5	55	94.5	62
	90.1	39	90.7	46	89.1	52	89.9	59
11	99.1	50	99.0	58	98.9	66	99.1	75
	94.8	45	95.2	53	95.6	61	94.9	68
	91.0	43	90.2	50	89.6	57	90.9	65
12	99.1	54	99.0	63	99.0	72	99.0	81
	95.2	49	94.7	57	95.5	66	95.3	74
	89.6	46	89.8	54	90.0	62	90.2	70
	Smaller sample size							
Larger sample size	9		10		11		12	
	Confidence level (%)	c	Confidence level (%)	c	Confidence level (%)	c	Confidence level (%)	c
9	98.9	69						
	95.0	63						
	90.6	60						
10	99.0	76	99.1	84				
	94.7	69	94.8	76				
	90.5	66	89.5	72				
11	99.0	83	99.0	91	98.9	99		
	95.4	76	94.9	83	95.3	91		
	90.5	72	90.1	79	89.9	86		
12	99.1	90	99.1	99	99.1	108	99.0	116
	95.1	82	95.0	90	94.9	98	94.8	106
	90.5	78	90.7	86	89.6	93	89.9	101

Answers to Odd-Numbered Exercises

Chapter 1

1. a. *Houston Chronicle*, *Des Moines Register*, *Chicago Tribune*, *Washington Post*
   b. Capital One, Campbell Soup, Merrill Lynch, Pulitzer
   c. Bill Jasper, Kay Reinke, Helen Ford, David Menendez
   d. 1.78, 2.44, 3.50, 3.04

3. a. In a sample of 100 phones, what are the chances that more than 20 need service while under warranty? What are the chances than none need service while still under warranty?
   b. What proportion of *all* phones of this brand and model will need service within the warranty period?

5. a. Two variables (at least) were recorded: skin color and hourly wages.
   b. Skin color is categorical (with four categories), while hourly wages is quantitative (units: $/h).

7. a. categorical   b. quantitative
   c. categorical   d. categorical
   e. categorical

9. a. No, the relevant conceptual population is all scores of all students who participate in the SI in conjunction with this particular statistics course.
   b. The advantage to randomly assigning students to the two groups is that the two groups should then be fairly comparable before the study. If the two groups perform differently in the class, we can reasonably attribute this to the treatments (SI and control). If it were left to students to choose, stronger or more dedicated students might gravitate toward SI, confounding the results.
   c. If all students were put in the treatment group there would be no results with which to compare the treatments.

11. One could generate a simple random sample of all single-family homes in the city or a stratified random sample by taking a simple random sample from each of the 10 district neighborhoods. From each of the homes in the sample the necessary variables would be collected. This would be an enumerative study because there exists a finite, identifiable population of objects from which to sample.

13. a. There could be several explanations for the variability of the measurements. Among them could be measuring error, (due to mechanical or technical changes across measurements), recording error, differences in weather conditions at time of measurements, etc.
    b. This could be called conceptual because there is no sampling frame.

15. This display brings out the gap in the data: There are no scores in the high 70s.

```
6L  |  034
6H  |  667899
7L  |  00122244
7H  |                          Stem = tens
8L  |  001111122344            Leaf = ones
8H  |  5557899
9L  |  03
9H  |  58
```

17. a.
```
0  |  123333333444444
0  |  55555667788888999
1  |  0000001111224
1  |  5789                    Stem: tens digit
2  |  0112                    Leaf: ones digit
2  |  6
3  |  334
3  |  7
4  |  2
4  |  68
5  |  012
5  |
6  |
6  |  6
7  |
7  |  6
8  |  1
```

b. Arguably, a representative crack depth might be around 9–10 μm.

c. This is somewhat subjective, but the display appears quite spread out.

d. No, the distribution is certainly not symmetric. Rather, crack depths appear to be strongly positively skewed.

e. Yes: All of the values 66.5, 76.1, and 81.1 μm appear to be high outliers. (Using an outlier convention described later in the chapter, even the values in the 50s would be considered outliers!)

19.
```
   American            |    | French
                       | 8  | 1
   755543211000        | 9  | 00234566
   9432                | 10 | 2356
   6630                | 11 | 1369
   850                 | 12 | 223558
   8                   | 13 | 7
                       | 14 |
                       | 15 | 8
   2                   | 16 |   Stem:tens digit
                       |    |   Leaf: ones digit
```

The American distribution is positively skewed, but the French distribution is fairly symmetric. Almost half of the American movies are in the 90s, but the French movies are more spread out.

21. a.

Value	Freq.	Rel. Freq. (=Freq./60)
0	7	.117
1	12	.200
2	13	.217
3	14	.233
4	6	.100
5	3	.050
6	3	.050
7	1	.017
8	1	.017

Note Relative frequencies add to 1.001, not 1, due to rounding.

b. The number of batches with at most 5 nonconforming items is $7 + 12 + 13 + 14 + 6 + 3 = 55$, which is a proportion of $55/60 = .917$. The proportion of batches with (strictly) fewer than 5 nonconforming items is $52/60 = .867$.

Notice that these proportions could also have been computed by using the relative frequencies: e.g., proportion of batches with 5 or fewer nonconforming items = $1 - (.05 + .017 + .017) = .916$; proportion of batches with fewer than 5 nonconforming items = $1 - (.05 + .05 + .017 + .017) = .866$.

c. The center of the histogram is somewhere around 2 or 3 and it shows that there is some positive skewness in the data. The histogram also shows that there is a lot of spread/variation in this data.

23. a. $589/1570 = .375$.

b. $1 - (589 + 190 + 176 + 157 + 115)/1570 = .218$.

c. $(115 + 89 + 57 + 55 + 33 + 31)/1570 = .242$.

d. The herd size distribution in the accompanying histogram is extremely positively skewed.

25. a. From a histogram, the number of subdivisions having no cul-de-sacs (i.e., $y = 0$) is 17/47 = .362, or 36.2%. The proportion having at least one cul-de-sac ($y \geq 1$) is (47 − 17)/47 = 30/47 = .638, or 63.8%. Note that subtracting the number of cul-de-sacs with $y = 0$ from the total, 47, is an easy way to find the number of subdivisions with $y \geq 1$.
   b. From a histogram, the number of subdivisions with at most 5 intersections (i.e., $z \leq 5$) is 42/47 = .894, or 89.4%. The proportion having fewer than 5 intersections ($z < 5$) is 39/47 = .830, or 83.0%.

27. a. The distribution of these by-state values is slightly positively skewed with one extremely high outlier (Washington DC, 54.6%) and two other potential outliers (Massachusetts, 40.5% and West Virginia, 19.2%). The "typical" state percentage appears to be between 25 and 30%.
   b. No: Since the population sizes of the 50 states + DC are not equal, the mean of these percentages would not equal the overall percentage. (If we knew all 51 population sizes, we could take the appropriate weighted average, effectively re-constructing the total count of people with 4-year degrees and dividing by the total population size.)

29. b. The transformation substantially changes the shape of the histogram. In particular, while the original variable x = number of defects was strongly positively skewed with an outlier, $\log_{10}(x)$ is reasonably symmetrically distributed with no outlier.

31. a. 7% of 464 students is roughly (.07)(464) = 32.48, or 32 students. [32/464 = .069, which rounds to .07.]
   b. 18% + 6% + 5% = 29%.
   c. No. Without an upper bound on the last category, we can't even make a density histogram of the data, because we don't know where the last rectangle should end.

33. a. The distribution is skewed to the right, or positively skewed. There is a gap in the histogram, and what appears to be an outlier in the 500 −< 550 interval.

Class interval	Frequency	Relative frequency
0 −< 50	9	0.18
50 −< 100	19	0.38
100 −< 150	11	0.22
150 −< 200	4	0.08
200 −< 250	2	0.04
250 −< 300	2	0.04
300 −< 350	1	0.02
350 −< 400	1	0.02
400 −< 450	0	0.00
450 −< 500	0	0.00
500 −< 550	1	0.02
	50	1.00

   b. The distribution of the natural logs of the original data is much more symmetric than the original.

Class interval	Frequency	Relative frequency
2.25 −< 2.75	2	0.04
2.75 −< 3.25	2	0.04
3.25 −< 3.75	3	0.06
3.75 −< 4.25	8	0.16
4.25 −< 4.75	18	0.36
4.75 −< 5.25	10	0.20
5.25 −< 5.75	4	0.08
5.75 < 6.25	3	0.06

   c. The proportion of lifetime observations in this sample that are less than 100 is .18 + .38 = .56, and the proportion that are at least 200 is .04 + .04 + .02 + .02 + .02 = .14.

35. a. The variable here is *helmet status*, a categorical variable. Its possible values are *no helmet, noncompliant helmet,* and *compliant helmet*.
   b.

Category	Frequency	Relative frequency
No helmet	731	.43
Noncompliant helmet	153	.09
Compliant helmet	816	.48
Total	1700	1.00

   c. .09 + .48 = .57.

39. a. The relative frequency distribution is as follows. The relative frequency distribution is almost unimodal and exhibits a large positive skew. The typical middle value is somewhere between 400 and 450, although the skewness makes it difficult to pinpoint more exactly than this.

Class	Rel. Freq.	Class	Rel. Freq.
0 −< 150	.193	1050 −< 1200	.029
150 −< 300	.183	1200 −< 1350	.005
300 −< 450	.251	1350 −< 1500	.004
450 −< 600	.148	1500 −< 1650	.001
600 −< 750	.097	1650 −< 1800	.002
750 −< 900	.066	1800 −< 1950	.002
900 −< 1050	.019		

b. The proportion of the fire loads less than 600 is .193 + .183 + .251 + .148 = .775 (the cumulative proportion for 600). The proportion of loads that are at least 1200 is .005 + .004 + .001 + .002 + .002 = .014 (the opposite of the cumulative proportion for 1200).

c. The proportion of loads between 600 and 1200 is 1 − .775 − .014 = .211.

41. a. $\bar{x} = (5 + 2 + \cdots + 5 + 0)/10 = 3.5$ yd.

b. The two middle values in order are 2 and 2, so $\tilde{x} = 2$ yards. Todd Gurley's mean rushing gain is artificially increased by the one 16-yard gain, while the median ignores this extreme value.

c. Deleting the 16-yard gain and the 1-yard loss (−1) amounts to trimming 1/10 observations from each end. So, we're talking about the 10% trimmed mean, and the average of the remaining 8 values is $\bar{x}_{tr(10)} = 2.5$ yards. As is typically the case, the trimmed mean falls between the median (2 yards) and the mean (3.5 yards).

43. a. With the one very high outlier (*Wall Street Journal* at over 2.2 million), we anticipate that the mean will be higher than the median.

b. $\bar{x} = \frac{1}{20}(2237601 + \cdots + 196286) = 403,456$. In order, the middle two values are 285,129 and 276,445, so $\tilde{x} = \frac{1}{2}(285129 + 276445) = 280,787$. Sure enough, the median circulation for the top 20 newspapers is substantially less than the mean, due to the one extremely high outlier.

45. Using software, $\tilde{x} = 92$, $\bar{x}_{tr(25)} = 95.07$, $\bar{x}_{tr(10)} = 102.23$, $\bar{x} = 119.3$. The mean is somewhat larger because of positive skewness. Trimming results in a value between the mean and median, and additional trimming gives a value closer to the median.

47. a. The *reported* values are (in increasing order) 110, 115, 120, 120, 125, 130, 130, 135, and 140. Thus the median of the reported values is 125.

b. 127.6 is reported as 130, so the median is now 130, a very substantial change. When there is rounding or grouping, the median can be highly sensitive to small change.

49. The mean cannot be calculated, because we need the exact value of the two 100+ observations. We can, however, compute median = $(57 + 79)/2 = 68.0$, 20% trimmed mean = 66.2, 30% trimmed mean = 67.5.

51. a. Manufacturer is a categorical variable.

b. Since Honda is the most frequent manufacturer, arguably Honda is the most representative "value" of this categorical variable.

c. No. Any numerical coding of these six categories artificially imposes an order on the manufacturers. For instance, sorting alphabetically and sorting by popularity would result in different codings and thus different means and medians. Only the *mode* (i.e., part **b**) makes sense as a representative value.

53. a. range = 49.3 − 23.5 = 25.8.

   b. $\Sigma x_i = 310.3$, $\bar{x} = 31.03$, $S_{xx} = \Sigma(x_i - \bar{x})^2$ $= 443.801$, $\Sigma x_i^2 = 10{,}072.41$, $s^2 = \frac{\Sigma(x_i-\bar{x})^2}{n-1}$ $= \frac{443.801}{9} = 49.3112$.

   c. $s = \sqrt{49.3112} = 7.022$.

   d. $s^2 = \frac{\Sigma x^2 - (\Sigma x)^2/n}{n-1} = \frac{10{,}072.41 - (310.3)^2/10}{9}$ $= 49.3112$.

55. a. $\bar{x} = \Sigma x_i/n = 14438/5 = 2887.6$. The sorted values are: 2781 2856 2888 2900 3013, so the sample median is $\tilde{x} = 2888$.

   b. Subtracting a constant from each observation shifts the data, but does *not* change its sample variance. For example, by subtracting 2700 from each observation we get the values 81, 200, 313, 156, and 188, which are smaller (fewer digits) and easier to work with by hand. The sum of squares of this transformed data is 204210 and its sum is 938, so the computational formula for the variance gives $s^2 = [204210 - (938)^2/5]/(5 - 1) = 7060.3$.

57. $s = 24.4$. In general, the size of a typical deviation from the sample mean (370.7 s) is about 24.4 s. Some observations may deviate from 370.7 by a little more than this, some by less.

59. $1,961,160

61. −3.5. One sample for which these are the deviations is 3.8, 4.4, 4.5, 4.8, and 0.

63. a. $q_1 = 149.5$, $q_3 = 1175$, iqr = 1175 − 149.5 = 1025.5

   b. A high outlier is anything exceeding $q_3 +$ 1.5iqr = 1175 + 1.5(1025.5) = 2713.25, and an extremely high outlier is anything over $q_3 + $ 3iqr = 1175 + 3(1025.5) = 4251.5.

   c. A boxplot shows a positively skewed award distribution, with a median award of $750 thousand and no apparently outliers.

65. a. 27.82, 26, 27.38

   b. From software, the quartiles are roughly 23 and 32, so iqr = 9. Mild outliers are outside 23 − 1.5(9) = 9.5 and 32 + 1.5(9) = 45.5. Extreme outliers are outside 23 − 3(9) = −4 and 32 + 3(9) = 59. Hence, there is one low mild outlier and there are three high mild outliers. *Note*: Depending on how the quartiles and iqr are calculated, the observation 46 might or might not be deemed an outlier.

67. The most noticeable feature of the comparative boxplots is that machine 2's sample values have considerably more variation than does machine 1's sample values. However, a typical value, as measured by the median, seems to be about the same for the two machines. The only outlier that exists is from machine 1.

69. All of the Indian salaries are below the first quartile of Yankee salaries. There is much more variability in the Yankee salaries. Neither team has any outliers.

71. Outliers occur in the 6 a.m. data. The distributions at the other times are fairly symmetric. Variability and the typical values in the data increase a little at the 12 noon and 2 p.m. times. Clearly the 6 a.m. vehicles warrant further investigation!

73. a.

Males			Females	
	2	6		Stem: ones digit
	2			Leaf: tenths digit
1	3	0011		
	3	22		
5444	3	5		
776	3			
988	3	8		
00	4			
	4	3		

   b. $\tilde{x} = 3.70$ cm for males and 3.15 cm for females.

   c. Males' aortic root diameters are greater, on average, than females' in this sample (see the medians above). But the women in the sample exhibited much more

variability in aortic root diameter than did the men, including some potential high and low outliers.

75. There are no outliers in the three data sets. However, as a comparative boxplot shows, the three data sets differ with respect to their central values (the medians are different) and the data for flow rate 160 is somewhat less variable than the other data sets. Flow rates 125 and 200 also exhibit a small degree of positive skewness.

77. a. HC data: $s = 9.59$. CO data: $s = 59.41$. Since the CO data are on a much larger scale, it makes sense that their standard deviation should be larger—standard deviation reflects *absolute* scale.

  b. The mean of the HC data is $96.8/4 = 24.2$; the mean of the CO data is $735/4 = 183.75$. Therefore, the coefficient of variation of the HC data is $9.59/24.2 = .3963$, or 39.63%. The coefficient of variation of the CO data is $59.41/183.75 = .3233$, or 32.33%. Thus, even though the CO data has a larger standard deviation than does the HC data, it actually exhibits *less* variability (in percentage terms) around its average than does the HC data.

79. 10.70; 10.60; 10.65

81. The IQ distribution for these 33 children is reasonably symmetric, with a mean IQ score of 113.7 and a standard deviation of 12.7. The sample includes three outliers (using the 1.5iqr rule): a low outlier at 82 and two high outliers at 140 and 146.

83. a. The typical radon level in houses where a child had cancer seems somewhat higher than in no-cancer households. Both distributions are positively skewed. Radon levels of 55, 55, and 85 Bq/m^3 are potential high outliers among the no-cancer households, while an extreme outlier of 210 Bq/m^3 was recorded in one household with a childhood cancer.

Cancer		No cancer
9987653	0	33566777889999
88876666555533211111000	1	11111223477
73322110	2	11449999
9843	3	389
5	4	
7	5	55
	6	
	7	
HI : 210	8	5

Stem : Tens digit Leaf : Ones digit

  b. $s = 31.7$ Bq/m^3 for the cancer households and 17.0 Bq/m^3 for the no-cancer households, suggesting greater variability in the first group. This seemingly contradicts the graph, where the radon distribution on the left appears more concentrated than the one on the right.

  c. iqr $= 11.0$ for cancer households and 18.0 for non-cancer households. Now the non-cancer households exhibit greater variability in radon levels, which is more consistent with our graph. The culprit here is presumably the extreme value of 210, which greatly influences the standard deviation of the cancer group but has no effect on the iqr of that sample.

85. The healthy individuals have higher receptor binding measure on average than the individuals with PTSD. There is also more variation in the healthy individuals' values. The distribution of values for the healthy is reasonably symmetric, while the distribution for the PTSD individuals is negatively skewed.

87. a. Mode $= .93$. It occurs four times in the data set.

  b. The *modal category* is the one with the highest (relative) frequency.

89. The measures that are sensitive to outliers are: the mean and the midrange. The mean is sensitive because all values are used in computing it. The midrange is sensitive because it uses only the most extreme values in its computation. The median, the trimmed mean, and the midquarter are not sensitive to outliers. The median is the most resistant to outliers because it uses only the middle value (or values) in its computation. The trimmed mean is somewhat resistant to

outliers. The larger the trimming percentage, the more resistant the trimmed mean becomes. The midquarter, which uses the quartiles, is reasonably resistant to outliers because both quartiles are resistant to outliers.

91. a. $s_y^2 = s_x^2$ and $s_y = s_x$ b. $s_z^2 = 1$ and $s_z = 1$

93. b. .552, .102 c. 30 d. 19

95. a. There may be a tendency to a repeating pattern.
    b. The value .1 gives a much smoother series.
    c. The smoothed value depends on all previous values of the time series, but the coefficient decreases with k.
    d. As t gets large, the coefficient $(1 - \alpha)^{t-1}$ decreases to zero, so there is decreasing sensitivity to the initial value.

Chapter 2

1. a. $A \cap B'$
    b. $A \cup B$
    c. $(A \cap B') \cup (B \cap A')$

3. a. $\mathscr{S}$ = {1324, 1342, 1423, 1432, 2314, 2341, 2413, 2431, 3124, 3142, 4123, 4132, 3214, 3241, 4213, 4231}
    b. A = {1324, 1342, 1423, 1432}
    c. B = {2314, 2341, 2413, 2431, 3214, 3241, 4213, 4231}
    d. $A \cup B$ = {1324, 1342, 1423, 1432, 2314, 2341, 2413, 2431, 3214, 3241, 4213, 4231}
       $A \cap B = \varnothing$
       A' = {2314, 2341, 2413, 2431, 3124, 3142, 4123, 4132, 3214, 3241, 4213, 4231}

5. a. A = {*SSF, SFS, FSS*}
    b. B = {*SSS, SSF, SFS, FSS*}
    c. C = {*SSS, SSF, SFS*}
    d. C' = {*SFF, FSS, FSF, FFS, FFF*}
       $A \cup C$ = {*SSS, SSF, SFS, FSS*}
       $A \cap C$ = {*SSF, SFS*}
       $B \cup C$ = {*SSS, SSF, SFS, FSS*}
       $B \cap C$ = {*SSS SSF, SFS*}

7. a. {111, 112, 113, 121, 122, 123, 131, 132, 133, 211, 212, 213, 221, 222, 223, 231, 232, 233, 311, 312, 313, 321, 322, 323, 331, 332, 333}
    b. {111, 222, 333}
    c. {123, 132, 213, 231, 312, 321}
    d. {111, 113, 131, 133, 311, 313, 331, 333}

9. a. {*BBBAAAA,   BBABAAA,   BBAABAA, BBAAABA,   BBAAAAB,   BABBAAA, BABABAA,   BABAABA,   BABAAAB, BAABBAA,   BAABABA,   BAABAAB, BAAABBA,   BAAABAB,   BAAAABB, ABBBAAA,   ABBABAA,   ABBAABA, ABBAAAB,   ABABBAA,   ABABABA, ABABAAB,   ABAABBA,   ABAABAB, ABAAABB,   AABBBAA,   AABBABA, AABBAAB,   AABABBA,   AABABAB, AABAABB,   AAABBBA,   AAABBAB, AAABABB,   AAAABBB*}
    b. {*AAAABBB,   AAABABB,   AAABBAB, AABAABB,   AABABAB*}

13. a. .07
    b. .30
    c. .57

15. a. They are awarded at least one of the first two projects, .36.
    b. They are awarded neither of the first two projects, .64.
    c. They are awarded at least one of the projects, .53.
    d. They are awarded none of the projects, .47.
    e. They are awarded only the third project, .17.
    f. Either they fail to get the first two or they are awarded the third, .75.

17. a. .572
    b. .879

19. a. SAS and SPSS are not the only packages.
    b. .7
    c. .8
    d. .2

21. a. .8841
    b. .0435

23. a. .10
    b. .18, .19
    c. .41
    d. .59
    e. .31
    f. .69

25. a. 1/15
    b. 6/15
    c. 14/15
    d. 8/15

27. a. .98
    b. .02
    c. .03
    d. .24

29. a. 1/9
    b. 8/9
    c. 2/9

31. a. 20
    b. 60
    c. 10

33. a. 243
    b. 3645, 10

35. .0679

37. a. 8008   b. 3300
    c. 5236   d. .4121, .6538

39. .20

41. .0456

43. a. .0839   b. .2498   c. .1998

45. 1/15, 1/3, 2/3

49. a. .447, .5, .2
    b. $P(A|C) = .4$, the fraction of ethnic group
       C that has blood type A.
       $P(C|A) = .447$, the fraction of those with
       blood group A that are of ethnic group C.
    c. .211

51. a. Of those with a Visa card, .5 is the
       fraction who also have a Master Card.

    b. Of those with a Visa card, .5 is the
       fraction who do not have a Master Card.
    c. Of those with Master Card, .625 is the
       fraction who also have a Visa Card.
    d. Of those with Master Card, .375 is the
       fraction who do not have a Visa Card.
    e. Of those with at least one of the two
       cards, .769 is the fraction who have a
       Visa card.

53. .217, .178

55. .436, .581

57. .0833

59. a. .102   b. 1

65. a. .067   b. .509

69. a. .765   b. .2353

71. .466, .288, .247

73. a. *BB* or *Bb*, with probability 1/2 each
    b. 4/7 c. 2/3

75. a. Because of independence, the condi-
       tional probability is the same as the
       unconditional probability, .3.
    b. .82
    c. .146

79. .349, .651, $(1 - p)^n$, $1 - (1 - p)^n$

81. .99999969, .2262

83. .9981

85. a. yes   b. no

87. a. $2p - p^2$
    b. $1 - (1 - p)^n$
    c. $(1 - p)^3$
    d. $.9 + .1(1 - p)^3$
    e. .0137

89. .8588, .9897

91. $2p/(1 + p)$

93. a. exact answer = .46   b. se $\approx$ .005

95. .8159 (answers will vary)

97. $\approx .39$, $\approx .88$ (answers will vary)

99. $\approx .91$ (answers will vary)

101. $\approx .02$ (answers will vary)

103. b. $\approx .37$ (answers will vary)
     c. $\approx 176{,}000{,}000$ (answers will vary; exact $= 176{,}214{,}841$)

105. a. $\approx .20$ b. $\approx .56$ (answers will vary)

107. a. $\approx .5177$ b. $\approx .4914$ (answers will vary)

109. $\approx .2$ (answers will vary)

111. b. $\pi \approx 4 \cdot \hat{P}(A)$

113. a. 10,626   b. 255,024   c. 127,512

115. a. 1/3, .444
     b. .15
     c. .291

117. .45, .32

119. a. 1/120
     b. 1/5
     c. 1/5

121. .905

123. a. .904   b. .766

125. .008

127. .362, .348, .290

129. a. $P(G|R_1 < R_2 < R_3) = 2/3$, so classify as granite if $R_1 < R_2 < R_3$.
     b. $P(G|R_1 < R_3 < R_2) = .294$, so classify as basalt if $R_1 < R_3 < R_2$.
        $P(G|R_3 < R_1 < R_2) = 1/15$, so classify as basalt if $R_3 < R_1 < R_2$.
     c. .175
     d. $p > 14/17$

131. a. $1/24$ b. $15/24$ c. $1 - e^{-1}$

133. $s = 1$

137. a. $P(B_0|\text{survive}) = b_0/[1 - (b_1 + b_2)cd]$
        $P(B_1|\text{survive}) = b_1(1 - cd)/[1 - (b_1 + b_2)cd]$
        $P(B_2|\text{survive})$
        $= b_2(1 - cd)/[1 - (b_1 + b_2)cd]$
     b. .712, .058, .231

Chapter 3

1.

$\mathcal{S}$:	FFF	SFF	FSF	FFS	FSS	SFS	SSF	SSS
X:	0	1	1	1	2	2	2	3

3. M = the absolute value of the difference between the outcomes with possible values 0, 1, 2, 3, 4, 5 or 6; $W = 1$ if the sum of the two resulting numbers is even and $W = 0$ otherwise, a Bernoulli random variable.

5. No, X can be a Bernoulli random variable where a success is an outcome in B, with B a particular subset of the sample space.

7. a. Possible values are 0, 1, 2, …, 12; discrete
   b. With N = # on the list, values are 0, 1, 2, …, N; discrete
   c. Possible values are 1, 2, 3, 4, …; discrete
   d. $\{x: 0 < x < \infty\}$ if we assume that a rattlesnake can be arbitrarily short or long; not discrete
   e. With c = amount earned per book sold, possible values are 0, c, $2c$, $3c$, …, $10{,}000c$; discrete
   f. $\{y: 0 < y < 14\}$ since 0 is the smallest possible pH and 14 is the largest possible pH; not discrete
   g. With m and M denoting the minimum and maximum possible tensions, respectively, possible values are $\{x: m < x < M\}$; not discrete
   h. Possible values are 3, 6, 9, 12, 15, … i.e. 3(1), 3(2), 3(3), 3(4), …giving a first element, etc.; discrete

9. a. X is a discrete random variable with possible values $\{2, 4, 6, 8, …\}$
   b. X is a discrete random variable with possible values $\{2, 3, 4, 5, …\}$

11. a. .10
    c. .45, .25

13. a. .70
    b. .45
    c. .55
    d. .71
    e. .65
    f. .45

15. a. (1, 2) (1, 3) (1, 4) (1, 5) (2, 3) (2, 4) (2, 5) (3, 4) (3, 5) (4, 5)

    b. $p(0) = .3, p(1) = .6, p(2) = .1, p(x) = 0$ otherwise

    c. $F(0) = .30, F(1) = .90, F(2) = 1$. The cdf is

$$F(x) = \begin{cases} 0 & x < 0 \\ .30 & 0 \le x < 1 \\ .90 & 1 \le x < 2 \\ 1 & 2 \le x \end{cases}$$

17. a. .81

    b. .162

    c. The fifth battery must be an A, and one of the first four must also be an A, so $p(5) = P(AUUUA$ or $UAUUA$ or $UUAUA$ or $UUUAA) = .00324$

    d. $P(Y = y) = (y - 1)(.1)^{y-2}(.9)^2, y = 2, 3, 4, 5,\dots$

19. b. $p(1) = .301, p(2) = .176, p(3) = .125, p(4) = .097, p(5) = .079, p(6) = .067, p(7) = .058, p(8) = .051, p(9) = .046$. Lower digits (such as 1 and 2) are much more likely to be the lead digit of a number than higher digits (such as 8 and 9).

    c. $F(1) = .301, F(2) = .477, F(3) = .602, F(4) = .699, F(5) = .778, F(6) = .845, F(7) = .903, F(8) = .954, F(9) = 1$. So, $F(x) = 0$ for $x < 1$; $F(x) = .301$ for $1 \le x < 2$; $F(x) = .477$ for $2 \le x < 3$; etc.

    d. .602, .301

21. $F(x) = 0, x < 0; .10, 0 \le x < 1; .25, 1 \le x < 2; .45, 2 \le x < 3; .70, 3 \le x < 4; .90, 4 \le x < 5; .96, 5 \le x < 6; 1.00, 6 \le x$

23. a. $p(1) = .30, p(3) = .10, p(4) = .05, p(6) = .15, p(12) = .40$

    b. .30, .60

25. a. $p(x) = (1/3)(2/3)^{x-1}, x = 1, 2, 3, \dots$

    b. $p(y) = (1/3)(2/3)^{y-2}, y = 2, 3, 4, \dots$

    c. $p(0) = 1/6, p(z) = (25/54)(4/9)^{z-1}, z = 1, 2, 3, 4, \dots$

29. a. .60

    b. \$110

31. a. 16.38, 272.298, 3.9936

    b. \$458.46

    c. \$33.97

    d. 13.66

33. Yes, because $\Sigma(1/x^2)$ is finite.

35. \$700

37. Since \$142.92 > \$100, you expect to win more if you gamble.

39. a. –\$1/19, –\$1/19

    b. The expected return for a \$1 wager on roulette is the same no matter how you bet.

    c. \$5.76, \$2.76, \$1.00

    d. Low-risk/low-reward bets (such as a color) have smaller standard deviation than high-risk/high-reward bets (such as a single number).

43. a. 32.5

    b. 7.5

    c. $V(X) = E[X(X - 1)] + E(X) - [E(X)]^2$

45. a. 1/4, 1/9, 1/16, 1/25, 1/100

    b. $\mu = 2.64, \sigma = 1.54, P(|X - \mu| \ge 2\sigma) = .04 < .25, P(|X - \mu| \ge 3\sigma) = 0 < 1/9$ The actual probability can be far below the Chebyshev bound, so the bound is conservative.

    c. 1/9, equal to the Chebyshev bound

    d. $p(-1) = .02, p(0) = .96, p(1) = .02$

47. $M_X(t) = .5e^t/(1 - .5e^t), E(X) = 2, V(X) = 2$

49. a. $.01e^{9t} + .05e^{10t} + .16e^{11t} + .78e^{12t}$

    b. 11.71, .3659

51. $p(0) = .2, p(1) = .3, p(3) = .5, E(X) = 1.8, V(X) = 1.56$

55. a. 5, 4   b. 5, 4

57. $p(y) = (.25)^{y-1}(.75)$ for $y = 1, 2, 3, \dots$

59. $M_Y(t) = e^{t^2/2}, E(X) = 0, V(X) = 1$

63. a. .124 b. .279 c. .635 d. .718

65. a. .873

    b. .007

    c. .716

    d. .277

    e. 1.25, 1.09

67. a. .786   b. .169   c. .382

69. a. .403   b. .787   c. .774

71. .1478

73. .407, assuming batteries' voltage levels are independent

75. a. .0104 c. .00197 d. 1500, 260

77. a. .017   b. .811, .425   c. .006, .902, .586

79. For $p = .9$ the probability is higher for B (.9963 versus .99 for A)
For $p = .5$ the probability is higher for A (.75 versus .6875 for B)

81. a. 20, 16 (binomial, $n = 100$, $p = .2$)
b. 70, 21

83. a. $p = 0$ or 1   b. $p = .5$

85. When $p = .5$, the true probability for $k = 2$ is .0414, compared to the bound of .25.
When $p = .5$, the true probability for $k = 3$ is .0026, compared to the bound of .1111.
When $p = .75$, the true probability for $k = 2$ is .0652, compared to the bound of .25.
When $p = .75$, the true probability for $k = 3$ is .0039, compared to the bound of .1111.

89. a. .932   b. .065   c. .068   d. .491   e. .251

91. a. .011   b. .441   c. .554, .459   d. .944

93. a. .219   b. .558

95. .857

97. a. .122, .808, .283 b. 12, 3.464 c. .530, .011

99. a. .099   b. .135   c. 2

101. a. 4   b. .215   c. 1.15 years

103. a. .221   b. 6, 800, 000   c. $p(x; 1608.5)$

109. a. .114   b. .879   c. .121   d. use Bin(15, .1)

111. a. $h(x; 15, 10, 20)$ b. .0325   c. .6966

113. a. $h(x; 10, 10, 20)$ b. .0325 c. $h(x; n, n, 2n)$, $E(X) = n/2$, $V(X) = n^2/[4(2n-1)]$

115. a. $nb(x; 2, .5) = (x+1).5^{x+2}, x = 0, 1, 2, 3, \ldots$
b. 3/16
c. 11/16
d. 4, 2

117. $2 + 2 + 2 = 6$

119. a. .2817   b. .7513   c. .4912, .9123

121. a. 160, 21.9   b. .6756

125. mean ≈ 0.5968, sd ≈ 0.8548 (answers will vary)

127. $\approx .9090$ (answers will vary)

129. a. mean ≈ 13.5888, sd ≈ 2.9381
b. $\approx .1562$ (answers will vary)

131. mean ≈ 3.4152, variance ≈ 5.97 (answers will vary)

133. b. 142 tickets

135. a. $\approx .2291$ b. $\approx \$8696$ c. $\approx \$7811$
d. $\approx .2342, \$7767, \$7571$ (answers will vary)

137. b. $\approx .9196$ (answers will vary)

139. b. 3.114, .405, .636

141. a. $b(x; 15, .75)$
b. .6865   c. .313
d. 45/4, 45/16 e. .309

143. a. .013 b. 19 c. .266 d. Poisson(500)

145. a. $p(x; 2.5)$ b. .067   c. .109

147. 1.813, 3.05

149. $p(2) = p^2, p(3) = (1 - p)p^2, p(4) = (1 - p)p^2, p(x) = [1 - p(2) - \cdots - p(x - 3)](1 - p)p^2, x = 5, 6, 7, \ldots$.
Alternatively, $p(x) = (1 - p)p(x - 1) + p(1 - p)p(x - 2), x = 5, 6, 7, \ldots .99950841$

151. a. 0029   b. 0767, .9702

153. a. .135   b. .00144   c. $\sum\limits_{x=0}^{\infty} [p(x; 2)]^5$

155. 3.590

157. a. No   b. .0273

159. b. $.5\mu_1 + .5\mu_2$
c. $.5\mu_1 + .5\mu_2 + .25(\mu_1 - \mu_2)^2$
d. $p(x; \mu_1, \mu_2) = .6 \, p(x; \mu_1) + .4 \, p(x; \mu_2)$

161. .5

165. $X \sim b(x; 25, p)$, $E(h(X)) = 500p + 750$, $\sigma_{h(X)} = 100\sqrt{p(1-p)}$. Independence and constant probability might not be valid because of the effect that customers can have on each other. Also, store employees might affect customer decisions.

167. $p(0) = .07776$, $p(1) = .10368$, $p(2) = .19008$, $p(3) = .20736$, $p(4) = .17280$, $p(5) = .13824$, $p(6) = .06912$, $p(7) = .03072$, $p(8) = .01024$

Chapter 4

1. a. .25 b. .5 c. 7/16

3. b. .5 c. 11/16 d. .6328

5. a. 3/8 b. 1/8 c. .2969 d. .5781

7. a. $f(x) = .10$ for $25 \le x \le 35$ and $= 0$ otherwise
b. .2 c. .4 d. .2

9. a. .699 b. .301, .301 c. .166

11. a. 1/4 b. 3/16 c. 15/16 d. $\sqrt{2}$ e. $f(x) = x/2$ for $0 \le x < 2$, and $f(x) = 0$ otherwise

13. a. 3 b. 0 for $x \le 1$, $1 - 1/x^3$ for $x > 1$
c. 1/8, .088

15. a. $F(x) = 0$ for $x \le 0$, $F(x) = x^3/8$ for $0 < x < 2$, $F(x) = 1$ for $x \ge 2$
b. 1/64   c. .0137, .0137   d. 1.817

17. b. 90th percentile of $Y = 1.8$(90th percentile of X) + 32   c. 100pth percentile of $Y = a$(100pth percentile of X) + b for $a > 0$.

19. a. 35, 25 b. .865

21. a. .8182, .1113 b. .314

23. a. $A + (B - A)p$
b. $(A + B)/2$
c. $(B^{n+1} - A^{n+1})/[(n+1)(B - A)]$

25. 314.79

27. 248, 3.6

29. 1/4, 1/16

31. a. $v/20$, $v/800$ b. 100.2π versus 100π c. $80\pi^2$

33. $M_Y(t) = (e^{5t} - e^{-5t})/10t$, $Y \sim \text{Unif}[-5, 5]$

35. a. $M_Y(t) = .04e^{10t}/(.04 - t)$ for $t < .04$, mean = 35, variance = 625
b. $M(t) = .04/(.04 - t)$ for $t < .04$, mean = 25, variance = 625
c. $M_Y(t) = .04/(.04 - t)$; Y is a shifted exponential rv

39. a. .4850 b. .3413 c. .4938 d. .9876 e. .9147 f. .9599 g. .9104 h. .0791 i. .0668 j. .9876

41. a. 1.34 b. −1.34 c. .674 d. −.674 e. −1.555

43. a. .9772 b. .5 c. .9104 d. .8413 e. .2417 f. .6826

45. a. .7977 b. .0004 c. The top 5% are the values above .3987.

47. The second machine

49. a. .2514, ~ 0 b. 39.985 ksi

51. .0510

53. a. .8664 b. .0124 c. .2718

55. a. .794 b. 5.88 c. 7.94 d. .265

57. a. $\Phi(1.72) - \Phi(.55)$ b. $\Phi(.55) - [1 - \Phi(1.72)]$; No, due to symmetry.

59. a. .4584 b. 135.8 kph c. .9265 d. .3173 e. .6844

61. a. .7286 b. .8643, .8159

63. a. .9932 b. .9875 c. .8064

65. a. .0392 b. ~ 1

69. a. .15872   actual   .15866
b. .0013495   actual   .0013499
c. .999936655   actual   .999936658
d. .00000028669   actual   .00000028665

71. a. 120 b. 1.329 c. .371 d. .735 e. 0

73. a. 5, 4 b. .715 c. .411

75. a. 1 b. 1 c. .982 d. .129

77. a. .449, .699, .148 b. .050, .018

79. a. $\cap A_i$ b. Exponential with $\lambda = .05$
    c. Exponential with parameter $n\lambda$

81. a. Gamma, $\alpha = 3$, $\beta = 1/\lambda$ b. .8165

85. a. .275, .599, .126   b. 20.418, 17.365
    c. 15.84

89. a. .9295   b. .2974   c. 98.184

91. a. 7.53, 9.966   b. .7823, .1469
    c. .6925; lognormal is not symmetric

93. a. 149.157, 223.595   b. .957   c. .0416
    d. 148.41   e. 9.57   f. 125.90

95. $\alpha = \beta$

97. b. $\Gamma(\alpha + \beta)\Gamma(m + \beta)/[\Gamma(\alpha + \beta + m)\Gamma(\beta)]$,
    $\beta/(\alpha + \beta)$

99. Yes, since the pattern in the plot is quite
    linear.

101. Yes

103. Yes, because the plot is reasonably straight

105. Form a new variable, the logarithms of a
     TN value, and then construct a normal plot
     for its values. Because of the linearity of
     this plot, normality is plausible.

107. The pattern in the normal probability plot
     is curved downward, consistent with a
     right-skewed distribution. It is not plausi-
     ble that shower flow rate has a normal
     population distribution.

109. The plot deviates from linearity, especially
     at the low end, where the smallest three
     observations are too small relative to the
     others. The plot works for any λ because λ
     is a scale parameter.

111. $f_Y(y) = 2/y^3$, $y > 1$

113. $f_Y(y) = ye^{-y^2/2}$, $y > 0$

115. $f_Y(y) = 1/16$, $0 < y < 16$

117. $f_Y(y) = 1/[\pi(1 + y^2)]$

119. $Y = X^2/16$

121. $f_Y(y) = \frac{2}{\sqrt{2\pi}}e^{-y^2/2}$ for $y > 0$

125. a. $F(x) = x^2/4$, $x = 2\sqrt{u}$
     c. sample mean and sd = 1.331 and 0.471
     (answers will vary), $\mu = 4/3$ and
     $\sigma = \sqrt{2}/3$

129. b. sample mean = 15.9188, close to 16
     (answers will vary)

131. $1480

133. b. $F(x) = 1 - 16/(x + 4)^2$, $x \geq 0$; $F(x) = 0$,
     $x < 0$ c. .247   d. 4   e. 16.67

135. a. .6563 b. 41.55 c. .3197

137. a. .0003 b. .0888

139. a. $F(x) = 1.5(1 - 1/x)$, $1 \leq x \leq 3$; $F(x) = 0$,
     $x < 1$; $F(x) = 1$, $x > 3$
     b. .9, .4   c. 1.6479 d. .5333 e. .2662

141. a. 1.075, 1.075 b. .0614, .3331 c. 2.476

143. b. $95,600, .3300

145. b. $F(x) = .5e^{2x}$, $x \leq 0$; $F(x) = 1 - .5e^{-.2x}$,
     $x > 0$ c. .5, .6648, .2555, .6703

147. a. $k = (\alpha - 1)5^{\alpha-1}$ for $\alpha > 1$ b. $F(x) = 0$,
     $x \leq 5$; $F(x) = 1 - (5/x)^{\alpha-1}$, $x > 5$
     c. $5(\alpha - 1)/(\alpha - 2)$

149. b. .4602, .3636 c. .5950 d. 140.178

151. a. Weibull b. .542

153. a. λ b. $\alpha x^{\alpha-1}/\beta^\alpha$
     c. $F(x) = 1 - e^{-\alpha(x-x^2/(2\beta))}$, $0 \leq x \leq \beta$;
     $F(x) = 0$, $x < 0$; $F(x) = 1 - e^{-\alpha\beta/2}$,
     $x > \beta$, $f(x) = \alpha(1 - x/\beta) e^{-\alpha(x-x^2/(2\beta))}$,
     $0 \leq x \leq \beta$; $f(x) = 0$, $x < 0$, $f(x) = 0$, $x > \beta$
     This gives total probability less than 1, so
     some probability is located at infinity (for
     items that last forever).

157. $F(q^*) = .818$

Chapter 5

1. a. .20 b. .42 c. The probability of at least one hose being in use at each pump is .70.

   d.
x	0	1	2
$p_X(x)$	.16	.34	.50

y	0	1	2
$p_Y(y)$	.24	.38	.38

   $P(X \leq 1) = .50$

   e. dependent, $.30 = P(X = 2 \text{ and } Y = 2) \neq P(X = 2)\,P(Y = 2) = (.50)(.38)$

3. a. .15 b. .40 c. $.22 = P(A) = P(|X_1 - X_2| \geq 2)$
   d. .17, .46

5. a. .0305 b. .1829
   c. probability = .1073, marginal evidence

7. a. .054 b. .00018

9. a. .030 b. .120 c. .10, .30 d. .38 e. yes, $p(x,y) = p_X(x) \cdot p_Y(y)$

11. a. 3/380,000 b. .3024 c. .3593
    d. $10kx^2 + .05, \quad 20 \leq x \leq 30$ e. no

13. a. $p(x, y) = e^{-\mu_1 - \mu_2}\mu_1^x\mu_2^y/x!y!$
    b. $e^{-\mu_1 - \mu_2}[1 + \mu_1 + \mu_2]$
    c. $\frac{e^{-\mu_1-\mu_2}}{m!}(\mu_1 + \mu_2)^m$, Poisson with parameter $\mu_1 + \mu_2$

15. a. $e^{-x-y}, x \geq 0, y \geq 0$ b. .3996 c. .5940
    d. .3298

17. a. $F(y) = 1 - 2e^{-2\lambda y} + e^{-3\lambda y}$ for $y \geq 0$, $F(y) = 0$ for $y < 0$; $f(y) = 4\lambda e^{-2\lambda y} - 3\lambda e^{-3\lambda y}$ for $y \geq 0$
    b. $2/(3\lambda)$

19. a. .25 b. $1/\pi$ c. $2/\pi$ d. $f_X(x) = 2\sqrt{r^2 - x^2}/(\pi r^2)$ for $-r \leq x \leq r, f_Y(y) = 2\sqrt{r^2 - y^2}/(\pi r^2)$ for $-r \leq y \leq r$, no

21. 1/3

23. a. .11 b.
x	0	1	2	3
$p_X(x)$	.78	.12	.07	.03

y	0	1	2
$p_Y(y)$	.77	.14	.09

c. no d. 0.35, 0.32 e. 95.72

25. .15

27. L^2

29. 1/4 h

31. –2/3

33. –.1082, –.0131

35. .238, .51

37. $V(h(X, Y)) = E(h^2(X, Y)) - [E(h(X, Y))]^2$, 13.34

41. $\rho = 1$ when $a > 0$

43. a. 87,850, 4370.37 b. yes, no c. .0027

45. .0336, .2310

47. .0314

49. a. 45 min b. 68.33 c. –1, 13.67 d. –5, 68.33

51. a. 50, 10.308 b. .0075 c. 50 d. 111.5625
    e. 131.25

53. a. .9616 b. .0623

55. a. .5, $n(n + 1)/4$ b. .25, $n(n + 1)(2n + 1)/24$

57. 10:52.76

61. a. Bin(10, 18/38) b. Bin(15, 18/38)
    c. Bin(25, 18/38) f. no

65. c. Gamma$(n, 1/\lambda)$

67. a. 2 c. 0, 2n, $1/(1 - t^2)^n$ d. $1/(1 - t^2/2n)^n$

69. a. $f_X(x) = 2x, 0 < x < 1$
    b. $f_{Y|X}(y|x) = 1/x, 0 < y < x < 1$
    c. .6
    d. no, the domain is not a rectangle
    e. $E(Y|X = x) = x/2$ f. $V(Y|X = x) = x^2/12$

71. a. $f_X(x) = 2e^{-2x}, 0 < x < \infty$
    b. $f_{Y|X}(y|x) = e^{-y+x}, 0 < x < y < \infty$
    c. $P(Y > 2|x = 1) = 1/e$

d. no, the domain is not rectangular

e. $E(Y|X = x) = x + 1$ f. $V(Y|X = x) = 1$

73. a. $x/2$, $x^2/12$ b. $1/x$, $0 < y < x < 1$ c. $-\ln(y)$, $0 < y < 1$ d. 1/4, 7/144 e. 1/4, 7/144

75. a. $p_{Y|X}(0|1) = 4/17$, $p_{Y|X}(1|1) = 10/17$, $p_{Y|X}(2|1) = 3/17$

   b. $p_{Y|X}(0|2) = .12$, $p_{Y|X}(1|2) = .28$, $p_{Y|X}(2|2) = .60$ c. .40

   d. $p_{X|Y}(0|2) = 1/19$, $p_{X|Y}(1|2) = 3/19$, $p_{X|Y}(2|2) = 15/19$

77. a. $x^2/2$, $x^4/12$ b. $1/x^2$, $0 < y < x^2 < 1$

   c. $(1/\sqrt{y}) - 1$, $0 < y < 1$

79. a. $p(1,1) = p(2,2) = p(3,3) = 1/9$, $p(2,1) = p(3,1) = p(3,2) = 2/9$

   b. $p_X(1) = 1/9$, $p_X(2) = 3/9$, $p_X(3) = 5/9$

   c. $p_{Y|X}(1|1) = 1$, $p_{Y|X}(1|2) = 2/3$, $p_{Y|X}(2|2) = 1/3$, $p_{Y|X}(1|3) = .4$, $p_{Y|X}(2|3) = .4$, $p_{Y|X}(3|3) = .2$

   d. $E(Y|X=1) = 1$, $E(Y|X=2) = 4/3$, $E(Y|X=3) = 1.8$, no

   e. $V(Y|X=1) = 0$, $V(Y|X=2) = 2/9$, $V(Y|X=3) = .56$

81. a. $p_{X|Y}(1|1) = .2$, $p_{X|Y}(2|1) = .4$, $p_{X|Y}(3|1) = .4$, $p_{X|Y}(2|2) = 1/3$, $p_{X|Y}(3|2) = 2/3$, $p_{X|Y}(3|3) = 1$

   b. $E(X|Y=1) = 2.2$, $E(X|Y=2) = 8/3$, $E(X|Y=3) = 3$

   c. $V(X|Y=1) = .56$, $V(X|Y=2) = 2/9$, $V(X|Y=3) = 0$

83. a. $p_X(x) = .1$, $x = 0, 1, 2, \ldots, 9$; $p_{Y|X}(y|x) = 1/9$, $y = 0, 1, 2, \ldots, 9$, $y \neq x$; $p_{X,Y}(x, y) = 1/90$, $x, y = 0, 1, 2, \ldots, 9$, $y \neq x$

   b. $E(Y|X=x) = 5 - x/9$, $x = 0, 1, 2, \ldots, 9$

85. a. $.6x$, $.24x$ b. 60 c. 60

87. 176, 12.68

89. a. $1 + 4p$, $4p(1 - p)$ b. 2598, 16,518,196

   c. $2598(1 + 4p)$, $\sqrt{16518196 + 93071200p - 26998416p^2}$

   d. 2598 and 4064, 7794 and 7504, 12,990 and 9088

91. a. normal, mean = 984, variance = 38,988

   b. .1379 c. 1237

93. a. $N(158, 8.72)$ b. $N(170, 8.72)$ c. .4090

95. a. $.8875x + 5.2125$ b. 111.5775

   c. 10.563 d. .0951

97. a. $2x - 10$ b. 9 c. 3 d. .0228

99. a. .1410 b. .1165

   With positive correlation, the deviations from their means of X and Y are likely to have the same sign.

101. a. $\frac{1}{4\pi} e^{-(y_1^2 + y_2^2)/4}$ b. $\frac{1}{\sqrt{4\pi}} e^{-y_1^2/4}$ c. Yes

103. a. $f(y) = y(2 - y)$, $0 \leq y \leq 1$

   b. $f(w) = 2(1 - w)$, $0 \leq w \leq 1$

105. $4y_3[\ln(y_3)]^2$ for $0 < y_3 < 1$

109. a. $g_5(y) = 5x_4/10^5$, 25/3 b. 20/3 c. 5

   d. 1.409

111. $g_{Y_5|Y_1}(y_5|4) = [2/3][(y_5 - 4)/6]^3$, $4 < y_5 < 10$; 8.8

113. $1/(n+1)$, $2/(n+1)$, $3/(n+1)$, …, $n/(n+1)$

115. $\frac{\Gamma(n+1)\Gamma(i+1/\theta)}{\Gamma(i)\Gamma(n+1+1/\theta)}$,

   $\frac{\Gamma(n+1)\Gamma(i+2/\theta)}{\Gamma(i)\Gamma(n+1+2/\theta)} - \left[\frac{\Gamma(n+1)\Gamma(i+1/\theta)}{\Gamma(i)\Gamma(n+1+1/\theta)}\right]^2$

117. a. .0238 b. $2025

121. a. $n(n - 1)[F(y_n) - F(y_1)]^{n-2}f(y_1)f(y_n)$ for $y_1 < y_n$

   b. $f_W(w) = \int n(n - 1)[F(w + w_1) - F(w_1)]^{n-2}f(w_1)f(w + w_1)dw_1$

   c. $n(n-1)w^{n-2}(1 - w)$ for $0 \leq w \leq 1$

123. $f_T(t) = e^{-t/2} - e^{-t}$ for $t > 0$

125. a. 3/81,250

   b. $f_X(x) = \int_{20-x}^{30-x} kxy\,dy = k(250x - 10x^2)$ $0 \leq x \leq 20$

   $\int_0^{30-x} kxy\,dy = k(450x - 30x^2 + \frac{1}{2}x^3)$. $20 \leq x \leq 30$

   $f_Y(y) = f_X(y)$

   dependent

   c. .3548 d. 25.969 e. -32.19, $-.894$

   f. 7.651

127. 7/6

131. c. If $p(0) = .3$, $p(1) = .5$, $p(2) = .2$, then 1 is the smaller of the two roots, so extinction is certain in this case with $\mu < 1$. If $p(0) = .2$, $p(1) = .5$, $p(2) = .3$,

then 2/3 is the smaller of the two roots, so extinction is not certain with $\mu > 1$.

133. a. $P((X, Y) \in A) = F(b, d) - F(b, c) - F(a, d) + F(a, b)$

   b. $P((X, Y) \in A) = F(10, 6) - F(10, 1) - F(4, 6) + F(4, 1)$
   $P((X, Y) \in A) = F(b, d) - F(b, c - 1) - F(a - 1, d) + F(a - 1, b - 1)$

   c. At each (x^*, y^*), $F(x^*, y^*)$ is the sum of the probabilities at points (x, y) such that $x \leq x^*$ and $y \leq y^*$

   | | x | |
$F(x, y)$	100	250
200	.50	1
y 100	.30	.50
0	.20	.25

   d. $F(x, y) = .6x^2y + .4xy^3$, $0 \leq x \leq 1$;
   $0 \leq y \leq 1$; $F(x, y) = 0$, $x \leq 0$;
   $F(x, y) = 0$, $y \leq 0$;
   $F(x, y) = .6x^2 + .4x$, $0 \leq x \leq 1, y > 1$;
   $F(x, y) = .6y + .4y^3$, $x > 1, 0 \leq y \leq 1$;
   $F(x, y) = 1$, $x > 1, y > 1$
   $P(.25 \leq x \leq .75, .25 \leq y \leq .75) = .23125$

   e. $F(x, y) = 6x^2y^2$, $x + y \leq 1, 0 \leq x \leq 1$;
   $0 \leq y \leq 1, x \geq 0, y \geq 0$
   $F(x, y) = 3x^4 - 8x^3 + 6x^2 + 3y^4 - 8y^3 + 6y^2 - 1$, $x + y > 1, x \leq 1, y \leq 1$
   $F(x, y) = 0, x \leq 0; F(x, y) = 0, y \leq 0;$
   $F(x, y) = 3x^4 - 8x^3 + 6x^2$, $0 \leq x \leq 1, y > 1$
   $F(x, y) = 3y^4 - 8y^3 + 6y^2$, $0 \leq y \leq 1, x > 1$
   $F(x, y) = 1$, $x > 1, y > 1$

135. a. $2x, x$ b. 40 c. 100

137. $\dfrac{2}{(1 - 1000t)(2 - 1000t)}$, 1500 hours

141. a. 2360, 73.7021 b. .9713

143. .8340

145. a. $\dfrac{\sigma_W^2}{\sigma_W^2 + \sigma_E^2}$ b. .9999

147. 26, 1.64

Chapter 6

1. a.
$\bar{x}$	25	32.5	40	45	52.5	65
$p(\bar{x})$	.04	.20	.25	.12	.30	.09

   $E(\overline{X}) = 44.5 = \mu$

   b.
s^2	0	112.5	312.5	800
$p(s^2)$	.38	.20	.30	.12

   $E(S^2) = 212.25 = \sigma^2$

3.
x/n	0	.1	.2	.3	.4
$p(x/n)$	0.0000	0.0000	0.0001	0.0008	0.0055

.5	.6	.7	.8	.9	1.0
0.0264	0.0881	0.2013	0.3020	0.2684	0.1074

5. a.
$\bar{x}$	1	1.5	2	2.5	3	3.5	4
$p(\bar{x})$	.16	.24	.25	.20	.10	.04	.01

   b. $P(\overline{X} \leq 2.5) = .85$

   c.
r	0	1	2	3
$p(r)$	.30	.40	.22	.08

   d. .24

7.
$\bar{x}$	$p(\bar{x})$	$\bar{x}$	$p(\bar{x})$	$\bar{x}$	$p(\bar{x})$
0.0	0.000045	1.4	0.090079	2.8	0.052077
0.2	0.000454	1.6	0.112599	3.0	0.034718
0.4	0.002270	1.8	0.125110	3.2	0.021699
0.6	0.007567	2.0	0.125110	3.4	0.012764
0.8	0.018917	2.2	0.113736	3.6	0.007091
1.0	0.037833	2.4	0.094780	3.8	0.003732
1.2	0.063055	2.6	0.072908	4.0	0.001866

11. a. 12, .01
    b. 12, .005
    c. With less variability, the second sample is more closely concentrated near 12.

13. a. No, the distribution is clearly not symmetric. A positively skewed distribution—perhaps Weibull, lognormal, or gamma.
    b. .0746
    c. .00000092. No, 82 is not a reasonable value for μ.

15. a. .8366 b. no

17. 43.29

19. a. .9772, .4772 b. 10

21. a. .9838 b. .8926 c. .9862 and .8934, both quite close

27. $1/\overline{X}$

29. Because χ_v^2 is the sum of v independent random variables, each distributed as χ_1^2, the Central Limit Theorem applies.

35. a. 3.2 b. 10.04, the square of (a)

39. a. 4.32

41. a. $v_2/(v_2 - 2)$, $v_2 > 2$
    b. $2v_2^2(v_1 + v_2 - 2) \ /[v_1(v_2 - 2)^2(v_2 - 4)]$, $v_2 > 4$

49. a. The approximate value, .0228, is smaller because of skewness in the chi-squared distribution
    b. This approximation gives the answer .03237, agreeing with the software answer to this number of decimals.

53. a. .9686 b. .90 c. .87174

55. .048

57. a. .9544 for all n b. .8839, .9234, .9347; increases with n toward (a)

59. a. 2.6, 1.2 b. 390, 14.7 c. ~ 1

61. .9686

63. .0722

65. a. .5774, .8165, .9045
    b. 1.312, 4.303, 18.216

67. a. .049 b. .09

Chapter 7

1. a. 113.73, $\overline{X}$ b. 113, $\tilde{X}$
   c. 12.74, S, an estimator for the population standard deviation

d. The sample proportion of students exceeding 100 in IQ is 30/33 = .91
e. .112, $S/\overline{X}$

3. a. 1.3481, $\overline{X}$ b. .0846 c. 1.3481, $\overline{X}$
   d. 1.78, $\overline{X} + 1.282S$ e. .6736

5. $\hat{\theta}_1 = N\overline{X}$, $\hat{\theta}_2 = T - N\overline{D}$, $\hat{\theta}_3 = T \cdot \overline{X}/\overline{Y}$;
   1,703,000, 1,591,300, 1,601,438.281

7. a. 120.6 b. 1,206,000, 10,000$\overline{X}$ c. .8
   d. 120, $\tilde{X}$

9. a. $\overline{X}$, 2.113 b. $\sqrt{\mu/n}$, .119

11. b. $\sqrt{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2}$
    c. In part (b) replace p_1 with X_1/n_1 and replace p_2 with X_2/n_2
    d. $-.245$ e. .0411

13. c. $\sqrt{n^2/[(n - 1)^2(n - 2)\overline{x}^2]}$

17. a. $\hat{\theta} = \sum X_i^2/(2n)$ b. 74.505

19. 4/9

21. a. $\hat{p} = 2\hat{\lambda} - .30 = .20$
    b. $\hat{p} = (100\hat{\lambda} - 9)/70$

25. a. .15 b. yes c. .4437

27. a. $\hat{\theta} = (2\overline{x} - 1)/(1 - \overline{x}) = 3$
    b. $\hat{\theta} = [-n/\Sigma \ln(x_i)] - 1 = 3.12$

29. $\hat{p} = r/x = .15$ This is the number of successes over the number of trials, the same as the result in Exercise 25. It is not the same as the estimate of Exercise 19.

31. a. $(2\pi\theta)^{-n/2}e^{-\Sigma x_i^2/2\theta}$
    b. $-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\theta) - \frac{\Sigma x_i^2}{2\theta}$
    c. $\sum x_i^2/n$
    d. $n/\sum x_i^2$

33. a. $\sum x_i^2/2n$=74.505, the same as Exercise 17
    b. $\sqrt{2\hat{\theta}\ln(2)}$=10.16

35. $\hat{\lambda} = -\ln(\hat{p})/24 = .0120$

37. a. $\overline{X}$ b. $\tilde{X}$

39. No, statistician A does not have more information.

41. $\prod_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i$

43. $\sum x_i$

45. $(\min\{X_i\}, \max\{X_i\})$

47. $2X(n - X)/[n(n - 1)]$

49. $\overline{X}$, by the Rao-Blackwell Theorem, because $\overline{X}$ is sufficient for μ

51. a. $1/p^2(1 - p)$
    b. $n/p^2(1 - p)$
    c. $p^2(1 - p)/n$

53. a. If we ignore the boundary, $1/\theta^2$ b. θ^2/n
    c. Both are less than θ^2/n; Cramér-Rao does not apply because the boundaries of the uniform variable X include θ itself

55. a. $\bar{x}$ b. $N(\mu, \sigma/\sqrt{n})$ c. Yes   d. They agree

57. a. $2/\sigma^2$ b. Yes

59. a. $1/p, (1 - p)/np^2$ b. $(1 - p)/np^2$ c. Yes

61. $\hat{\lambda} = 6/(6t_6 - t_1 - \cdots - t_5) =$
    $6/(x_1 + 2x_2 + \cdots + 6x_6) = .0436$, where
    $x_1 = t_1, x_2 = t_2 - t_1, \ldots, x_6 = t_6 - t_5$

63. 2.912, 2.242

67. 5.93, 11.66

69. b. no, $E(\hat{\sigma}^2) = \sigma^2/2$, so $2\hat{\sigma}^2$ is unbiased

73. .448, .4364

75. $d(X) = (-1)^X, d(200) = 1, d(199) = -1$

77. $\hat{\beta} = \sum x_i y_i / \sum x_i^2 = 30.040$, the estimated minutes per item; $\hat{\sigma}^2 = \frac{1}{n} \sum (y_i - \hat{\beta} x_i)^2 = 16.912; 25\hat{\beta} = 751$

Chapter 8

1. a. 99.5% b. 85% c. 2.97 d. 1.15

3. a. A narrower interval has a lower probability b. No, μ is not random
   c. No, the interval refers to μ, not individual observations
   d. No, a probability of .95 does not guarantee 95 successes in 100 trials

5. a. (4.52, 5.18)   b. (4.12, 5.00)   c. 55 d. 94

7. Increase n by a factor of 4. Decrease the width by a factor of 5.

9. a. $\bar{x} - z_\alpha \sigma/\sqrt{n}$ b. 4.418   c. 59.70

11. 950; .8724 (normal approximation), .8731 (binomial)

13. a. 1.341 b. 1.753 c. 1.708 d. 1.684 e. 2.704

15. a. 2.228 b. 2.131 c. 2.947 d. 4.604 e. 2.492
    f. 2.715

17. a. Yes b. (4.89, 5.79) c. (.5868, .6948) fl oz

19. a. (63.12, 66.66) b. No, data indicates the population is not normal

21. a. (29.26, 40.78) b. (–3.61, 73.65); times are normally distributed; no

23. a. (18.94, 24.86)   b. narrower   c. narrower
    d. (12.09, 31.71)

25. a. Assuming normality, a 95% lower confidence bound is 8.11. When the bound is calculated from repeated independent samples, roughly 95% of such bounds should be below the population mean.
    b. A 95% lower prediction bound is 7.03. When the bound is calculated from repeated independent samples, roughly 95% of such bounds should be below the value of an independent observation.

27. a. 378.85   b. 413.14   c. (340.16, 401.22)

29. (–8228.0, 116,042.4); yes (note that negative values in PI make its validity suspect)

31. a. (169.36, 179.37)
    b. (134.30, 214.43), which includes 152
    c. The second interval is much wider, because it allows for the variability of a single observation.
    d. The normal probability plot gives no reason to doubt normality. This is especially important for part (b), but the large sample size implies that normality is not so critical for (a).

33. a. (18.413, 20.102) b. 18.852 c. data indicates the population distribution is not normal

35. (18.3, 19.7) days

37. 0.056; yes, though potentially not by much

39. 97

41. a. 80% b. 98% c. 75%

43. (.798, 845)

45. a. .504 b. Yes, since $p > .504 > .5$

47. .584

49. (.513, .615)

51. .441; yes

53. a. 381 b. 339

55. (.028, .167)

57. a. 22.307 b. 34.381 c. 44.313 d. 49.925
    e. 11.523 f. 10.519

59. (0.4318, 1.4866), (.657, 1.219)

61. b. (2.34, 5.60) c. False

63. a. (7.91, 12.00) b. Yes
    c. The accompanying R code assumes the data has been read in as the vector x.

```
N=5000
xbar=rep(0,N)
for (i in 1:N){
 resample = sample(x,length(x),
replace = T)
 xbar[i]=mean(resample)
}
```

d. (7.905, 12.005) for one simulation; bootstrap distribution is somewhat skewed, so validity is questionable
   e. (8.204, 12.091) for one simulation
   f. Bootstrap percentile interval; population and bootstrap distributions are both skewed

65. a. (26.61, 32.94)
    b. Because of outliers, weight gain does not seem normally distributed. However, with $n = 68$, the effects of the CLT might be enough to validate use of t procedures anyway.
    d. (26.66, 32.90) for one simulation; yes, because histogram is bell-shaped
    e. (26.69, 32.88) for one simulation
    f. All three are close, so one-sample t should be considered valid

67. a. (38.46, 38.84)
    b. Although a normal probability plot is not perfectly straight, there is not enough deviation to reject normality.
    d. (38.47, 38.83) for one simulation; possibly invalid because bootstrap distribution is somewhat skewed
    e. (38.51, 38.81) for one simulation
    f. All three intervals are surprisingly similar.
    g. Yes: all CIs are well above normal body temperature of 37°C

69. a. (170.75, 183.57)
    b. Plot is reasonably linear, CI in part (a) is legitimate.
    d. (170.76, 183.46) for one simulation; very similar to (a)
    e. (171.04, 183.00) for one simulation
    f. All three are valid, so use the shortest: (171.04, 183.00). None of the CIs capture the true μ value

71. 246

73. a. .614   b. 4727.8, no
    c. Yes: .66(7700) = 5082 > 5000

75. a. (.163, .174)   b. (.089, .326)

77. (.1295, .2986)

79. a. At 95% confidence, average TV viewing time for the population of all 0–11 months old children is between 0.8 and 1.0 hours per day. (Similar interpretation for others.)
  b. Samples with larger standard deviations and/or smaller sample sizes will result in wider intervals.
  c. Yes: none of the intervals overlap

81. c. $\sigma^2/\Sigma x_i^2$, $\sigma/\sqrt{\Sigma x_i^2}$ d. Spread out: variance is inversely proportional to sum of squares of x values
  e. $\hat{\beta} \pm t_{.025,n-1}s/\sqrt{\Sigma x_i^2}$; (29.93, 30.15)

85. a. $\chi^2_{.95,2n}/2\Sigma X_i$, .0098
  b. $\exp(-t \cdot \chi^2_{.05,2n}/2\Sigma X_i) = .058$

87. a. $(\bar{x} - t_{.025,n-1,\delta} \cdot s/\sqrt{n}, \bar{x} - t_{.975,n-1,\delta} \cdot s/\sqrt{n})$
  b. (3.01, 4.46)

89. a. $1/2^n$ b. $n/2^n$ c. $(n+1)/2^n$, $1 - (n+1)/2^{n-1}$, (29.9, 39.3) with confidence level .9785

91. a. $P(A_1 \cap A_2) = .95^2 = .9025$
  b. $P(A_1 \cap A_2) \geq .90$
  c. $P(A_1 \cap A_2) \geq 1 - 2\alpha$;
  $P(A_1 \cap A_2 \cap \cdots \cap A_k)$
  $\geq 1 - k\alpha$

Chapter 9

1. a. yes b. no c. no d. yes e. no f. yes

5. H_0: $\sigma = .05$ versus H_a: $\sigma < .05$. Type I error: Conclude that the standard deviation is less than .05 mm when it is really equal to .05 mm. Type II error: Conclude that the standard deviation is .05 mm when it is really less than .05.

7. A type I error here involves saying that the plant is not in compliance when in fact it is. A type II error occurs when we conclude that the plant is in compliance when in fact it isn't. A government regulator might regard the type II error as being more serious.

9. a. R_1
  b. Reject H_0

  c. A type I error involves saying that the two companies are not equally favored when they are. A type II error involves saying that the two companies are equally favored when they are not.
  d. Bin(25, .5); .0433
  e. $\beta(.3) = \beta(.7) = .488$;
  $\beta(.4) = \beta(.6) = .845$;
  power = .512 for .3 and .7, .155 for .4 and .6

11. a. H_0: $\mu = 10$ versus H_a: $\mu \neq 10$
  b. .01
  c. .5319. .0078
  d. $c = 2.58$
  e. $c = 1.96$
  f. $\bar{x} = 10.02$, so do not reject H_0

13. c. .0004, ~ 0, P(type I error) $\leq \alpha = .01$ when $\mu < \mu_0$

15. a. .0301 b. .003 c. .004

17. Test H_0: $\mu = .5$ versus H_a: $\mu \neq .5$

  a. Do not reject H_0 because $t_{.025,12} = 2.179 > |1.6|$
  b. Do not reject H_0 because $t_{.025,12} = 2.179 > |-1.6|$
  c. Do not reject H_0 because $t_{.005,24} = 2.797 > |-2.6|$
  d. Reject H_0 because $t_{.005,24} = 2.797 < |-3.9|$

19. a. Do not reject H_0 because $|-2.27| < 2.576$
  b. .2266
  c. 22

21. $z = -2.14$, so reject H_0 at .05 level but not at .01 level

23. Because $t = 2.24 > 1.708 = t_{.05,25}$, reject H_0: $\mu = 360$. Yes, this suggests contradiction of prior belief.

25. a. Because $|-1.40| < 2.064$, H_0 is not rejected at the .05 level.
  b. 600 lies in the CI for μ

27. a. no, $t = -.02$ b. .58 c. $n = 20$ total observations

29. Since $1.04 < 2.132$, we do not reject H_0 at the .05 significance level.

31. a. Because $t = .50 < 1.89 = t_{.05,7}$ do not reject H_0.
    b. .73

33. Because $t = -1.24 > -1.40 = -t_{.10,8}$, we do not have evidence to question the prior belief.

37. a. $1 - F\left(t_{\alpha,n-1}; n-1, \dfrac{\mu' - \mu_0}{\sigma/\sqrt{n}}\right)$

    b. $F\left(-t_{\alpha/2,n-1}; n-1, \dfrac{\mu' - \mu_0}{\sigma/\sqrt{n}}\right)$
    $+ 1 - F\left(t_{\alpha/2,n-1}; n-1, \dfrac{\mu' - \mu_0}{\sigma/\sqrt{n}}\right)$

39. Since $|-2.469| \geq 1.96$, reject H_0.

41. a. Do not reject H_0: $p = .10$ in favor of H_a: $p > .10$ because 16 or more blistered plates would be required for rejection at the .05 level. Because H_0 is not rejected, there could be a type II error.
    b. $\beta(.15) = .4920$ when $n = 100$;
       $\beta(.15) = .2743$ when $n = 200$
    c. 362

43. a. Do not reject H_0: $p = .02$ in favor of H_a: $p < .02$ because $z = -1.01$ is not in the rejection region at the .05 level. There is no strong evidence suggesting that the inventory be postponed.
    b. $\beta(.01) = .195$
    c. $1 - \beta(.05) \approx 0$

45. a. Test H_0: $p = .05$ versus H_a: $p \neq .05$. Since $z = 3.07 > 2.58$, H_0 is rejected. The company's premise is not correct.
    b. $\beta(.10) = .033$

47. Using $n = 25$, the probability of 5 or more leaky faucets is .0980 if $p = .10$, and the probability of 4 or fewer leaky faucets is .0905 if $p = .3$. Thus, the rejection region is 5 or more, $\alpha = .0980$, and $\beta = .0905$.

49. a. reject b. reject   c. do not reject

d. reject e. do not reject

51. a. .0778 b. .1841 c. .0250 d. .0066
    e. .5438

53. a. $P = .0403$ b. $P = .0176$ c. $P = .1304$
    d. $P = .6532$ e. $P = .0021$ f. $P = .00022$

55. Based on the given data, there is no reason to believe that pregnant women differ from others in terms of serum receptor concentration.

57. a. Because the P-value is .166, no modification is indicated.
    b. .9974

59. Because $t = -1.759$ and the P-value = .082, which is less than .10, reject H_0: $\mu = 3.0$ against a two-tailed alternative at the 10% level. However, the P-value exceeds .05, so do not reject H_0 at the 5% level. There is just a weak indication that the percentage is not equal to 3% (lower than 3%).

61. a. Test H_0: $\mu = 10$ versus H_a: $\mu < 10$
    b. Because the P-value is .017 < .05, reject H_0, suggesting that the pens do not meet specifications.
    c. Because the P-value is .045 > .01, do not reject H_0, suggesting there is no reason to say the lifetime is inadequate.
    d. Because the P-value is .0011, reject H_0. There is good evidence showing that the pens do not meet specifications.

63. Do not reject H_0 at .01 or .05, reject H_0 at .10

65. b. 36.614 c. yes

67. a. $\Sigma x_i > c$ b. yes

69. Yes, the test is UMP for the alternative H_a: $\theta > .5$ since the tests for H_0: $\theta = .5$ versus H_a: $\theta = p_0$ all have the same form for $p_0 > .5$.

71. b. .0502 c. .04345, .05826, no d. .05114; not most powerful

73. $-2\ln(\Lambda) = 3.041$, P-value = .081

75. a. .98, .85, .43, .004, .0000002
    b. .40, .11, .0062, .0000003
    c. Because the null hypothesis will be rejected with high probability, even with only slight departure from the null hypothesis, it is not very useful to do a .01 level test.

77. a. $\frac{S^2 - \sigma_0^2}{\sqrt{2\sigma_0^4/(n-1)}}$
    b. P-value $= P(Z \leq -3.59) \approx .0002$, so reject H_0.

79. a. 16.803
    b. reject H_0 because 15 is not > 16.803
    c. no
    d. reject H_0 at .10, uncertain at .01

81. The following R code performs the bootstrap simulation described in this section.

```
mu0 = 113; N = 5000
x = c(117.6,109.5,111.6,109.2,119.1,110.8)
w = x - mean(x) + mu0
wbar = rep(0,N) # allocating space for
bootstrap means
for (i in 1:N){
   resample  =  sample(w,length(w),re-
place=T)
 wbar[i] = mean(resample)
}
```

The P-value is estimated by the proportion of these $\bar{w}_i^*$ values that are at or below the observed $\bar{x}$ value of 112.9667. In one run of this code, that proportion was .5018, so do not reject H_0.

83. a. H_0: the reformulated drug *is no safer than* the original, recalled drug.
    H_a: the reformulated drug is safer than the recalled drug.
    b. Type I error: The FDA rejects H_0 and concludes the new drug is safer, when in fact it isn't. Type II error: The FDA fails to recognize that H_a is true, yet the new drug is indeed safer.
    c. Type I (arguably); lower α

85. Yes, only 25 required

87. $t = 6.4$, P-value ≈ 0, reject H_0

89. a. no
    b. $t = .44$, P-value $= .33$, do not reject H_0

91. Assuming normality, calculate $t = 1.70$, which gives a two-tailed P-value of .102. Do not reject the null hypothesis H_0: $\mu = 1.75$.

93. The P-value for a lower tail test is .0014, so it is reasonable to reject the idea that $p = .75$ and conclude that fewer than 75% of mechanics can identify the problem.

95. Because the P-value is .013 $>$ .01, do not reject the null hypothesis at the .01 level.

97. a. For testing H_0: $\mu = \mu_0$ versus H_a: $\mu > \mu_0$ at level α, reject H_0 if $2\Sigma x_i/\mu_0 > \chi^2_{\alpha,2n}$
    For testing H_0: $\mu = \mu_0$ versus H_a: $\mu < \mu_0$ at level α, reject H_0 if $2\Sigma x_i/\mu_0 < \chi^2_{1-\alpha,2n}$
    For testing H_0: $\mu = \mu_0$ versus H_a: $\mu \neq \mu_0$ at level α, reject H_0 if $2\Sigma x_i/\mu_0 > \chi^2_{\alpha/2,2n}$ or if $2\Sigma x_i/\mu_0 < \chi^2_{1-\alpha/2,2n}$
    b. Because $\Sigma x_i = 737$, the test statistic value is $2\Sigma x_i/\mu_0 = 19.65$, which gives a P-value of .52. There is no reason to reject the null hypothesis.

99. a. yes

Chapter 10

1. a. $-.4$ b. .0724, .269
   c. Although the CLT implies that the distribution will be approximately normal when the sample sizes are each 100, the distribution will not necessarily be normal when the sample sizes are each 10.

3. a. $z = 4.84 \geq 1.96$, reject H_0
   b. (1251, 2949)

5. a. H_a says that the average calorie output for sufferers is more than 1 cal/cm^2/min below that for non-sufferers. Reject H_0 in favor of H_a because $z = -2.90 < -2.33$
   b. .0019
   c. .82, .18
   d. 66

7. a. We must assume here that the population elapsed time distributions are both normal.
   b. $z = 1.32 < 2.576$, do not reject H_0

9. 22, no

11. b. It decreases.

13. a. 17   b. 21   c. 18   d. 26

15. $H_0: \mu_1 - \mu_2 = 0$ versus $H_a: \mu_1 - \mu_2 < 0$; $t = -15.83$, reject H_0

17. a. $t = -18.64 \leq -1.686$, strongly reject H_0
    b. $t = 15.66 \geq 1.680$, again strongly reject H_0
    c. at most .10

19. a. (219.6, 538.4)
    b. $t = 2.20$, P-value = .014, reject H_0

21. a. No, mean < sd so positively skewed; no
    b. ($115, $375)

23. Because $t = -3.35 < -3.30 = t_{.001,42}$, yes, there is evidence that experts do hit harder.

25. b. No c. Because $|t| = |-.38| < 2.23 = t_{.025,10}$, no, there is no evidence of a difference.

27. Because the one-tailed P-value is .0004 < .01, conclude at the .01 level that the difference is as stated. This could result in a type I error.

29. Yes, because $t = 2.08$ with P-value = .046.

31. b. (127.6, 202.0) c. 131.75

33. Because $t = 1.82$ with P-value .046 < .05, conclude at the .05 level that the difference exceeds 1.

35. a. The slender distribution appears to have a lower mean and lower variance.
    b. With $t = 1.88$ and a P-value of .097, there is no significant difference at the .05 level.

37. With $t = 2.19$ and a two-tailed P-value of .031, there is a significant difference at the .05 level but not the .01 level.

41. a. $(\bar{x} - \bar{y}) \pm t_{\alpha/2,m+n-2} \cdot s_p \sqrt{1/m + 1/n}$
    b. (−455, 45)
    c. (−448, 38)

43. $t = 3.88 \geq 3.365$, so reject H_0

45. a. (.000046, .000446); yes, because 0 does not fall in the CI
    b. $t = 2.68$, P-value = .01, reject H_0

47. a. yes   b. $10,524   c. $|t| = |-1.21| < 1.729$, so do not reject H_0; yes

49. a. two-sample t
    b. $t = 2.47 \geq 1.681$, reject H_0
    c. paired t
    d. $t = -4.34 \leq -1.717$, reject H_0

51. b. (12.67, 25.16)

53. $t = -2.2$, P-value = .028, reject H_0

57. a. Because $|z| = |-4.84| > 1.96$, conclude that there is a difference. Rural residents are more favorable to the increase.
    b. .9967

59. (.016, .171)

61. a. (−.294, −.207)

63. $H_0: p_1 - p_2 = 0$ versus $H_a: p_1 - p_2 < 0$, $z = -2.01$, P-value = .022, reject H_0

65. a. p_1 = the proportion of all students who would agree to be surveyed by Melissa, p_2 = the proportion of all students who would agree to be surveyed by Kristine; $z = 3.00$, P-value = .003, reject H_0
    b. No

67. 769

69. a. $H_0: p_3 = p_2$ versus $H_a: p_3 > p_2$
    b. $\hat{p}_3 - \hat{p}_2 = (X_3 - X_2)/n$
    c. $(X_3 - X_2)/\sqrt{X_2 + X_3}$
    d. $z = 2.68$, P-value = .0037, reject H_0 at .01 but not at .001.

71. a. 3.69   b. 4.82   c. .207   d. .271
    e. 4.30   f. .212   g. .95   h. .94

73. $f = 1.814 < F_{.10,9,7} = 2.72$, so P-value $> .10$ $> .01$, do not reject H_0

75. $f = 4.38 < F_{.01,11,9} = 5.18$, do not reject H_0

77. (0.87, 2.41)

79. a. (.158, .735)
    b. Bootstrap distribution of differences looks quite normal.
    c. (.171, .723) for one simulation
    d. (.156, .740) for one simulation
    e. All three intervals are quite similar.
    f. Students on lifestyle floors appear to have a higher mean GPA, somewhere between $\sim .16$ higher and $\sim .73$ higher.

81. a. (0.593, 1.246); normal probability plots show departures from normality, CI might not be valid.
    b. The R code below assumes two vectors, L and N, contain the original data.

```
ratio = rep(0,5000)
for (i in 1:5000){
  L.resamp = sample(L,length(L),
replace=T)
  N.resamp = sample(N,length(N),
replace=T)
 ratio[i] = sd(L.resamp)/sd(N.re-
samp)
}
```

CI = (0.568, 1.289) for one simulation

83. a. The bootstrap distribution of differences of medians is definitely not normal.
    b. (0.38, 10.44) for one simulation
    c. (0.4706, 10.0294) for one simulation

85. a. $t = 2.62$, df = 17, P-value $= .018$, reject H_0 at the .05 level
    b. In the R code below, the data is read is as a data frame called df with two columns, Time and Group. The first lists the times for each rat, while the second has B and C labels.

```
N = 5000
diff = rep(0,N)
for (i in 1:N){
```

```
 resample = sample(df$Time, length
(df$Time), replace=T)
     C.resamp  =  resample[df
$Group=="C"]
     B.resamp  =  resample[df
$Group=="B"]
 diff[i] = mean(C.resamp) - mean(B.
resamp)
}
```

P-value $= 2$(proportion of simulated differences $> 10.59 - 5.71$) $= .02$ for one simulation

    c. Results of (a) and (b) are similar

87. a. $f = 4.46$; $F_{.95,6,5} = 0.228 < 4.46 < F_{.05,6,5}$ $= 4.95$, so do not reject H_0 at .10 level.
    b. Use code similar to Exercise 85, but change the last line to calculate the ratio of resampled variances. Observed ratio $=$ 4.48, proportion of ratio values ≥ 4.48 was .086 for one simulation, so P-value $=$ $2(.086) = .172$, and H_0 is again not rejected.

89. a. Use the code from Exercise 85. Observed difference $= 3.47$, proportion of simulated differences ≥ 3.47 was .019 for one simulation, so P-value $= 2(.019) =$ .038. Reject H_0 at the .05 level.
    b. Results are similar; not surprising since both methods are valid.

91. a. ($6.40, $11.85)
    b. Use code provided in Chapter 8; bootstrap distribution of $\bar{d}$ is not normal.
    c. ($6.44, $11.81) for one simulation
    d. ($6.23, $11.51) for one simulation
    e. (a) and (c) are similar, while (d) is shifted to the left; (d) is most trustworthy
    f. On average, books cost between $6.23 and $11.51 more with Amazon than at the campus bookstore!

95. The difference is significant at the .05, .01, and .001 levels.

99. b. No, given that the 95% CI includes 0, the test at the .05 level does not reject equality of means.

101. (−299.2, 1517.8)

103. (1020.2, 1339.9) Because 0 is not in the CI, we would reject equality of means at the .01 level.

105. Because $t = 2.61$ and the one-tailed P-value is .007, the difference is significant at the .05 level using either a one-tailed or a two-tailed test.

107. a. μ_1 = true mean AEDI score improvement for all 2001 students; H_0: $\mu_1 = 0$ versus H_a: $\mu_1 > 0$; $t = 2.41$, df = 36, P-value = .011, reject H_0.
    b. Similarly, $t = 2.19$, P-value = .020, reject H_0.
    c. H_0: $\mu_1 - \mu_2 = 0$ versus H_0: $\mu_1 - \mu_2 < 0$, $t = -0.23$, P-value = .411, do not reject H_0. The data does not suggest an "Enron effect."

109. Because $t = 7.50$ and the one-tailed P-value is .0000001, the difference is highly significant, assuming normality.

111. The two-sample t test is inappropriate for paired data. The paired t gives a mean difference .3, $t = 2.67$, and the two-tailed P-value is .045, so the means are significantly different at the .05 level. We are concluding tentatively that the label understates the alcohol percentage.

113. Because the paired $t = 3.88$ and the two-tailed P-value is .008, the difference is significant at the .05 and .01 levels, but not at the .001 level.

115. a. $t = 11.86 > 2.33$, reject H_0 at .01 level.
    b. $t = 8.99$, again clearly reject H_0.
    c. Yes, because students were randomly assigned to experimental groups.

117. .902, .826, .029, .00000003

119. Because $z = 4.25$ and the one-tailed P-value is .00001, the difference is highly significant and companies do discriminate.

121. With $Z = (\overline{X} - \overline{Y})/\sqrt{\overline{X}/n + \overline{Y}/m}$, the result is $z = -5.33$, two-tailed P-value = .0000001, so one should conclude that there is a significant difference in parameters

Chapter 11

1. a. Reject H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ in favor of H_a: $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$ not all the same, because $f = 5.57 > 2.69 = F_{.05,4,30}$.
    b. Using Table A.8, $.001 < P\text{-value} < .01$. (The P-value is .0018)

3. SSTr = 2304, SSE = 4200, $f = 5.76 \geq F_{.05,2,21} = 3.47$, reject H_0

5. a. SSTr = 9982.4, MSTr = 1109.16
   b.

Source	df	SS	MS	f
Brand	9	9982.4	1109.16	8.53
Error	30	3900.0	130.00	
Total	39			

   $8.53 \geq F_{.01,9,30} = 3.07$, so reject H_0.

7.

Source	df	SS	MS	f
Type	3	127375	42458	25.09
Error	20	33839	1692	
Total	23	161214		

   P-value $\approx$ .000, so reject H_0.

9. a. SSTr = 270, MSTr = 90, SSE = 17446, MSE = 167.75
   b. $f = 0.54 < F_{.05,3,104} \approx 2.69$, so H_0 is not rejected at the .05 level.

11. b. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ vs H_a: not all μ's are equal, $f = [18797.5/4] / [251.5/15] = 4699.38/16.77 = 280.28$, strongly reject H_0.

15. $Q_{.05,5,15} = 4.37$, and $4.37\sqrt{272.8/4} = 36.09$

3	1	4	2	5
437.5	462.0	469.3	512.8	532.1

17.

3	1	4	2	5
427.5	462.0	469.3	502.8	532.1

19.

4	3	1	2
562.02	698.07	713.00	756.93

21. (6.401, 10.589)

23. a.

SOO	BSP	CW	SWE	BMIPS	ST	BSF	N	GMIPS	GS
117	122	127	129	141	142	144	147	148	175

b. $(-16.79, -0.73)$

25. $422.16 < SSE < 431.88$

27. $SSTr = 465.5$, $SSE = 124.5$, $f = 17.12 \geq F_{.05,3,14} = 3.34$, so reject H_0 at .05 level.

29. a. With large sample sizes, normality is less important. $11.32/9.13 < 2$ indicates equal variances is plausible.

   b. $SSTr = 2445.7$, $SSE = 118,632.6$, $f = 20.92 \geq F_{.05,4,1015} = 2.38$, so H_0 is rejected.

   c. $Q_{.05,5,1015} \approx 3.86$, $d_{ij} = 3.86\sqrt{\frac{116.9}{2}\left(\frac{1}{J_i} + \frac{1}{J_j}\right)}$

Graduate	Freshman	Sophomore	Junior	Senior
45.55	48.95	51.45	52.89	52.92

31. μ_i = true mean impact of social media, as a percentage of sales, for the ith category; $SSTr = 3804$, $SSE = 76973$, $f = 8.85 \geq F_{.01,2,358} \approx 4.66$, so H_0 is rejected at the .01 significance level.

33. a. The distributions of the polyunsaturated fat percentages for each of the four regimens must be normal with equal variances.

   b. $SSTr = 8.334$, $SSE = 77.79$, $f = 1.714 < F_{.10,3,50} = 2.20$, so P-value $> .10$ and H_0 is not rejected.

35. μ_i = true mean change in CMS under the ith treatment; $SSTr = 19.84$, $SSE = 16867.6$, $f = 0.1129 < F_{.05,2,96} \approx 3.09$, so H_0 is not rejected at the .05 level.

37. When H_0 is true, all the α_i's are 0, and $E(MSTr) = \sigma^2$. Otherwise, $E(MSTr) > \sigma^2$.

39. $\lambda = 10$, $F_{.05,3,14} = 3.344$. From R, $\beta =$ pf(3.344,df1=3,df2=14,ncp=10)= .372, and so power $= 1 - \beta = 1 - .372 = .628$.

41. a. The sample standard deviations are very different.

b. For the transformed data, $\bar{y}_{..} = 2.46$, $SSTr = 26.104$, $SSE = 7.748$, $f = 70.752$, so H_0 is clearly rejected.

c.

PW	MK	NW	50K	FR
1.30	2.28	2.62	2.75	2.82

43. $h(x) = \arcsin(\sqrt{x/n})$

45. a. $MSA = 7.65$, $MSE = 4.93$, $f_A = 1.55$. Since $1.55 < F_{.05,4,12} = 3.26$, don't reject H_{0A}.

   b. $MSB = 14.70$, $f_B = 2.98 < F_{.05,2,12} = 3.49$, don't reject H_{0B}.

47. a.

Source	df	SS	MS	f
Method	5	596,748	119,349.6	9.67
Block	16	529,100	3306.9	0.27
Error	80	987,380	12342.3	
Total	101	2,113,228		

   b. H_0: $\alpha_1 = \cdots = \alpha_5 = 0$ versus H_a: not all α's are 0. Since $9.67 \geq F_{.01,5,80} = 3.255$, we reject H_0 at the .01 level.

   c. $I = 6$, $J = 17$, $MSE = 12342.3$, $Q_{.01,6,80} \approx 4.93$, $HSD = 4.93\sqrt{12342.3/17} = 132.8$.

Wrist acc.	Hip acc.	Pedometer	Wrist + LFE	Hip + LFE	Hand tally
449	466	557	579	606	668

49. a.

Source	df	SS	MS	f	P-value
Spindle speed	2	16106	8052.8	10.47	0.026
Feed rate	2	2156	1077.8	1.40	0.346
Error	4	3078	769.4		
Total	8	21339			

   b. The test statistic value and P-value for H_{0A}: $\alpha_1 = \alpha_2 = \alpha_3 = 0$ versus H_{aA}: not all α's $= 0$ are $f = 10.47$ and $P = .026$. Since $.026 \leq .05$, we reject H_{0A} at the .05 level and conclude that mean temperature varies with spindle speed.

   c. The test statistic and P-value for H_{0B}: $\beta_1 = \beta_2 = \beta_3 = 0$ versus H_{aB}: not all β's $= 0$ are $f = 1.40$ and $P = .346$. Since $.346 > .05$, do not reject H_{0B} at the .05 level; conclude that feed rate has no statistically significant effect on mean temperature.

51. c.

Source	df	SS	MS	f
Flavor	2	20,797	10,398.5	35.0
Block (Subject)	53	135,833	2,562.9	8.6
Error	106	31,506	297.2	
Total	161	188,136		

With $f = 35.0 \geq F_{.01,2,106} \approx 4.81$, $H_{0A}: \alpha_1 = \alpha_2 = \alpha_3 = 0$ is rejected at the .01 level.

d. Yes

53.

Source	df	SS	MS	f	P-value
Current	2	106.78	53.39	0.19	0.833
Voltage	2	56.05	28.03	0.10	0.907
Error	4	1115.75	278.94		
Total	8	1278.58			

According to the ANOVA table, neither factor has a statistically significant effect at the .10 level: both P-values are > .10.

55. With $f = 8.69 > 6.01 = F_{.01,2,18}$, there are significant differences among the three treatment means. The normal plot of residuals shows no reason to doubt normality, and the plot of residuals against the fitted values shows no reason to doubt constant variance. There is no significant difference between treatments B and C, but Treatment A differs (it is lower) significantly from the others at the .01 level.

57. Because $f = 8.87 > 7.01 = F_{01,4,8}$, reject the hypothesis that the variance for B is 0.

61. a.

```
Source        df   SS         MS        F
A             2    30763      15381.5   3.79
B             3    34185.6    11395.2   2.81
Interaction   6    43581.2    7263.5    1.79
Error         24   97436.8    4059.9
Total         35   205966.6
```

b. Because $1.79 < 2.51 = F_{.05,6,24}$, there is no significant interaction.

c. Because $3.79 > 3.40 = F_{.05,2,24}$, there is a significant difference among the A means at the .05 level.

d. Because $2.81 < 3.01 = F_{.05,6,24}$, there is no significant difference among the B means at the .05 level.

e. Using $d = 64.93$,

3	1	2
3960.2	4010.88	4029.10

63. a. With $f = 1.55 < 2.81 = F_{.10,2,12}$, there is no significant interaction at the .10 level.

b. With $f = 376.27 > 18.64 = F_{.001,2,12}$, there is a significant difference between the formulation means at the .001 level. With $f = 19.27 > 12.97 = F_{.001,1,12}$, there is a significant difference among the speed means at the .001 level.

c. Main effects Formulation: (1) 11.19, (2) −11.19 Speed: (60) 1.99, (70) −5.03, (80) 3.04

65. a. Factor #1 = firing distance, levels = 25 yd, 50 yd. Factor #2 = bullet brand, levels = Federal, Remington, Winchester. Treatments: (25, Fed), (25, Rem), (25, Win), (50, Fed), (50, Rem), and (50, Win).

b. The interaction plot suggests a huge distance effect. There appears to be very little bullet manufacturer effect. The nonparallel pattern suggests perhaps a slight interaction effect.

c.

Source	df	SS	MS	f	P-value
Distance	1	568.97	568.969	242.56	0.000
Bullet	2	2.97	1.487	0.63	0.531
Distance * bullet	2	2.48	1.242	0.53	0.589
Error	444	1041.49	2.346		
Total	449	1615.92			

67.

```
Source        DF    SS        MS        F
pen           3     1387.5    462.50    0.34
surface       2     2888.1    1444.04   1.07
Interaction   6     8100.3    1350.04   1.97
Error         12    8216.0    684.67
Total         23    20591.8
```

With $f = 1.97 < 2.33 = F_{.10,6,12}$, there is no significant interaction at the .10 level.

With $f = .34 < 3.29 = F_{.10,3,6}$, there is no significant difference among the pen means at the .10 level.

With $f = 1.07 < 3.46 = F_{.10,2,6}$, there is no significant difference among the surface means at the .10 level.

69.

Source	df	Adj SS	Adj MS	f	P-value
Distance	2	562424	281212	360.70	0.000
Temperature	2	11757	5879	7.54	0.004
Distance * Temperature	4	21715	5429	6.96	0.001
Error	18	14033	780		
Total	26	609930			

The ANOVA table indicates a highly statistically significant interaction effect ($f = 6.96$, P-value = .001). The interaction by itself indicates that both nozzle-bed distance and temperature play a significant role in determining strut width. Apply Tukey's method here to the nine (distance, temperature) pairs to identify honestly significant differences.

Distance*Temperature	N	Mean	Grouping		
0.2 220	3	935.000	A		
0.2 180	3	860.000	A	B	
0.2 200	3	806.667		B	
0.3 200	3	676.667			C
0.3 220	3	643.333			C
0.3 180	3	610.000			C D
0.4 220	3	538.333			D E
0.4 200	3	511.667			E
0.4 180	3	505.000			E

73. a. MSAB/MSE b. MSA/MSAB, MSB/MSAB

75.

Source	df	SS	MS	f
Treatment	3	81.1944	27.0648	22.36
Block	8	66.5000	8.3125	6.87
Error	24	29.0556	1.2106	
Total	35	176.7500		

Since $22.36 > F_{.05,3,24} = 3.01$, reject H_{0A}. There is an effect due to treatments. Next, $Q_{.05,4,24} = 3.90$, so Tukey's HSD is $3.90\sqrt{1.2106/9} = 1.43$.

1	4	3	2
8.56	9.22	10.78	12.44

77. a. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs. H_a: at least two of the μ_i's are different; $f = 3.68 < F_{.01,3,20} = 4.94$, thus fail to reject H_0. The means do not appear to differ.

b. We reject H_0 when the P-value $\le \alpha$. Since $.029 > .01$, we still fail to reject H_0.

79. SSTr = 6.172, SSE = 1045.75, $f = 3.086/18.674 = 0.165 < F_{.05,2,56} = 3.16$, so do not reject H_0.

81.

Source	DF	SS	MS	F
Diet	4	.929	.232	2.15
Error	25	2.690	.108	
Total	29	3.619		

Because $f = 2.15 < 2.76 = F_{.05,4,25}$, there is no significant difference among the diet means at the .05 level.

b. $(-.144, .474)$ Yes, the interval includes 0.

c. .53

83. a. SSTr = 19812.6, SSE = 1083126, $f = 9906.3/7125.8 = 1.30 < F_{.05,2,152} = 3.06$, and H_0 is not rejected.

b. $Q_{.05,3,152} \approx 3.347$ and $d_{ij} = 3.347\sqrt{\frac{7125.8}{2}\left(\frac{1}{J_i} + \frac{1}{J_j}\right)}$ for each pair. Then $d_{12} = 42.1$, $d_{13} = 37.2$, and $d_{23} = 40.7$. None of the sample means are nearly this far apart, so Tukey's method provides no statistically significant differences. This is consistent with the results in part (a).

Chapter 12

1. a. Both the BMI and peak foot pressure distributions appear positively skewed with some gaps and possible high outliers.

Stem-and-leaf of BMI			Stem-and-leaf of Foot pressure		
1	12	8	7	3	0012344
6	13	00588	16	3	566666678
13	14	2456689	18	4	11
19	15	000569	(8)	4	56789999
21	16	69	16	5	34
21	17	01156	14	5	577778
16	18	677	8	6	024
13	19		5	6	6
13	20	0156	4	7	4
9	21	0126	3	7	
5	22	4	3	8	1
4	23	1	2	8	59
3	24	27			
1	25				
1	26	5			
Leaf Unit = 0.1			*Leaf Unit = 10*		

    b. No

    c. The scatterplot suggests some positive association between BMI and peak foot pressure, but the relationship does not appear to be very strong, and there are many outliers from the overall pattern.

3. Yes.

5. b. Yes

    c. The relationship of y to x is roughly quadratic.

7. a. 48.75 mpg
    b. −.0085 mpg
    c. −4.25 mpg
    d. 4.25 mpg

9. a. .095 m^3/min   b. −.475 m^3/min
    c. .83 m^3/min, 1.305 m^3/min
    d. .4207, .3446   e. .0036

11. a. −.01 h, −.10 h b. 3.0 h, 2.5 h c. .3653
    d. .4624

13. a. $y = .63 + .652x$
    b. 23.46, -2.46

    c. 392, 5.72
    d. 95.6%
    e. $y = 2.29 + .564x$, $R^2 = 68.8\%$

15. a. $y = -14.6497 + .09092x$
    b. 1.8997
    c. −.9977, −.0877, .0423, .7823
    d. 42%

17. a. Yes
    b. slope, .827; intercept, −1.13
    c. 40.22
    d. 5.24
    e. 97.5%

19. a. $y = 75.212 - .20939x$, 54.274
    b. 79.1%
    c. 2.56

21. b. $y = -0.398 + 3.080x$
    c. A 1-cm increase in palprebal fissure width corresponds to an estimated 3.080 cm^2 increase in average/expected OSA.
    d. 3.452 cm^2
    e. 3.452 cm^2

25. new slope $= 1.8\hat{\beta}_1$, new intercept $= 1.8\hat{\beta}_0 + 32$

29. $\hat{\beta}_0^* = \overline{Y}$ and $\hat{\beta}_1^* = \hat{\beta}_1$

31. a. .0756
    b. .813
    c. The $n = 7$ sample is preferable (larger S_{xx}).

33. $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$, $t = 22.64$, P-value ≈ 0, so there is a useful linear relationship.
    CI = (.748, .906)

35. a. $\hat{\beta}_1 = 1.536$, and a 95% CI is (.632, 2.440)
    b. Yes, for the test of H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, we find $t = 3.62$, with P-value .0025. At the .01 level conclude that there is a useful linear relationship.
    c. Because 5 is beyond the range of the data, predicting at a dose of 5 might involve too much extrapolation.
    d. The observation does not seem to be exerting undue influence.

37. a. Yes, for the test of H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, we find $t = -6.73$, with

P-value $< 10^{-9}$. At the .01 level conclude that there is a useful linear relationship.

b. $(-2.77, -1.42)$

43. a. 600 is closer to $\bar{x} = 613.5$ than is 750
    b. $(2.258, 3.188)$
    c. $(1.336, 4.110)$
    d. at least 90%

45. a. $y = -1.5846 + 2.58494x$, 83.73%
    b. $(2.16, 3.01)$
    c. $(-0.125, 0.058)$
    d. $(-0.559, 0.491)$
    e. H_0: $\mu_{Y|.7} = 0$ versus H_a: $\mu_{Y|.7} \neq 0$; reject H_0 because 0 is not in the confidence interval $(0.125, 0.325)$ for $\mu_{Y|.7}$

47. $(86.3, 123.5)$

49. a. $t = 4.88$, P-value ≈ 0, so reject H_0. A useful relationship exists.
    b. $(64.2, 161.3)$
    c. $(10228, 11362)$
    d. $(8215, 13379)$
    e. Wider, because 85 is farther from the mean x-value of 68.65 than is 70.
    f. No, extrapolation
    g. $(8707, \quad 10633); \quad (10020, \quad 11574);$ $(10845, 13004)$

51. a. Yes
    b. $t = 6.45$, P-value $= .003$, reject H_0. A useful relationship exists.
    c. $(.05191, .11544)$
    d. $(.00048, .16687)$

55. a. For the test of H_0: $\rho = 0$ versus H_a: $\rho > 0$, we find $r = .7482$, $t = 3.91$, with P-value $< .05$. At the .05 level conclude that there is a positive correlation.
    b. $R^2 = .56$; it is the same no matter which variable is the predictor.

57. a. $t = 1.74 < 2.179$, so do not reject H_0: $\rho = 0$.
    b. $R^2 = 20\%$

59. a. $(.829, .914)$
    b. $z = 2.412$, P-value $= .008$, so reject H_0: $\rho = 0$

c. $R^2 = 77.1\%$
d. Still 77.1%

61. a. Reject the null hypothesis in favor of the alternative.
    b. No, with a large sample size a small r can be significant.
    c. Because $t = 2.200 > 1.96 = t_{.025,9998}$ the correlation is statistically (but not necessarily practically) significant at the .05 level.

65. a. $.184, -.238, -.426$
    b. The mean that is subtracted is not the mean $\bar{x}_{1,n-1}$ of $x_1, x_2, \ldots, x_{n-1}$, or the mean $\bar{x}_{2,n}$ of $x_2, x_3, \ldots, x_n$. Also, the denominator of r_1 is not $\sqrt{\sum_1^{n-1} (x_i - \bar{x}_{1,n-1})^2}\sqrt{\sum_2^n (x_i - \bar{x}_{2,n})^2}$. However, if n is large then r_1 is approximately the same as the correlation. A similar relationship applies to r_2.
    c. No
    d. After performing one test at the .05 level, doing more tests raises the probability of at least one type I error to more than .05.

67. The plot suggests that the regression model assumptions of linearity/model adequacy and constant error variance are both plausible.

69. a. The plot does not show curvature, but equal variance is not satisfied.
    b. The standardized residual plot is similar to (a). The normality plot suggests normality of the true errors is plausible.

71. a. For testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, $t = 10.97$, with P-value .0004. At the .001 level conclude that there is a useful linear relationship.
    b. The residual plot shows curvature, so the linear relationship of part (a) is questionable.
    c. There are no extreme standardized residuals, and the plot of standardized residuals is similar to the plot of ordinary residuals.

73. a. The plot indicates there are no outliers, but there appears to be higher variance for middle values of filtration rate.
   b. e_i/e_i^*'s range between .57 and .65, which are close to s_e.
   c. Similar to the plot in (a).

75. The first data set seems appropriate for a straight-line model. The second data set shows a quadratic relationship, so the straight-line relationship is inappropriate. The third data set is linear except for an outlier, and removal of the outlier will allow a line to be fitted. The fourth data set has only two values of x, so there is no way to tell if the relationship is linear.

77. a. $24,000
   b. $16,300

79. b. 9.193
   c. $f = 82.75$, P-value ≈ 0, so at least one of the four predictors is useful, but not necessarily all four.
   d. Compare each to $t_{.00125,95} = 3.106$. Predictors 1, 2, and 4 are useful.

81. a. $y = -77 + 4.397x_1 + 165x_2$ c. $2,781,500

83. a. $R^2 = 34.05\%$, $s_e = 0.967$
   b. $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ versus H_a: not all three β's are 0; $f = 2.065 < 3.49$, so do not reject H_0 at the .05 level
   c. Yes

85. a. $y = 148 - 133x_1 + 128.5x_2 + 0.0351x_3$
   b. For x_1: $t = -0.26$, $P = .798$. For x_2: $t = 9.43$, $P < .0001$. For x_3: $t = 1.42$, $P = .171$. Only x_2 is a statistically significant predictor of y.
   c. $R^2 = 86.33\%$, $R_a^2 = 84.17\%$
   d. $R^2 = 86.28\%$, $R_a^2 = 84.91\%$

87. a. $f = 87.6$, P-value ≈ 0, strongly reject H_0
   b. $R_a^2 = 93.5\%$
   c. (9.095, 11.087)

89. a. Always decreases
   b. 61.432 GPa
   c. $f = 227.88 > 5.568$, so H_0 is rejected.

d. (55.458, 67.406)
e. (53.717, 69.147)

91. a. Both plots exhibit curvature.
   b. No
   c. The plots suggest that all model assumptions are satisfied.
   d. All second-order terms should be retained.

93. a. H_0: $\beta_1 = \beta_2 = 0$ versus H_a: not both β_1 and β_2 are 0, $f = 22.91 \geq F_{.05,2,9} = 4.26$, so reject H_0. Yes, there is a useful relationship.
   b. H_0: $\beta_2 = 0$ versus H_a: $\beta_2 \neq 0$, $t = 4.01$, P-value $< .005$, reject H_0. Yes.
   c. (.5443, 1.9557)
   d. (2.91, 6.29)

95. a. The quadratic terms are important in providing a good fit to the data.
   b. A 95% PI is (.560, .771).

97. a. $r_{RI} = .843$ (P-value $= .000$), $r_{RA} = .621$ (.001), $r_{IA} = .843$ (.000)
   b. Rating $= 2.24 + 0.0419$ IBU $- 0.166$ ABV. Because the predictors are highly correlated, one is redundant.
   c. Linearity is an issue.
   e. The regression is quite effective, with $R^2 = 87.2\%$. The ABV coefficient is not significant, so ABV is not needed. The highly significant positive coefficient for IBU and negative coefficient for its square show that Rating increases with IBU, but the rate of increase is lower at higher IBU.

99. a. $\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix}$ $\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 4 \end{bmatrix}$,

$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$ $\mathbf{b} = \begin{bmatrix} 6 \\ 2 \\ 4 \end{bmatrix}$

b. $\mathbf{b} = \begin{bmatrix} 1.5 \\ .5 \\ 1 \end{bmatrix}$

c. $\hat{\mathbf{y}} = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 3 \end{bmatrix}$ $\mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}$

SSE = 4, MSE = 4

d. $(-12.2, 13.2)$

e. For the test of H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, we find $|t| = .5 < t_{.025,1} = 12.7$, so do not reject H_0 at the .05 level. The x_1 term does not play a significant role.

f.
Source	DF	SS	MS	F
Regression	2	5	2.5	0.625
Error	1	4	4.0	
Total	3	9		

With $f = .625 < 199.5 = F_{.05,2,1}$, there is no significant relationship at the .05 level.

101. a. $\mathbf{X'X} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$,

$(\mathbf{X'X})^{-1} = \frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix}$

b. $\mathbf{X'y} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$,

$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \bar{y} - (S_{xy}/S_{xx})\bar{x} \\ S_{xy}/S_{xx} \end{bmatrix}$

103. $\hat{\beta}_0 = \bar{y}$, $s_e = \sqrt{\sum (y - \bar{y})^2/(n-1)}$, $\bar{y} \pm t_{.025,n-1}s/\sqrt{n}$

105. a. $\hat{\beta}_0 = \frac{1}{m+n}\sum_1^{m+n} y_i = \bar{y}$, $\hat{\beta}_1 = \frac{1}{m}\sum_1^m y_i - \frac{1}{n}\sum_{m+1}^{m+n} y_i = \bar{y}_1 - \bar{y}_2$

b. $\hat{\mathbf{y}} = [\bar{y}_1 \ldots \bar{y}_1 \bar{y}_2 \ldots \bar{y}_2]'$, SSE $= \sum_1^m (y_i - \bar{y}_1)^2 + \sum_{m+1}^n (y_i - \bar{y}_2)^2$, $s_e = \sqrt{\text{SSE}/(m+n-2)}$, $s_{\hat{\beta}_1} = s_e\sqrt{1/m + 1/n}$

d. $\hat{\beta}_0 = 128.166$, $\hat{\beta}_1 = -14.333$, $\hat{\mathbf{y}} = [121, 121, 121, 135.33, 135.33, 135.33]'$, SSE $= 116.666$, $s_e = 5.4$, 95% CI for β_1 $(-26.58, -2.09)$

109. a. With $f = 12.04 > 9.55 = F_{.01,2,7}$, there is a significant relationship at the .01 level. To test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, we find $|t| = 2.96 > t_{.025,7} = 2.36$, so reject

H_0 at the .05 level. The foot term is needed. To test H_0: $\beta_2 = 0$ versus H_a: $\beta_2 \neq 0$, we find $|t| = 0.02 < t_{.025,7} = 2.36$, so do not reject H_0 at the .05 level. The height term is not needed.

b. The highest leverage is .88 for the fifth point. The height for this student is given as 54 inches, too low to be correct for this group of students. Also this value differs by 8" from the wingspan, an extreme difference.

c. Point 1 has leverage .55, and this student has height 75, foot length 13, both quite high.

Point 2 has leverage .31, and this student has height 66 and foot length 8.5, at the low end.

Point 7 has leverage .31 and this student has both height and foot length at the high end.

d. Point 2 has the most extreme residual. This student has a height of 66" and a wingspan of 56" differing by 10", so the extremely low wingspan is probably wrong.

e. For this data set it would make sense to eliminate points 2 and 5 because they seem to be wrong. However, outliers are not always mistakes and one needs to be careful about eliminating them.

111. a. $p(10) = .060$, $p(50) = .777$
b. $odds(10) = .0639$, $odds(50) = 3.49$
d. \$37.50

113. a. H_0: $\beta_1 = 0$ versus H_0: $\beta_1 \neq 0$, $z = -2.026$, H_0 is rejected at $\alpha = .05$
b. $(.675, .993)$

115. a. $\hat{\beta}_0 = -.0573$ and $\hat{\beta}_1 = .00430$
c. H_0: $\beta_1 = 0$ versus H_0: $\beta_1 \neq 0$, $z = 0.74$, H_0 is not rejected.

117. a. .912
b. .794

119. c. .484
d. $z_1 = -0.38$, $z_2 = 1.75$, so do not reject H_0: $\beta_1 = 0$ but do reject H_0: $\beta_2 = 0$ in favor of H_a: $\beta_2 \neq 0$.

121. a. Flood damage increases with flood level, but there are two "jumps" at 2–3 ft and 5–6 ft.
   b. No

123. a. 50.73% b. .7122
   c. To test H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, we have $t = 3.93$, with P-value .0013. At the .01 level conclude that there is a useful linear relationship.
   d. (1.056, 1.275)
   e. $\hat{y} = 1.014$, $y - \hat{y} = -.214$

125. No, if the relationship of y to x is linear, then the relationship of y^2 to x is quadratic.

127. a. Yes
   b. $\hat{y} = 98.293$, $y - \hat{y} = .117$
   c. $s_e = .155$
   d. $R^2 = .794$
   e. 95% CI for β_1: (.0613, .0901)
   f. The new observation is an outlier, and has a major impact:
      The equation of the line changes from $y = 97.50 + .0757x$ to $y = 97.28 + .1603x$
      s_e changes from .155 to .291
      R^2 changes from .794 to .616

129. a. The paired t procedure gives $t = 3.54$ with a two-tailed P-value of .002, so at the .01 level we reject the hypothesis of equal means.
   b. The regression line is $y = 4.79 + .743x$, and the test of H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$, gives $t = 7.41$ with a P-value of <.000001, so there is a significant relationship. However, prediction is not perfect, with $R^2 = .753$, so one variable accounts for only 75% if the variability in the other.

133. a. linear
   b. After fitting a line to the data, the residuals show a lot of curvature.
   c. Yes, $\ln(y) = 3.1564 + 0.004811x$, $\hat{\alpha} = 23.486$, $\hat{\beta} = 0.004811$
   d. (54.42, 112.36)

135. a. A linear relationship is plausible.
   b. $y = 31.04 - 5.79x$; model utility $t = -4.25$, P-value ≈ 0, so pH is a statistically useful predictor of mean crown dieback.
   c. PI = (1.42, 14.33), CI = (6.42, 9.32)
   d. PI = (4.69, 18.00), CI = (9.18, 13.52)

137. a. $y = 84.82 + .1643x_1 - 79.67x_2$ and $R_a^2 = .654$
   b. $R_a^2 = .831$ with interaction, .7207 for full second-order model. The model with an interaction term but without quadratic terms is preferred.
   c. $y = 6.22 + 5.779x_1 + 51.33x_2 - 9.357x_1x_2$, 39.32 MPa
   d. First-order: $R_a^2 = 66.22\%$; with interaction, $R_a^2 = 68.27\%$; full second-order: $R_a^2 = 70.42\%$. These suggest that the full second-order model is "best" for predicting adsorbability.

Chapter 13

1. a. reject H_0 b. do not reject H_0
   c. do not reject H_0 d. do not reject H_0

3. Do not reject H_0 because $\chi^2 = 1.57 < 7.815 = \chi^2_{.05,3}$.

5. Because $\chi^2 = 6.61$ with P-value .68, do not reject H_0.

7. Do not reject H_0 because $\chi^2 = 4.41 < 7.779 = \chi^2_{.10,5-1}$.

9. a. [0, .223), [.223, .510), [.510, .916), [.916, 1.609). [1.609, ∞)
   b. Because $\chi^2 = 1.25$ with P-value >.10, do not reject H_0.

11. a. $(-\infty, -.967)$, $[-.967, -.431)$, $[-.431, 0)$, $[0, .431)$, $[.431, .967)$, $[.967, \infty)$
   b. $(-\infty, .49806)$, $[.49806, .49914)$, $[.49914, .50)$, $[.50, .50086)$, $[.50086, .50194)$, $[.50194, \infty)$
   c. Because $\chi^2 = 5.53$ with P-value >.10, do not reject H_0.

13. a. With $\theta = P(\text{male})$, $\hat{\theta} = .504$, $\chi^2 = 3.45$, df = $4 - 1 - 1 = 2$. Do not reject H_0 because $\chi^2 \geq \chi^2_{.05,2} = 5.992$.
   b. No, because the expected count for the last category is too small.

15. $\hat{\mu} = 3.167$ which gives $\chi^2 = 103.9$ with P-value $< .001$, so reject the assumption of a Poisson model.

17. The observed test statistic value is $\chi^2 = 6.668 < 10.645$ (df = $9 - 1 - 2 = 6$), so H_0 is not rejected at the .10 level.

19. $\chi^2 = 2.788 \approx z^2$; P-values are the same. Reject H_0 at the .10 level but not at the .05 level.

21. a. $\chi^2 = 4.504 < \chi^2_{.05,4} = 9.488$, so H_0 is not rejected.

23. $\chi^2 = 9.858 < \chi^2_{.05,6} = 12.592$, so H_0 is not rejected at the .05 level.

25. a. Reject H_0 because $\chi^2 = 11.954$ at 2 df and P-value $= .003$.
   b. Very large sample sizes make the test capable of detecting even slight deviations from H_0.

27. $\hat{e}_{ijk} = n_{i..} \cdot n_{.j.} \cdot n_{..k}/n^2$; df = $IJK - (I + J + K) + 2 = 28$.

29. a. Because $.6806 < \chi^2_{.10,2} = 4.605$, H_0 is not rejected.
   b. Now $\chi^2 = 6.806 \geq 4.605$, and H_0 is rejected.
   c. 677

31. a. With $\chi^2 = 6.45$ and P-value .040, reject independence at the .05 level.
   b. With $z = -2.29$ and P-value .022, reject independence at the .05 level.
   c. Because the logistic regression takes into account the order in the professorial ranks, it should be more sensitive, so it should give a lower P-value.
   d. There are few female professors but many assistant professors, and the assistant professors will be the professors of the future.

33. $\chi^2 = 5.934$, df = 2, P-value = .059. So, at the .05 level, we (barely) fail to reject H_0.

35. $\chi^2 = 29.775 \geq \chi^2_{.05,(4-1)(3-1)} = 12.592$. So, H_0 is rejected at the .05 level.

37. a. H_0: The population proportion of Late Game Leader Wins is the same for all four sports; H_a: The proportion of Late Game Leader Wins is not the same for all four sports. With $\chi^2 = 10.518 > 7.815 = \chi^2_{.05,3}$, reject the null hypothesis at the .05 level. Sports differ in terms of coming from behind late in the game.
   b. Yes (baseball)

39. $\chi^2 = 881.36$, df = 16, P-value is effectively zero. USA respondents were more amenable to torture than the Europeans, while South Korean respondents were vastly more likely than anyone else to say it's "sometimes" okay to torture terror suspects.

41. a. No, $\chi^2 = 9.02 > 7.815 = \chi^2_{.05,3}$.
   b. With $\chi^2 = .157 < 6.251 = \chi^2_{.10,3}$, there is no reason to say the model does not fit.

43. a. $H_0: p_0 = p_1 = \cdots = p_9 = .10$ versus H_a: at least one $p_i \neq .10$, with df = 9.
   b. $H_0: p_{ij} = .01$ for i and $j = 0, 1, 2, \ldots, 9$ versus H_a: at least one $p_{ij} \neq .01$, with df = 99.
   c. No, there must be more observations than cells to do a valid chi-square test.
   d. The results give no reason to reject randomness.

Chapter 14

1. $(y_4, y_{15}) = (\$55,000, \$61,000)$

3. (Y_{14}, Y_{27})

5. P-value = $1 - B(18; 25, .5) = .007$, so reject H_0 at the .05 level.

7. P-value $= 1 - B(16; 24, .5) = .032$, so reject H_0 at the .05 level.

9. Assuming distribution of differences is symmetric, let p = the true proportion of individuals who would perceive a longer time for the shorter exam (positive difference) in this experiment. Hypotheses are equivalent to H_0: $p = .5$ versus H_a: $p > .5$. P-value ≈ 0, so H_0 is strongly rejected.

11. $s_+ = 27$, and since 27 is neither ≥ 64 nor ≤ 14, we do not reject H_0.

13. $s_+ = 22 < 24$, so H_0 is not rejected at the .05 level.

15. Test H_0: $\mu_D = 0$ versus H_0: $\mu_D \neq 0$. $s_+ = 72 \geq 64$, so H_0 is rejected at level .05.

17. a. Test H_0: $\mu_D = 0$ versus H_a: $\mu_D < 0$. $s_+ = 2 \leq 6$, and so H_0 is rejected at the .055 level.
    b. Test H_0: $\mu_D = 0$ versus H_a: $\mu_D > 0$. Because $s_+ = 28 < 30$, H_0 cannot be rejected at this level.

19. a. Assume that the population distribution of differences is at least symmetric. With $s_+ = 3 \leq 6$, H_0 is rejected. P-value $= .021$.
    b. Assume that the population distribution of differences is normal. $t = -2.54$, df $= 7$, P-value $= .019$, so H_0 is rejected.
    c. The P-values were .035 (sign test), .021 (Wilcoxon signed-rank test), and .019 (paired t test). As is typical, the P-value decreases with more powerful tests. But, all three tests agree that H_0 is rejected at the .05 level, and the sign test has the fewest assumptions.

21. (7.22, 7.73)

23. (−.1745, −.0110)

25. With $w = 38$, reject H_0 at the .05 level because the rejection region is $\{w \geq 36\}$.

27. Test H_0: $\mu_1 - \mu_2 = 1$ versus H_a: $\mu_1 - \mu_2 > 1$. After subtracting 1 from the original process measurements, we get $w = 65$. Do not reject H_0 because $w < 84$.

29. b. Test H_0: $\mu_1 - \mu_2 = 0$ vs H_a: $\mu_1 - \mu_2 < 0$. With a P-value of .0027 we reject H_0 at the .01 level.

31. Test H_0: $\mu_1 - \mu_2 = 0$ vs H_a: $\mu_1 - \mu_2 > 0$. W has mean $m(m + n + 1)/2 = 59.5$ and variance $mn(m + n + 1)/12 = 89.25$. $z = 2.33$, P-value $= .01$, so H_0 is rejected at the .05 level.

33. Pain: $z = -1.40$, P-value $= .0808$. Depression: $z = -2.93$, P-value $= .0017$. Anxiety: $z = -4.32$, P-value $< .0001$. Fail to reject first H_0, reject last two. Chance of at least one type I error is no more than .03.

35. (16, 87)

37. $h = 21.43$, df $= 3$, P-value $< .0001$, so H_0 is strongly rejected.

39. $h = 9.85$, df $= 2$, P-value $= .007 < .01$, so H_0 is rejected.

43. a. Rank averages of the three positions/ rows are $\bar{r}_{1.} = 12/6 = 2, \bar{r}_{2.} = 13/6 = 2.1\bar{6}, \bar{r}_{3.} = 11/6 = 1.8\bar{3}$; $Fr = 0.333$, df $= 2$, P-value $\approx .85$, so H_0 is certainly not rejected.
    b. $Fr = 6.34$, df $= 2$, P-value $= .042$, so reject H_0 at the .05 level.
    c. $Fr = 1.85$, df $= 2$, P-value $= .40$, so H_0 is not rejected.

45. H_0: $\mu_1 = \cdots = \mu_{10}$ versus H_a: not all μ_i's are equal. $Fr = 78.67$, df $= 9$, P-value ≈ 0, so H_0 is resoundingly rejected. The four algorithms inspired by quantum computing (Q's in the name) have much lower rank means, suggesting they are far better at minimizing entropy.

49. Test H_0: $\mu_1 - \mu_2 = 0$ vs H_a: $\mu_1 - \mu_2 > 0$. Rank sum for first sample $= 26$, P-value $= .014$. Reject H_0 at .05 but not .01.

51. mean = 637.5, variance = 10731.25

    a. $z = -0.21$, P-value = .83. The data do not contradict a claim the sensory test is reliable.

    b. $z = 1.70$, P-value = .09. Reject H_0 at .10 level. This might indicate a lack of reliability of the sensory test for the population of healthy patients.

53. Test H_0: $\mu_1 = \cdots = \mu_5$ versus H_a: not all μ_i's are equal. $h = 20.21 \geq 13.277$, so H_0 is rejected at the .01 level.

55. μ_i = population mean skin potential (mV) with the ith emotion ($i = 1$ for fear, etc.). The hypotheses are H_0: $\mu_1 = \cdots = \mu_4$ versus H_a: not all μ_i's are equal. $Fr = 6.45 < \chi^2_{.05,3} = 7.815$, so we fail to reject H_0 at the .05 level.

57. Because $w' = 26 < 27$, do not reject the null hypothesis at the 5% level.

Chapter 15

1. a. $\pi(.50) = .80$ and $\pi(.75) = .20$

    b. $\pi(.50|\text{HHHTH}) = .6124$ and $\pi(.75|\text{HHHTH}) = .3876$

3. a. Gamma(9, 5/3)

    b. Gamma(145, 5/53)

5. $\alpha_1 = \alpha_0 + \Sigma x_i$, $\beta_1 = 1/(n + 1/\beta_0)$

7. $\alpha_1 = \alpha_0 + nr$, $\beta_1 = \beta_0 + \Sigma x_i - nr$

9. Normal, $\mu_1 = \dfrac{\tau_0 \mu_0 + \Sigma \ln x_i}{\tau_0 + n}$, $\tau_1 = \tau_0 + n$

11. a. 13.68

    b. (11.54, 15.99)

15. Normal, mean = 116.77, variance = 10.227, same as previous

17. (.485, .535)

19. a. $\alpha_0 \beta_0$ b. $\dfrac{\alpha_0 + \Sigma X_i}{n + 1/\beta_0}$

References

Overview and Descriptive Statistics

Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, *Graphical Methods for Data Analysis*, Brooks/Cole, Pacific Grove, CA, 1983. A highly recommended presentation of graphical and pictorial methodology in statistics.

Cleveland, William S., *The Elements of Graphing Data* (2nd ed.), Hobart Press, Summit, NJ, 1994. A very nice survey of graphical methods for displaying and summarizing data.

Freedman, David, Robert Pisani, and Roger Purves, *Statistics* (4th ed.), Norton, New York, 2007. An excellent, very nonmathematical survey of basic statistical reasoning and concepts.

Hoaglin, David, Frederick Mosteller, and John Tukey, *Understanding Robust and Exploratory Data Analysis*, Wiley-Interscience, New York, 2000. Discusses why, as well as how, exploratory methods should be employed; it is good on details of stem-and-leaf displays and boxplots.

Moore, David S. and William I. Notz, *Statistics: Concepts and Controversies* (9th ed.), Freeman, San Francisco, 2016. An extremely readable and entertaining paperback that contains an intuitive discussion of problems connected with sampling and designed experiments.

Peck, Roxy, et al. (eds.), *Statistics: A Guide to the Unknown* (4th ed.), Thomson-Brooks/Cole, Belmont, CA, 2005. Contains many short, nontechnical articles describing various applications of statistics.

Utts, Jessica, *Seeing Through Statistics* (4th ed), Cengage Learning, Boston, 2014. The focus is on statistical literacy and critical thinking; a wonderful exposition.

Probability and Probability Distributions

Balakrishnan, N., Norman L. Johnson, and Samuel Kotz, *Continuous Univariate Distributions, Vol 1* (3rd ed.), Wiley, Hoboken, NJ, 2016. Encyclopedic, not for bedtime reading.

Carlton, Matthew A. and Jay L. Devore, *Probability with STEM Applications,* Wiley, Hoboken, NJ, 2021. An expansion of the material in Chapters 2â€"6 of this book (*Modern Mathematical Statistics with Applications*) plus additional material.

Gorroochurn, Prakash, *Classic Problems of Probability*, Wiley, Hoboken, NJ, 2012. An entertaining excursion through 33 famous probability problems.

Johnson, Norman L, Adrienne W. Kemp, and Samuel Kotz, *Univariate Discrete Distributions* (3rd ed.), Wiley, Hoboken, NJ, 2005. Encyclopedic, not for bedtime reading.

Olofsson, Peter, *Probabilities*: *The Little Numbers That Rule Our Lives* (2nd ed.), Wiley, Hoboken, NJ, 2015. A very non-technical and thoroughly charming introduction to the quantitative assessment of uncertainty.

Ross, Sheldon, *A First Course in Probability* (10th ed.), Prentice Hall, Upper Saddle River, NJ, 2018. Rather tightly written and more mathematically sophisticated than this text but contains a wealth of interesting examples and exercises.

Ross, Sheldon, *Introduction to Probability Models* (12th ed.), Academic Press, Cambridge, MA, 2019. Another tightly written exposition of somewhat more advanced topics, again with a wealth of interesting examples and exercises.

Winkler, Robert, *Introduction to Bayesian Inference and Decision* (2nd ed.), Probabilistic Publishing, Sugar Land, Texas, 2003. A very good introduction to subjective probability.

Basic Inferential Statistics

Casella, George and Roger L. Berger, *Statistical Inference* (2nd ed.), Cengage Learning, Boston, 2001. The focus is on the theory of mathematical statistics; exposition is appropriate for advanced undergraduates and MS. level students. Hopefully there will be a new edition soon.

Daniel, Cuthbert and Fred S. Wood, *Fitting Equations to Data: Computer Analysis of Multifactor Data* (2nd ed.), Wiley-Interscience, New York, 1999. Contains many insights and methods that evolved from the authors' extensive consulting experience.

DeGroot, Morris, and Mark Schervish, *Probability and Statistics* (4th ed.), Pearson, Englewood Cliffs, NJ, 2011. Includes an excellent discussion of both general properties and methods of point estimation; of particular interest are examples showing how general principles and methods can yield unsatisfactory estimators in particular situations.

Davison, A.C. and D.V. Hinkley, *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, UK, 1997. General principles and methods interspersed with many examples.

Efron, Bradley, and Robert Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993. The first general accessible exposition, and still informative and authoritative.

Good, Philip, *A Practitioner's Guide to Resampling for Data Analysis, Data Mining, and Modeling*, Chapman and Hall, New York, 2019. Obviously brand new, and hopefully informative about recent bootstrap methodology.

Meeker, William Q., Gerald J. Hahn, and Luis A. Escobar, *Statistical Intervals*: A Guide for Practitioners and Researchers (2nd ed.), Wiley, Hoboken, NJ, 2016. Everything you ever wanted to know about statistical intervals (confidence, prediction, tolerance, and others).

Rice, John, *Mathematical Statistics and Data Analysis* (3rd ed.), Cengage Learning, Boston, 2013. A nice blending of statistical theory and data analysis.

Wilcox, Rand, *Introduction to Robust Estimation and Hypothesis Testing* (4th ed.), Academic Press, Cambridge, MA, 2016. Presents alternatives to the inferential methods based on t and F distributions.

Specific Topics in Inferential Statistics

Agresti, Alan, *An Introduction to Categorical Data Analysis* (3rd ed.), Wiley, New York, 2018. An excellent treatment of various aspects of categorical data analysis by one of the most prominent researchers in this area.

Chatterjee, Samprit and Ali Hadi, *Regression Analysis by Example* (5th ed.), Wiley, Hoboken, NJ, 2012. A relatively brief but informative discussion of selected topics.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, *Bayesian Data Analysis* (3rd ed.), Chapman and Hall, New York, 2013. A comprehensive survey of theoretical, practical, and computational issues in Bayesian inference; its authors have made many contributions to Bayesian methodology.

Hollander, Myles, and Douglas Wolfe, *Nonparametric Statistical Methods* (3rd ed.), Wiley, Hoboken, NJ, 2013. A very good reference on distribution-free (i.e. non-parametric) methods with an excellent collection of tables.

Lehmann, Erich, *Nonparametrics: Statistical Methods Based on Ranks* (revised ed.), Springer, New York, 2006. An excellent development of the most important "classical" (i.e. pre-bootstrap) methods, presented with a great deal of insightful commentary.

Miller, Rupert, *Beyond ANOVA: The Basics of Applied Statistics*, Wiley, Hoboken, NJ, 1986. An excellent source of information about assumption checking and alternative methods of analysis.

Montgomery, Douglas, *Design and Analysis of Experiments* (9th ed.), Wiley, New York, 2019. An up-to-date presentation of ANOVA models and methodology.

Kutner, Michael, Christopher Nachtsheim, John Neter, and William Li, *Applied Linear Statistical Models* (5th ed.), McGraw-Hill, New York, NY, 2005. The first 14 chapters constitute an extremely readable and informative survey of regression analysis. The second half of the book contains a well-presented survey of ANOVA. The level throughout is comparable to that of the present text (generally without proofs); the comprehensive discussion makes the book an excellent reference.

Ott, R. Lyman, and Michael Longnecker, *An Introduction to Statistical Methods and Data Analysis* (7th ed.), Cengage Learning, Boston, 2015. Includes several chapters on ANOVA and regression methodology that can profitably be read by students desiring a non-mathematical exposition; there is a good chapter on various multiple comparison methods.

Rencher, Alvin C. and William F. Christensen, *Methods of Multivariate Analysis* (3rd ed.), Wiley, Hoboken, NJ, 2012. Arguably the definitive text on multivariate analysis, including extensive information on the multivariate normal distribution.

Statistical Computing and Simulation

Datar, Radhika and Harish Garg, *Hands On Exploratory Data Analysis with R*, Packt Publishing, Birmingham, UK, 2019. Explains how the R software package can be used to explore various types of data.

Law, Averill M., *Simulation Modeling and Analysis* (5th ed.), McGraw-Hill, New York, 2014. An authoritative survey of various aspects and methods of simulation.

Index