Chapter 13
Analysis of Variance

Many experiments examine the effects of multiple experimental conditions. When each measured response from a subject is a single-number, the data are usually analyzed with *analysis of variance* (*ANOVA*). The name has a certain logic because, as we will see, the technique rests on a breakdown of sums of squares (assessing variation), but the null hypothesis typically takes the theoretical means to be equal among the experimental conditions, specifying no treatment effect, so that one may think of the methodology as an investigation of means. The general ideas developed in Chapters 10 and 11 carry over to ANOVA. One additional, very important notion involves the structure of the experiment. This is spelled out in Section 13.1. In Section 13.2 we indicate the way standard ANOVA models may be considered special cases of linear regression, as treated in Section 12.5. This is important conceptually and computationally. In Section 13.3 we take up nonparametric methods in ANOVA and in Section 13.4 we discuss causality and the role of randomization, which is especially relevant in clinical studies.

13.1 One-Way and Two-Way ANOVA

ANOVA can take many forms, depending on the design of the experiment and the resulting structure of the data. We consider here only the two simplest kinds of ANOVA and introduce them with a pair of examples.

Example 13.1 Stimulation and development of motor control Zelazo et al. (1972) conducted a study to see whether stimulation of infants during the first eight weeks of life could make them walk earlier. The stimulation involved a simulation of walking in which a parent held the baby in a manner that would make it respond reflexively with walking-type leg movements. The data in Table 13.1 are ages in months at which 24 infants were judged to begin walking.[1] Each 1-week-old infant

[1] For pedagogical simplicity, we wanted the number of subjects per group to be equal. This is not required for ANOVA; it merely makes things a bit easier to discuss. In the original data there were

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_13,

Table 13.1 Data from motor control experiment of Zelazo et al. (1972).

Active-exercise Group	Passive-exercise Group	No-exercise Group	8-Week control Group
9.00	11.00	11.50	13.25
9.50	10.00	12.00	11.50
9.75	10.00	9.00	12.00
10.00	11.75	11.50	13.50
13.00	10.50	13.25	11.50
9.50	15.00	13.00	12.35

Entries are ages at which each of 24 infants began walking. The treatment group is "active-exercise" and the other three groups served as controls

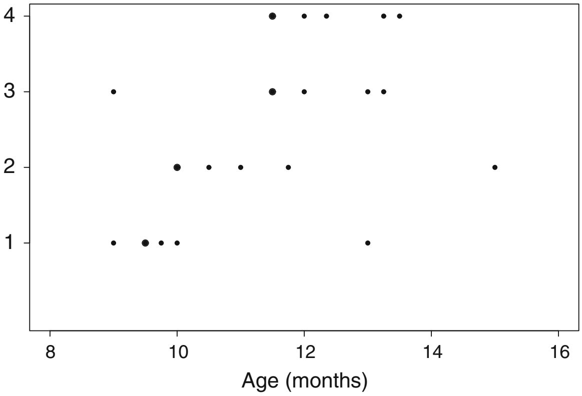

Fig. 13.1 Display of data from Table 13.1. The age of walking is shown for each of the *four* conditions, with *1* being active exercise, *2* being passive exercise, *3* being no exercise, and *4* being the 8-week control. Each larger plotted *dot* indicates the presence of 2 identical values of age within a given condition (so that for each condition there are 6 observations at 5 locations on the graph).

was assigned to one of four groups, namely, an experimental group (active-exercise) and three control groups (passive-exercise, no-exercise, 8-week control).[2] The issue is whether the active-exercise group walked earlier than the controls. From Fig. 13.1 it may be seen that the active-exercise group infants had somewhat earlier reported

(Footnote 1 continued)
only 5 subjects in the 8-week control group. We therefore added the 12.35 value to the 8-week control group.

[2] Infants in the active-exercise group received stimulation of the walking and placing reflexes during four 3-minute sessions that were held each day from the beginning of the second week until the end of the eighth week. The infants in the passive-exercise group received equal amounts of gross motor and social stimulation as those who received active-exercise, but unlike the active-exercise group, these infants had neither the walking nor placing reflex exercised. Infants in the no-exercise group did not receive any special training, but were tested along with the active-exercise and passive-exercise subjects. The 8-week control group was tested only when they were 8 weeks of age; this group served as a control for the possible helpful effects of repeated examination.

Table 13.2 Data from finger tapping experiment of Scott and Chen (1944).

Drug	Subject No.			
	1	2	3	4
Pl	11	56	15	6
Th	26	83	34	13
Ca	20	71	41	32

Entries are tapping rates. Each of 4 subjects received all 3 treatments (drugs): placebo, theobromine, and caffeine

ages of walking than those in the three control groups. However, there is quite a bit of variability, with one of the 6 infants in the active group being relatively late (13.0) and one in the no-exercise group being quite early (9.0). Thus, it's hard to tell whether there is a consistent pattern. □

Notice the layout of the data in the example above: it makes sense to display them in columns, with each column identified with a different treatment. The next example is different.

Example 13.2 Finger tapping in response to stimulants Scott and Chen (1944) conducted an experiment on finger tapping in response to orally-administered stimulants. Four subjects were each given three different treatments and then their finger-tapping rates were analyzed. The treatments were caffeine (Ca); 1-ethyltheobromine (Th: the stimulant in chocolate, similar to caffeine); and a placebo (Pl). The tapping rates (rate minus 440, with "rate" not defined but possibly taps per minute) are shown in Table 13.2.

In this case we would be interested in comparing the three treatments. The mean tapping rates for Pl, Th, and Ca are 22, 39, and 41. Is this evidence that theobromine and caffeine led to increased tapping rates? □

An important distinction between the two experiments above is that in the finger tapping experiment in Example 13.2 each subject received *all* of the treatments. Thus, the 12 data values were produced by only 4 subjects in the experiment, not 12. In the motor control experiment of Example 13.1, each subject received only one treatment, and the 24 data values came from 24 subjects. The two situations require related but different statistical methods. Table 13.1 is sometimes called a *one-way* table and is treated by *one-way ANOVA* while Table 13.2 is called a *two-way* table and is treated by *two-way ANOVA*.

13.1.1 ANOVA is based on a linear model.

The one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{13.1}$$

where Y_{ij} is the jth observation in the ith group, $\mu + \alpha_i$ is the mean for the ith group and ϵ_{ij} is the error for the jth observation in the ith group (the discrepancy between Y_{ij} and $\mu + \alpha_i$). Here, μ is the overall mean (the "grand mean") and α_i is the increment added to that overall mean in obtaining the mean for the ith group, so that

$$\frac{1}{I}\sum_{i=1}^{I} \mu + \alpha_i = \mu$$

and this implies

$$\sum_{i=1}^{I} \alpha_i = 0. \tag{13.2}$$

We take the number of groups to be I, so that $i = 1, 2, \ldots, I$, and write the number of observations in group i as n_i. In some places we also write the ith group mean as

$$\mu_i = \mu + \alpha_i.$$

The one-way ANOVA assumptions are

(i) the ANOVA model (13.1) holds;
(ii) the errors satisfy $E(\epsilon_i) = 0$ for all i;
(iii) the errors ϵ_i are independent of each other;
(iv) $V(\epsilon_i) = \sigma^2$ for all i (homogeneity of error variances), and
(v) $\epsilon_i \sim N(0, \sigma^2)$ (normality of the errors).

Note that these are the same assumptions as those used in linear regression (apart from the replacement of (12.5) with (13.1); see p. 315). As a result, residual analysis may be used in very much the same way as in regression. Indeed, mathematically, analysis of variance may be considered a special case of linear regression. We return to this in Section 13.2.

The purpose of this model is to provide a basis for statistical comparison of the group means $\mu + \alpha_i$. That is, we ask whether there is evidence that the means are different and, if so, we can estimate how different they are. Formally, we want to test the null hypothesis that the groups means are equal:

$$\mu + \alpha_1 = \mu + \alpha_2 = \cdots = \mu + \alpha_I.$$

The usual way the hypothesis is stated is as follows:

$$\mathrm{H}_0 : \alpha_i = 0 \tag{13.3}$$

for all i, which implies that the group means are equal. It also satisfies the condition that the grand mean μ remains the expectation of Y_{ij} under H_0.

13.1.2 One-way ANOVA decomposes total variability into average group variability and average individual variability, which would be roughly equal under the null hypothesis.

At the beginning of Section 12.5.2 we wrote the basic signal and noise decomposition for regression,

$$SST = SSR + SSE.$$

In ANOVA we decompose the variability in the data similarly into two pieces, replacing SSR with a treatment or "group" sum of squares SS_{group}. To test H_0 defined by (13.3) we compute a measure of the *average* amount of variability due to the groups, and an *average* amount of variability due to error, then compare these. Under the null hypothesis that the group means are equal, there should be no systematic variability due to groups, so that the variability we see in our "average variability due to groups" is the result of background variability in the measurements themselves, that is, the error variability. In other words, the average variability due to groups should be about the same size as the average variability due to error. Thus, to test H_0 we use a ratio of these measures of average variability and when the ratio is much larger than 1 there is evidence against H_0, in favor of there being differences among the groups. We first specify and illustrate the procedure and then indicate its motivation as a likelihood ratio test.

We begin with the total sum of squares

$$SST = \sum_{i,j}(y_{ij} - \bar{y}_{..})^2$$

where the double dots in the subscript on $y_{..}$ indicate that the mean is being taken over all the values of y, averaging across both rows and columns. In the infant exercise example we average across all 24 values. We also define the error (residual) sum of squares to be

$$SSE = \sum_{i,j}(y_{ij} - \bar{y}_{i.})^2$$

where the single dot in the subscript on $y_{i.}$ indicates that the mean is being taken *within* the ith group. In the infant exercise example there would be 4 means $\bar{y}_{i.}$ for $i = 1, 2, 3, 4$ and each would be an average across all 6 values in the appropriate column. The group sum of squares is then

$$SS_{group} = SST - SSE.$$

We next obtain averages of the group and error sums of squares by dividing by their respective degrees of freedom, df_{group} and df_{error}. Because of the constraint (13.2) we have $df_{group} = I - 1$ and, with n being the total number of observations, this leaves $n - 1 - (I - 1) = n - I$ degrees of freedom for error, i.e., $df_{error} = n - I$.

Table 13.3 Group means and standard deviations for the data in Example 13.1.

Group	N	Mean	St. Dev.
Active exercise	6	10.1	1.5
Passive exercise	6	11.3	1.9
No exercise	6	11.7	1.5
8-week control	6	12.35	.86

Table 13.4 Analysis of Variance table for data in Example 13.1.

Source	DF	SS	MS	F	*p*-value
Groups	3	15.74	5.25	2.40	0.098
Error	20	43.69	2.18		
Total	23	59.43			

The table lists each source of variability, the degrees of freedom for that source, and the sum of squares. For the groups and errors sources the mean squares (given by (13.4)) are also shown, and the F-statistic (given by (13.5)) and p-value are shown on the groups line

The resulting averages, called the *group mean square* and the *mean squared error*, are defined by

$$\begin{aligned} MS_{group} &= SS_{group}/df_{group} \\ MSE &= SSE/df_{error}. \end{aligned} \tag{13.4}$$

Finally, we obtain from these the F-ratio

$$F = MS_{group}/MSE. \tag{13.5}$$

Under the null hypothesis this ratio follows an F_{ν_1,ν_2} distribution, where $\nu_1 = df_{group}$ and $\nu_2 = df_{error}$ which is used to compute the p-value. Equations (13.4) and (13.5) should be compared with Eq. (12.49).

Note that in a certain sense “analysis of variance” is a misnomer. We are really analyzing several means, and determining whether there’s evidence that they are different. However, the basic tool for doing so is a comparison of sums of squares, that is, a comparison of different sources of variability, which explains the terminology.

Example 13.1 (continued from p. 361) The means and standard deviations for the 4 groups are shown in Table 13.3, and the basic ANOVA breakdown is given in Table 13.4. The pooled standard deviation is $s = \sqrt{2.18} = 1.48$. Because $F = 2.40$ on 3 and 20 d.f. with $p = .098$ there is no evidence of any differences among the means. Although from the sample means it may appear that the mean age of walking is somewhat smaller for the first group than those for the control groups, according to the ANOVA F-test there is enough variability in the data that any differences among the means are consistent with chance fluctuation. As we mentioned on p. 361, there are a couple of points visible in Fig. 13.1 that increase the variability and, thus, the denominator of the F-ratio. We will analyze these data further on p. 368. □

We now indicate how the F-test in (13.4) and (13.5) arises as a likelihood ratio test by considering the simpler ANOVA problem in which σ is known. Let us write the group means in the form $\mu_i = \mu + \alpha_i$. The pdf for observation y_{ij} is

$$f(y_{ij}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_{ij}-\mu_i)^2}{\sigma^2}}$$

and from the joint pdf

$$f(y_{11}, y_{12}, \dots, y_{In_I}) = \prod_{ij} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_{ij}-\mu_i)^2}{\sigma^2}}$$

the loglikelihood function (after dropping the constant involving $\sqrt{2\pi}\sigma$) is

$$\ell(\mu_1, \dots, \mu_I) = -\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij} - \mu_i)^2. \tag{13.6}$$

Under H_0 we have $\mu_i = \mu$, for $i = 1, \dots, I$ and the loglikelihood function becomes

$$\ell(\mu) = -\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij} - \mu)^2. \tag{13.7}$$

When we maximize the loglikelihood in (13.6) we get

$$\hat{\mu}_i = \bar{y}_{i.}$$

and

$$\begin{aligned}\ell(\hat{\mu}_1, \dots, \hat{\mu}_I) &= -\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij} - \bar{y}_{i.})^2 \\ &= -\frac{1}{2\sigma^2} SSE.\end{aligned}$$

When we maximize the loglikelihood in (13.7) we get

$$\hat{\mu}_i = \bar{y}_{..}$$

and

$$\ell(\hat{\mu}) = -\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij} - \bar{y}_{..})^2$$
$$= -\frac{1}{2\sigma^2}SST.$$

The log of the likelihood ratio *LR* in (11.6) is

$$\log LR = \ell(\hat{\mu}) - \ell(\hat{\mu}_1, \ldots, \hat{\mu}_I)$$

and multiplying this by -2, and combining with (13.7) and (13.6) after inserting the MLEs we get

$$\begin{aligned} -2\log LR &= \frac{1}{\sigma^2}SST - \frac{1}{\sigma^2}SSE \\ &= \frac{SS_{group}}{\sigma^2}. \end{aligned} \tag{13.8}$$

From (13.8), the likelihood ratio test will reject H_0 when SS_{group} is sufficiently large relative to σ^2.

The ANOVA F-statistic (13.5) arises from[3] (13.8) when we estimate σ^2 by *MSE* and normalize SS_{group} by its degrees of freedom, which is done for mathematical convenience (the ratio of MS_{group} to *MSE* follows an F_{ν_1,ν_2} distribution).

13.1.3 When there are only two groups, the ANOVA F-test reduces to a t-test.

In the special case of only two groups with two means μ_1 and μ_2, the null hypothesis H_0: $\mu_1 = \mu_2$ may be tested with a t-test. This turns out to be equivalent to the ANOVA F test and, in fact, the square of the t-statistic is equal to the F-statistic (compare the similar statements about regression on p. 337).

Example 13.1 (continued from p. 366) From the pooled standard deviation $s = 1.48$ reported on p. 366 we get the standard error of each mean $SE = s/\sqrt{6} = .60$. Comparing the active exercise group mean with the eight-week control we have a difference of $12.35 - 10.1 = 2.25$. Using the pooled estimate s, this difference has a standard error of $SE(\bar{X}_4 - \bar{X}_1) = s\sqrt{\frac{1}{6} + \frac{1}{6}} = .853$ and the t ratio is

$$t_{obs} = 2.25/.853 = 2.6$$

[3] When σ is unknown the derivation is slightly different because σ must be included among the parameters in the loglikelihood function, so its MLE must be found and the likelihood ratio is different; but the end result is equivalent to the F-test.

analogously with Eq. (10.19). Here, however, we are using *all* the data from the 4 groups to compute s, rather than only the data from two groups we are currently comparing. Therefore, we have 20 degrees of freedom going into s and thus 20 degrees of freedom for the t-test (rather than 10 degrees of freedom if we were using only the 2 groups). We obtain $p = .017$.

An alternative analysis compares the active exercise group with the other three groups, all of which could be considered controls. In this case, we would combine the data from the 3 control groups and thereby end up with two groups: the active exercise group and a single control group, the latter now having 18 observations. We would then use the "two-sample t" analysis, as in (10.21). Carrying this out, we obtain (i) a test of the null hypothesis that the means for these two groups are equal, which we may write as H_0: $\mu_{active} - \mu_{controls} = 0$, and (ii) a 95 % CI for the difference between the means $\mu_{active} - \mu_{controls}$.

First, we find the two means and standard errors to be 10.12 ± 0.59 and $11.81 \pm .34$, which gives a t-ratio of 2.46 on 22 degrees of freedom and $p = .022$. Second, applying the formula for the 95 % CI in Eq. (7.31) we find our 95 % CI for the decrease in mean age of walking for the active group compared with controls to be (.26, 3.1) months.

The conclusions from this analysis are different from those on p. 366, based on the F-test. We summarize on p. 374. □

13.1.4 *Two-way ANOVA assesses the effects of one factor while adjusting for the other factor.*

On p. 363 we described the distinction between one-way and two-way tables by contrasting Examples 13.1 and 13.2. To introduce the two-way analysis let us first look further at the data in Example 13.2.

Example 13.2 (continued from p. 363) Figure 13.2 displays the tapping rates for the three drugs across the four subjects. We can see that the subjects have very different tapping rates, but for all four of them the placebo rate is noticeably lower than that obtained with theobromine or caffeine. Also, the comparison of rates for theobromine and caffeine is inconsistent across subjects. The quantitative analysis, below, will support these qualitative observations. □

The two-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

where Y_{ij} is the observation for the ith treatment on the jth subject, $\mu + \alpha_i + \beta_j$ is its mean, and ϵ_{ij} is the error for the ith treatment and jth subject. Here, α_i is the increment added to the overall mean μ in obtaining the mean for the ith treatment while β_j is the increment added to overall mean in obtaining the mean for the jth subject. We say that α_i is the *effect* for the ith treatment and β_j the effect for the

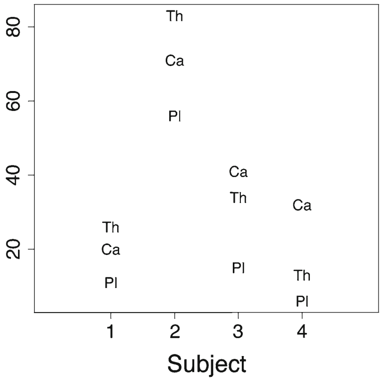

Fig. 13.2 Tapping rates displayed with identifiers "Pl" for placebo, "Ca" for caffeine, and "Th" for theobromine.

Table 13.5 Analysis of Variance table for data in Example 13.2.

Source	DF	SS	MS	F	p-value
Drugs	2	872	436	7.88	.021
Subjects	3	5478	1826	33	.0004
Error	6	332	55.3		
Total	11	6682			

The form of the table is similar to that in Table 13.4, except there are now F-ratios and p-values for both drugs and subjects

jth subject. A common terminology replaces the subjects with *blocks*, so that one would say β_j is the effect for the jth block. This terminology comes from the origin of ANOVA in agricultural field trials, where it referred to a block of land in a field.

As in one-way ANOVA, in two-way models the null hypothesis of interest is H_0: $\alpha_i = 0$ for all i. In the two-way case it is also possible to formulate the hypothesis that all the β_j's are zero, as well. This is not usually an object of investigation in experiments on multiple subjects because it would typically not be plausible for the subjects all to react the same way to the various treatments. However, statistics packages print out F-statistics and p-values for both hypotheses, so it's important to keep them straight (Table 13.5).

Example 13.2 (continued from p. 369) In the ANOVA for the finger tapping data there are two "factors" to be considered, drugs and subjects. Here, $F = 7.88$ on 2 and 6 d.f. with $p = .021$ indicates some evidence that the treatment means are different. There is also an F-ratio for subjects, which in fact is much larger and has a considerably smaller p-value: in this example, there is a very substantial difference among the subjects. In particular, the second subject has a much higher tapping rate than the others. The variability among subjects might be important to the conclusions one would wish to draw.

We may say something about the means, as well. For the three groups the mean tapping rates are, respectively, 22, 39, and 41. Standard errors are found by plugging in an estimate s of σ and again applying $SE = s/\sqrt{n}$. We have $s = \sqrt{MSE} = \sqrt{55.3} = 7.44$. Since there are 4 observations per treatment group, we use $n = 4$ and get 22 ± 3.7, 39 ± 3.7 and 41 ± 3.7. Clearly, the caffeine and theobromine groups have tapping rates substantially above that for the placebo group. □

13.1.5 When the variances are inhomogeneous across conditions a likelihood ratio test may be used.

The ANOVA F-test remains accurate for modest deviations from the homogeneity of variance assumption, which is assumption (iv) on p. 364. A rough rule of thumb is that as long as each ratio of pairs of standard deviations for two different groups is less than 3, the F-test should be accurate. However, in extreme cases where group i has a standard deviation σ_i that is much larger than the standard deviation σ_k for group k, there will be much more information in an observation y_{ij} about μ_i than in y_{kj} about μ_k. In such situations the usual F-statistic fails to take account of the differing contributions of data from different groups to the assessment of H_0 and it no longer has an F distribution. The problem may be fixed by re-deriving the likelihood ratio statistic and applying a permutation or bootstrap test. See Behseta et al. (2007) and references therein.

Example 4.7 (continued from p. 306) In examining directional information at each MEG brain source Wang et al. (2010) found grossly different standard deviations for the 4 different movement directions. They therefore applied the procedure of Behseta et al. (2007) to get likelihood ratio test statistics at every source and every time point. This was also used by Xu et al. (2011) within the permutation test described briefly on p. 306. □

13.1.6 More complicated experimental designs may be accommodated by ANOVA.

We have reviewed the fundamental ideas in ANOVA but have specified the procedures only in the two simplest cases involving one or two experimental factors. In many studies, especially involving human subjects, the designs can be more complicated. Sometimes they involve *multiple factors*, e.g., when there are 3 factors the analysis involves 3-way ANOVA. In Example 13.2 each subject's tapping rate was measured repeatedly, across 3 conditions. This is a special case of a *repeated measures* design. In many situations each subject is measured for all treatment conditions, but there is another factor, such as gender, that applies to groups of subjects. Such repeated-measures designs require specialized ANOVA methods. An additional possibility is that subjects, or other factors, may be considered themselves to provide an interesting source of variation. In this case their effects may be modeled as random variables. This generates *random-effects models* and they too require specialized techniques. We discuss random-effects models briefly in Chapter 16.

13.1.7 Additional analyses, involving multiple comparisons, may require adjustments to p-values.

Because ANOVA involves comparison of several means, many possible hypotheses may be of interest.

Example 13.1 (continued from p. 368) We have already looked at the data on development of motor control in two different ways. On p. 366 we used ANOVA to test the hypothesis of no differences among the mean age of walking, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. Then, on p. 368, we reported two further analyses. The first used a t-test to test the null hypothesis of no difference between the active exercise group and the eight-week control group mean ages of walking, H_0: $\mu_1 = \mu_4$ with a t-test. The second used a t-test to test the null hypothesis of no difference between the mean age of walking in the active exercise group and that in the three control groups combined, H_0: $\mu_1 = \frac{1}{3}(\mu_2 + \mu_3 + \mu_4)$. We also could have singled out the other control groups and tested H_0: $\mu_1 = \mu_2$ and H_0: $\mu_1 = \mu_3$. Furthermore, because the p-value quantifies the rarity, or surprise, of the results, we ought to ask what other results *might have been* as surprising as those we actually observed. What if the passive exercise group had produced apparent earlier walking, similar to the active exercise group, by comparison with the eight-week control group? Wouldn't that have been a result we would have found interesting? Once we admit that this, too, would have been reported as a finding, then we realize that we were, effectively, testing many possible null hypotheses. The problem of testing multiple hypotheses was discussed in Section 11.3. □

As illustrated in Example 13.1, above, ANOVA often generates many plausible null hypotheses and, in this context, the problem of multiple hypothesis testing is also called the problem of *multiple comparisons*. In Section 11.3 we presented the Bonferroni correction, which can be applied when the number of comparisons (null hypotheses) is easily enumerated. We commented that the Bonferroni method is conservative, in the sense of yielding adjusted p-values that sometimes seem unnecessarily large, making it relatively difficult to obtain statistically significant results. This has spawned a large literature on multiple comparison procedures, most of which aim to provide smaller p-values under specific circumstances, so that it becomes easier to declare statistical significance. For example, a method due to Dunnett assumes there is a single control group with mean μ_c and considers all null hypotheses of the form H_0: $\mu_i = \mu_c$, for $i \neq c$. When there are I means, there are $I - 1$ such null hypotheses and, under the standard ANOVA assumptions it is possible to find an exact p-value for this case. Similarly, when there is no single control group, a method due to Tukey examines all pairs of means, i.e., all null hypotheses of the form H_0: $\mu_i = \mu_j$ for distinct i and j. When there are I means, this narrows the number of hypotheses down to $\binom{I}{2}$ and, again, an exact p-value can be obtained.

We have two general comments on the problem of multiple comparisons in ANOVA. First, permutation tests discussed in Chapter 11 can be used to obtain p-values that take account of multiple testing procedures, as illustrated in Example 4.7 on p. 306. In Example 13.1, for instance, we might want to compare each of the 3 control groups to the active exercise group, using 3 t-tests. We then might focus on the t-test having the largest t-value. To obtain a p-value for this comparison we could create permutation pseudo-data and for each set of pseudo-data we could test all 3 null hypotheses of equality between mean of the active exercise group and the mean of each of the three control groups and we could store the largest of the 3 t-statistics based on the pseudo-data. A comparison of the largest t-statistic computed from the real data with those computed from the pseudo-data would give us a p-value, as in the cases examined in Section 11.2.1.

A second point is that multiple comparisons procedures in ANOVA are different than those arising in the neuroimaging of Example 11.3, which was used to motivate the multiple testing procedures discussed in Section 11.3.2. In neuroimaging there are typically thousands of null hypotheses, while in ANOVA, even when considering many possible combinations, the number is usually much smaller. The adjustments in ANOVA, including the Bonferroni correction, are therefore less severe. Importantly, when different multiple comparison methods lead to inconsistent conclusions it is an indication that the results are equivocal. In fact, in many ANOVA settings a very workable way to proceed is to begin by relying on the F test. If one obtains a significant F-statistic there is evidence for a difference among the means, and it therefore makes sense to go ahead and examine whichever means happen to look interesting, without worrying much about the process of selecting them. In other words, a widely-advocated method, sometimes called the *protected least-significant difference*, is to require a significant F statistic and then to report results from the many t tests, or any of them that seem to be of interest.

> *Details:* A *contrast* among the means is a linear combination $\sum_i c_i\mu_i$ for which $\sum c_i = 0$. For example, when $I = 4$, the contrast vector $c = (1, -1, 0, 0)$ would define the contrast $\mu_1 - \mu_2$. Corresponding to any contrast we have the null hypothesis that the contrast is zero, i.e.,
>
> $$H_0\colon \sum_{i=1}^{I} c_i\mu_i = 0. \tag{13.9}$$
>
> It is possible to define a test of this null hypothesis with a p-value that adjusts for examining all possible contrasts. In other words, the null hypothesis being tested is that H_0 in (13.9) holds for all contrast vectors c. This is usually called the *Scheffé* test. In terms of linear combinations of the means, this is a maximally protective procedure: it guards against spurious results from examining all possible linear

comparisons. Under the standard assumptions, it may be shown that the F-test is significant at level α if and only if there exists a linear contrast for which a test of H_0 defined by (13.9) is significant at level α according to the Scheffé test. □

Example 13.1 (continued from p. 372) Where does all this leave us in this example? We may summarize by saying that there is some evidence, but not strong evidence, that the active group mean age of walking is a bit younger than that for the control groups. The marginal nature of this evidence becomes clear when we ignore the special feature that the latter three groups are all controls and look for differences among all four groups: we find no evidence for this, according to the F-test. Given that it may be difficult to determine exactly when a given child walks, and it is not clear that the parents made this determination in the absence of knowledge about what to expect based on the experimental hypothesis, some skepticism would seem appropriate.[4] □

13.2 ANOVA as Regression

13.2.1 The general linear model includes both regression and ANOVA models.

We now return to the matrix formulation of multiple regression, discussed in Section 12.5.3, and show how linear regression may be used to solve problems of analysis of variance. The points are, first, it can be helpful conceptually to re-frame ANOVA as regression and, second, statistical software typically does this.

ANOVA concerns the comparison of means among several groups, corresponding to experimental conditions. Let us consider two simple examples. Suppose X is the $n \times 1$ vector of 1s

$$X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

We then compute $X^T X = n$ and $X^T Y = \sum y_i$ and find

$$(X^T X)^{-1} X^T y = \bar{y}.$$

Therefore, the sample mean may be found by applying regression with this very special version of the design matrix X.

[4] On the other hand, the paper by Zelazo et al. presented an additional measure where the results were more striking. On this subject, see Adolph (2002).

Next, consider two groups of m values $y_{11}, \ldots, y_{1m}$ and $y_{21}, \ldots, y_{2m}$, corresponding to two experimental conditions, having sample means $\bar{y}_1$ and $\bar{y}_2$. We define

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1m} \\ y_{21} \\ \vdots \\ y_{2m} \end{pmatrix} \tag{13.10}$$

and

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \tag{13.11}$$

where the first column contains m rows of 1s followed by m rows of 0s and the second column contains m rows of 0s followed by m rows of 1s. The first column of X is an *indicator variable*, indicating membership in the first group, i.e., the ith element of the first column of X is 1 if the ith element of y is in the first group and is 0 otherwise. The second column of X is an indicator variable indicating membership in the second group. We compute

$$X^T X = \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix}$$

$$X^T y = \begin{pmatrix} \sum y_{1i} \\ \sum y_{2i} \end{pmatrix}$$

and

$$(X^T X)^{-1} X^T y = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix}.$$

Thus, the sample means are obtained from multiple regression based on the design matrix in (13.11). In a similar manner we may use linear regression to compute means across several experimental conditions: for each condition we introduce an additional indicator variable as an additional column of the design matrix. The ANOVA from this regression becomes the same as the ANOVA table used in 1-way ANOVA. In this case of two conditions, the regression results would be equivalent to those from a t-test, as described in Section 13.1.3.

Before leaving the subject of indicator variables, let us make the further point that there are typically many reasonable choices of the way to code the columns of the X matrix. For example, if we reconsider two groups of m values $y_{11}, \ldots, y_{1m}$ and $y_{21}, \ldots, y_{2m}$, we could take

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{pmatrix}. \tag{13.12}$$

In this case, X is no longer made up of indicator variables, but its columns span the same space as that spanned by the indicator variables given in (13.11). That is, a vector $\boldsymbol{v}$ is a linear combination of the columns of X using (13.12) if and only if it is a linear combination of the columns of X using (13.11), though the coefficients of the linear combinations will be different in the two cases. Another way to say this is that the space of fitted values $V = \{X\beta^*,\ \beta^* \in R^2\}$, defined in Section 12.5.3, is the same regardless of whether the design matrix X takes the form of (13.11) or (13.12). Using (13.12) we obtain

$$X^T X = \begin{pmatrix} 2m & 0 \\ 0 & 2m \end{pmatrix}$$

$$X^T Y = \begin{pmatrix} \sum y_{1i} + \sum y_{2i} \\ \sum y_{1i} - \sum y_{2i} \end{pmatrix}$$

and

$$(X^T X)^{-1} X^T Y = \begin{pmatrix} \bar{y} \\ (\bar{y}_1 - \bar{y}_2)/2 \end{pmatrix}$$

where $\bar{y}$ is the overall mean. The second component $(\bar{y}_1 - \bar{y}_2)/2$ is often called a *contrast*, because it is "contrasting" the means of the groups. Generally speaking, a contrast vector (leading to a contrast estimate) is one whose components add to zero; see the discussion surrounding (13.9). In ANOVA settings, where there are multiple groups, it is often of interest to define an X matrix made up of contrast vectors, together with the vector 1_{vec} whose components are all equal to 1.[5]

A different way to represent ANOVA data is also useful, especially with statistical software. The input to software is typically a vector of data, such as represented

[5] It is also convenient to require the vectors to be orthogonal to one another, in which case they are called *orthogonal contrasts*. For orthogonal contrasts, each estimate is independent of the others. This is a topic discussed in many books on regression analysis and experimental design.

in (13.10), and the software must be informed which observations correspond to different groups. In conjunction with the data in (13.10) we define

$$L = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 2 \\ 2 \\ \vdots \\ 2 \end{pmatrix} \tag{13.13}$$

where the first m rows are 1s and the last m rows are 2s. The values 1 and 2 in the vector L in (13.13) are called the *levels* of the conditions or factor. In the case of the finger tapping data in Example 13.2 we could define $y = (11, 26, 15, 6, 26, 83, 34, 13, 20, 71, 41, 32)^T$ and then set

$$L = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 2 & 1 \\ 2 & 2 \\ 2 & 3 \\ 2 & 4 \\ 3 & 1 \\ 3 & 2 \\ 3 & 3 \\ 3 & 4 \end{pmatrix} \tag{13.14}$$

so that the first column of the level matrix L represents the “levels” of the drugs (1 for Placebo, 2 for Theobromine, 3 for Caffeine) and the second column represents “levels” of the subjects (1 for first subject, etc.). Statistical software used for 1-way or 2-way ANOVA requires some identifier of group structure, such as (13.13) and (13.14). It is possible to produce a design matrix X from a level matrix L, and vice-versa. ANOVA software often provides functions for this purpose.

13.2.2 In multi-way ANOVA, interactions are often of interest.

In Section 12.5.6 we described the way interactions between explanatory variables arise in multiple regression. Interactions play an important role in many ANOVA settings. Here we consider the simplest case of interactions between two conditions that each have two levels and then connect the ANOVA and regression contexts.

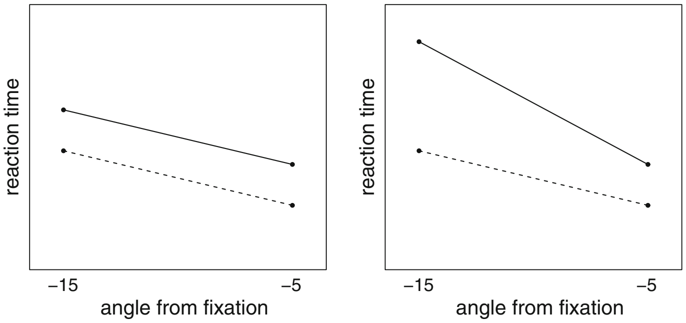

Fig. 13.3 Hypothetical plots of mean saccadic reaction time when angular distance from fixation to target is either −15 or −5 degrees, i.e., when the eyes fixate either 15 or 5 degrees to the *right* of the target. *Solid lines* correspond to patients; *dashed* correspond to controls. In the *left plot* the lines are parallel, indicating the reaction time is longer among patients by the same amount for both angular distances; there is no interaction between angular distance and subject classification. In the *right plot* the increase reaction time among patients is greater at −15 degrees than at −5 degrees, so the lines are no longer parallel; this represents an interaction between angular distance and subject classification.

Example 2.1 (continued) In the experiment on saccadic reaction time, Behrmann et al. (2002) sought to characterize the way eye saccades differed among patients with hemispatial neglect compared with control subjects.[6] We use this context to illustrate presence and absence of interaction. Let Y be saccadic reaction time, x_1 represent the distance from eye fixation to target, measured in degrees of angle to the right. When the target was on the left side of fixation, which was the neglected side for the patients, the angle was negative. We let $x_1 = 1$ when the target was at -15 degrees (15 degrees to the left of fixation) and $x_1 = 0$ when the target was at -5 degrees. We also let x_2 be an indicator variable indicating patients, i.e., $x_2 = 1$ for patients and $x_2 = 0$ for control subjects. These variables define 4 mean saccadic reaction times: μ_{11} is the mean reaction time among patients when the target was at -15 degrees; μ_{10} is the mean reaction time among controls when the target was at -15 degrees; μ_{01} is the mean reaction time among patients when the target was at -5 degrees; and μ_{00} is the mean reaction time among controls when the target was at -5 degrees. If patients and controls reacted similarly, except that patients had a fixed latency of response, then the means would satisfy

$$H_0\colon \mu_{11} - \mu_{10} = \mu_{01} - \mu_{00} \tag{13.15}$$

which is the null hypothesis of no interaction. The left side of Fig. 13.3 displays a possible set of four means satisfying H_0 in (13.15). On the other hand, if the patients also moved their eyes more slowly then their mean response would be even longer at -15 than at -5, and we would have

[6] The purpose of the study was to distinguish responses based on eye-centered coordinates, head-centered coordinates, and trunk-centered coordinates.

$$\mu_{11} - \mu_{10} > \mu_{01} - \mu_{00},$$

as shown on the right side of Fig. 13.3. The second case, but not the first, corresponds to the presence of an interaction effect between x_1 and x_2. Statistical evidence of an interaction effect would be found by obtaining a statistically significant interaction of x_1 and x_2. □

In Section 12.5.6 we said that in regression based on explanatory variables x_1 and x_2 the variable defined as the product $x_1 x_2$ represents the interaction between these variables. In the equation

$$y = a + bx_1 + cx_2 + dx_1x_2, \tag{13.16}$$

which was Eq. (12.70), we noted that when $d = 0$ the graphs of y versus x_1 for two different values of x_2 produce two parallel lines, but when $d \neq 0$ the two lines are no longer parallel. Figure 13.3 displays an example of this phenomenon. In ANOVA the variables correspond to the experimental design, as outlined briefly in Section 13.2.1, and interaction effects are found via least-squares regression.[7] We omit details. Here is a neuroimaging example.

Example 13.3 Neural correlates of delay of gratification Successful decision making often requires an ability to forgo immediate gain in favor of increased future reward. Casey et al. (2011) reported fMRI results for group of individuals who had been studied 40 years earlier, as preschool children, for their ability to delay gratification. Previously it had been shown that performance on a delay-of-gratification task during childhood predicted ability to perform on a go/no-go task as adults. The authors imaged their subjects during go/no-go tasks. One of their findings involved the inferior prefrontal gyrus, an area thought to be involved in impulse control during similar tasks. Based on the childhood results, the authors categorized the subjects has either "low" or "high" childhood ability to delay gratification. The question was whether the two groups had different neural activity in the inferior prefrontal gyrus 40 years later, and the experimental prediction was that in the low ability group neural activity in the inferior prefrontal gyrus would be similar on go and no-go trials, but for the high ability group there would be much stronger activity on no-go trials (when impulse control is operative) than on go trials. This corresponds to an interaction between trial type ("go" vs. "no-go") and subject group (low or high childhood ability). Let us write the means of the neural activity in go and no-go trials[8] for the low and high ability groups as $\mu_{\text{go}}^{\text{low}}$, $\mu_{\text{nogo}}^{\text{low}}$, $\mu_{\text{go}}^{\text{high}}$, $\mu_{\text{nogo}}^{\text{high}}$. The null hypothesis of no interaction would be

[7] ANOVA may also be applied, as a special case of regression, when one explanatory variable is quantitative and another variable is an ANOVA indicator variable. This is usually called *analysis of covariance* or ANCOVA. Its purpose is to adjust the ANOVA for effects of the quantitative variable. See p. 332.

[8] We are here simplifying by ignoring some aspects of the experimental design.

$$H_0\colon \mu_{\text{nogo}}^{\text{low}} - \mu_{\text{go}}^{\text{low}} = \mu_{\text{nogo}}^{\text{high}} - \mu_{\text{go}}^{\text{high}}.$$

Casey et al. found evidence against H_0, reporting a statistically significant interaction ($p = .014$) between trial type and subject group. □

In Example 13.3 it was hypothesized that for one group of subjects (the low ability group) the means under the two conditions (μ_{go} and μ_{nogo}) would be very close in magnitude while for the other group (the high ability group) they would be quite different. It would be tempting to test $H_0 : \mu_{\text{go}} = \mu_{\text{nogo}}$ for each of the two groups: if the test were significant for the second group but not for the first group one might then conclude that the two groups were different with regard to the two conditions. In fact, such reasoning is common in neuroscience and psychology (see Nieuwenhuis et al. 2011). Unfortunately, it is not correct. As pointed out in Section 10.4.8, a non-significant test does not itself provide evidence for H_0. Thus, in particular, a non-significant test of $H_0 : \mu_{\text{go}} = \mu_{\text{nogo}}$ does not provide evidence that the two means are approximately the same. Instead, a confidence interval or test for the interaction effect should be reported, as in Example 13.3.

13.2.3 ANOVA comparisons may be adjusted using analysis of covariance.

In comparing results under two or more experimental conditions it often happens that the subjects (or other experimental units) are not comparable with respect to some background variable, often called a *covariate*. For instance, suppose we have data under two conditions as in (13.10). As indicated in Section 13.2.1, the two means $\bar{y}_1$ and $\bar{y}_2$ may be compared by performing the regression of y on the X matrix given by (13.11), producing results that are equivalent to a t-test (and a t-based confidence interval). Now suppose we have an additional covariate u with values given by

$$u = \begin{pmatrix} u_{11} \\ \vdots \\ u_{1m} \\ u_{21} \\ \vdots \\ u_{2m} \end{pmatrix}. \tag{13.17}$$

If we regress y on both X and u we will obtain a comparison between the means under the two conditions *after adjusting for* the covariate u. As explained at the beginning of Section 12.5, this is a consequence of the regression formulation.

Example 13.4 Improving Working Memory in Children with ADHD Deficits in working memory (WM) are associated with ADHD. Klingsberg et al. (2005)

reported results of a randomized, controlled double-blind trial aimed at assessing the possible benefits of a computerized training program aimed at improving WM. (The virtues of randomized, double-blind trials are discussed briefly in Section 13.4.) The training program consisted of at least 25 sessions, each lasting roughly 40 minutes, in which subjects completed WM tasks. In the experimental condition the difficulty of the WM tasks was automatically adjusted to match the current assessment of the subject's WM. In the control condition difficulty remained at an initial low level. A total of 42 children with ADHD (ages 7–12) were randomly allocated to one of the two conditions and completed the entire protocol.

The key outcome was "span-board" task performance, a standard assessment of visuospatial WM. This was assessed at the subject's initial visit and then twice after training had been completed: both 5–6 weeks after the initial visit and, again, 3 months subsequent to this. Baseline score at the initial visit was used as a covariate, together with age and number of days of training. The authors reported a highly significant difference between span-board task performances under the experimental and control conditions, after adjusting for the covariates, with $p = .001$ at 5–6 weeks post initial visit and $p = .002$ at the second visit 3 months later. This constitutes strong evidence that WM can be improved by training among ADHD children. □

The use of covariates to adjust comparisons in the context of ANOVA is usually called *analysis of covariance*.

13.3 Nonparametric Methods

ANOVA assumption (v) on p. 364, normality, is often suspect. Because ANOVA is a special case of regression and, under weak conditions, the least-squares estimates are asymptotically normal according to (12.63), the ordinary ANOVA procedures work well with large samples even for non-normal data. Sometimes, however, the sample size may be modest while the data appear grossly non-normal. In the next two subsections we discuss two approaches to ANOVA for non-normal data. The first, in Section 13.3.1, is based on *ranks*, and the idea is to replace each data value by its rank within the whole data set. Rank-based procedures remove the assumption of a specific distributional form. The second approach involves permutation and bootstrap tests, as discussed in Sections 11.2.1 and 11.2.2. We describe these very briefly in Section 13.3.2.

The body of ANOVA methods under the assumption of normality are called *parametric*, meaning that they are based on probability models characterized by a small number of parameters. The methods in Sections 13.3.1 and 13.3.2 are *nonparametric*. Please note, however, that all these procedures continue to make the more consequential assumptions of additivity and independence of the errors.

Table 13.6 Data from Frezza et al. (1990) on first-pass alcohol metabolism.

Alcoholic Women	Non-alcoholic Women	Alcoholic Men	Non-alcoholic Men
0.6	0.4	1.5	0.3
0.6	0.1	1.9	2.5
1.5	0.2	2.7	2.7
	0.3	3.0	3.0
	0.3	3.7	4.0
	0.4		4.5
	1.0		6.1
	1.1		9.5
	1.2		12.3
	1.3		
	1.6		
	1.8		
	2.0		
	2.5		
	2.9		

13.3.1 Distribution-free nonparametric tests may be obtained by replacing data values with their ranks.

To describe rank-based ANOVA we begin with an example.

Example 13.5 Alcohol metabolism among men and women Women seem to have a lower tolerance for alcohol than men, and are more prone to develop alcohol-related diseases. When men and women of the same size and history of drinking consume equal amounts of alcohol, the alcohol in the bloodstream of the women tends to be higher. In research by Frezza et al. (1990), the “first-pass” metabolism of alcohol in the stomach was studied. The data shown in Table 13.6 come from 18 women and 14 men who volunteered to be studied. Each subject was given two doses of .3 g ethanol per kilogram of body weight, one orally and one intravenously on two different days. The difference in concentrations of alcohol in the blood (at some fixed time after administration), between the intravenous dose and the oral dose, provides a measure of first-pass metabolism in the digestive system and liver; this defines the response variable in the table, with units in mmols per liter per hour. If first-pass metabolism were more effective in men than women, the difference in levels following intravenous and oral administration would tend to be higher among men.

We begin by ignoring the distinction between alcoholic and non-alcoholic subjects. This reduces the data to two groups: women and men. The data in Table 13.6 are strikingly skewed toward high values. One possibility would be transform the data and apply the usual t-test. Instead, we describe a rank-based analysis.

Table 13.7 Data from Table 13.6 together with corresponding ranks, where the smallest observation has rank 1 and the largest has rank $n = 32$.

Case	Difference	Female	Rank
1	0.6	1	8.5
2	0.6	1	8.5
3	1.5	1	14.5
4	0.4	1	6.5
5	0.1	1	1.0
6	0.2	1	2.0
7	0.3	1	4.0
8	0.3	1	4.0
9	0.4	1	6.5
10	1.0	1	10.0
11	1.1	1	11.0
12	1.2	1	12.0
13	1.3	1	13.0
14	1.6	1	16.0
15	1.8	1	17.0
16	2.0	1	19.0
17	2.5	1	20.5
18	2.9	1	24.0
19	1.5	0	14.5
20	1.9	0	18.0
21	2.7	0	22.5
22	3.0	0	25.5
23	3.7	0	27.0
24	0.3	0	4.0
25	2.5	0	20.5
26	2.7	0	22.5
27	3.0	0	25.5
28	4.0	0	28.0
29	4.5	0	29.0
30	6.1	0	30.0
31	9.5	0	31.0
32	12.3	0	32.0

The data are printed out again in Table 13.7, with each rank listed at the end. The rank goes from 1 up to 32, with the smallest value getting the rank 1 and the largest value getting the rank 32. Ranks ending in .5 represent ties, i.e., cases in which some data value appears twice. The women in the study have a 1 in the “females” column. □

Rank-sum methods compare the ranks of the two groups. That is, if one group has values of its ranks that are sufficiently much larger than those of the other group, there will be evidence that the means of the two groups are different. More specifically, we

Table 13.8 Four observations from Table 13.7.

Case	Difference	Female	Rank
1	0.6	1	1
18	2.9	1	3
19	1.5	0	2
32	12.3	0	4

may find the sum of the ranks from one of the groups and see whether it is either much larger or much smaller than would be expected if, in fact, the two groups followed the same distribution. Based on the null hypothesis that the probability distributions for the two groups are the same, we can get a p-value. The test statistic W is the sum of the ranks from one of the two groups. This is the *rank-sum test*. It is sometimes called the Wilcoxon rank-sum test, and it is also often called the Mann-Whitney test. Let us write the distribution functions for males and females as $F_{\text{males}}(x)$ and $F_{\text{females}}(x)$. The rank-sum test tests the null hypothesis

$$H_0 \colon F_{\text{males}}(x) = F_{\text{females}}(x)$$

for all x.

To be specific about the procedure, suppose the alcohol metabolism data consisted only of the four observations in Table 13.8. In this case we would rank the data as 1, 3, 2, 4 (0.6 is the smallest, 2.9 is the third smallest, 1.5 is the second smallest, and 12.3 is the fourth smallest). Then we would add up the values of the ranks for the females to get the statistic $W = 1 + 3 = 4$.

Example 13.5 (continued) For the data in Table 13.7 we obtained the rank-sum test statistic $W_{obs} = 330$ with $p = .0002$. This may be compared with the usual t-based method gave $T_{obs} = 3.41$ with $p = .0042$. In this case, we get similar conclusions and are reassured that the assumption of normality is not crucial. In fact, if we first transform the data by taking logs, the usual t-test gives $p = .0002$. □

An analogous procedure for several groups is called the *Kruskal-Wallis test*. It may be used in place of the usual F-statistic from an ANOVA.

Example 13.5 (continued) When all four groups are used and the data are transformed by logs we find $p = .003$ from the usual ANOVA F-test. In fact, the residual analysis for the log-transformed data looks pretty good and we would find little reason to worry about the assumption of normality. However, using the Kruskal-Wallis test we get $p = .002$, which again corroborates the conclusion.

In using this example to describe rank-based methods we have concentrated on technique, but a more basic concern lurks here: we must wonder about the extent to which the volunteers represent the population as a whole, and whether the particular men and women in the study might for some reason self-select in a manner that was related to their alcohol metabolism. We return to such considerations in Section 13.4.

□

13.3.2 *Permutation and bootstrap tests may be used to test ANOVA hypotheses.*

In Section 11.2 we described how permutation and bootstrap tests may be used as alternatives to the t-distribution for computing a p-value in order to test H_0: $\mu_1 = \mu_2$ based on data involving sample sizes n_1 and n_2. The essential method was to (i) merge the data, then (ii) repeatedly resample the $n_1 + n_2$ data values, putting them arbitrarily into groups of size n_1 and n_2 to create pseudo-data, (iii) to each pseudo-data pair of samples apply the t-statistic, and finally (iv) see what proportion of the pseudo-data give t-statistic values greater than that observed in the real data. When the sampling is done without replacement the method is a permutation test, and with replacement it becomes a bootstrap test.

For one-way ANOVA the procedure is exactly analogous. For instance, with 3 conditions we would have data with sample sizes n_1, n_2, and n_3; we would follow step (i) then in (ii) resample the $n_1 + n_2 + n_3$ data values and put them into groups of sizes n_1, n_2, n_3; in (iii) we would get the F-statistic, and likewise in (iv) we would see what proportion of the pseudo-data F values exceed the F obtained for the real data.

Two-way ANOVA is more complicated because the two-way structure must be respected, but the concept is the same. See Manly (2007).

13.4 Causation, Randomization, and Observational Studies

13.4.1 *Randomization eliminates effects of confounding factors.*

Most studies aim to provide causal explanations of observed phenomena. To claim causality, investigators must argue that alternative explanations of an observed relationship are implausible.

Example 13.6 IQ and breast milk Lucas et al. (1992) obtained IQ test scores from 300 children who had been premature infants and initially fed milk by a tube. The children were 8 years old when they took the IQ test. The milk they had been fed by tube was either breast milk or prepared formula, or some combination of the two. Of interest was the relationship between IQ test scores and the proportion of milk the infants received that was breast milk. The amount of breast milk a baby had drunk was determined by whether or not the mother wished to feed the infant by breast milk, and how much milk the mother was able to express. □

In Example 13.6, immediately we must be aware of possible *confounding factors*. The decision to administer the treatment, i.e., to use breast milk or not, was the mother's; whatever might determine that decision *and also be related to subsequent IQ* would affect the observed relationship between IQ and consumption of breast

Table 13.9 Regression results from Lucas et al. (1992).

Explanatory variable	Estimated coefficient	p-Value
Social class	−3.5	.0004
Mother's education	2.0	.01
Female or not	4.2	.01
Days of ventilation	−2.6	.02
Received breast milk or not	8.3	<0.0001

The increase in IQ after adjusting for the other variables was 8.3 points (with $p < 0.0001$)

milk. If, for example, mothers who chose to breast feed were also more likely to provide intellectual stimulation to their young children, then the decision to breast feed could appear to raise IQ even though it was the increased stimulation that had the greater impact. The study would be free of these concerns if babies instead received a randomly-determined percentage of breast milk, but few mothers would give up this decision in order to be part of a scientific investigation.

Example 13.6 (continued) In an attempt to control confounding factors, and to reduce variability and make the comparisons more sensitive, the researchers performed a regression that included characteristics of both the mothers and the babies: social class (ordered from 1 to 5 with 5 being highest), mother's education (ordered from 1 to 5 with 5 being highest), whether or not the child was a female (1 if female, 0 if male), the number of days of ventilation of the baby after birth, and whether or not the baby received any breast milk (1 if yes, 0 if no). The results of the regression are shown in Table 13.9.

Let us begin by interpreting the main finding. If we hold fixed social class, mother's education, sex of the baby, and days of ventilation, there is a highly significant effect of whether or not the baby received breast milk, with breast milk increasing subsequent IQ, on average, by 8.3 points. This is quite a large effect. If it were felt appropriate to generalize from these data to the population at large, this effect would certainly be something the pediatric professions would pay attention to.

Should we believe that early consumption of breast milk would tend to increase IQ in the general population? □

To analyze the possibility of confounding factors it is useful to introduce some terminology and list some basic points.

In both experiments and observational studies, we are typically interested in effects of some explanatory variable or treatment on a response variable. A study is called an *experiment* when it imposes treatment conditions on some subjects; measurements on that subject are called the *response variable*. On the other hand, *observational studies* examine relationships between response variables and potential explanatory variables, which could become treatments, but there is no active administration of a treatment. A *confounding factor* (or *confounding variable*) is one that affects both the response variable and an explanatory variable; its effects on the response can not be distinguished from the effects of the explanatory variable of interest on the response.

The particular subjects being experimented upon may have special characteristics that make them different than those about which one may wish to draw conclusions. In many situations, carefully designed experiments can avoid these difficulties. *Randomization*, meaning the random allocation of the treatment to the subject provides a way of avoiding confounding variables; *double-blind* experiments can avoid hidden biases in the response measurements. It is also important to keep in mind that response variables and explanatory variables may not accurately capture what they are purported to be measuring. Strict adherence to the experimental *protocol* can also help avoid mismeasured variables. More generally, errors that can result from failure to adhere to protocol have been emphasized by Simmons et al. (2011).

Well-designed, randomized experiments can support causal explanations for associations between response and explanatory variables. More specifically, based on a well-designed experiment, it may be possible to say that, up to some degree of statistical uncertainty (represented by a standard error or confidence interval), a response will on average increase or decrease by a particular amount when an explanatory variable changes its value by some number of units (including being present rather than absent, as is the case for typical treatments).

In fact, it is possible to define a *causal effect*, and the corresponding association effect that would be observed in data. There is then a theorem saying that in a randomized experiment the causal effect is equal to the association effect (e.g., Wasserman (2004, Chapter 16)). In other words, for a randomized experiment, association *is* causation (see Section 12.4.2).

13.4.2 Observational studies can produce substantial evidence.

Although it is preferable to have data from a well-designed randomized experiment, there are situations in which it is impossible to randomly assign subjects to treatments. For example, one can not tell people whether they will be in "smoking" or "non-smoking" groups. Still, very convincing evidence can accumulate from observational studies—as in fact has happened in the case of smoking. Several observed patterns may increase the plausibility of an explanatory variable as a cause of a response variable:[9]

- The explanatory variable or treatment precedes observation of the response, and in terms of timing can thus act as a cause.
- Large effects are observed; this makes it less likely that the association is due to a confounding variable. One often-cited example is that mortality due to scrotum cancer among chimney sweeps was about 200 times above the population levels early in the 20th century.

[9] A widely-cited source for many of these ideas is Hill (1971).

- A quantitative “dose-response” relationship is observed, in which an increase in the explanatory variable increases (or decreases) the observed response, as opposed to simply an observation of an effect when a treatment is applied versus not applied.
- There is physiological evidence to support a theory that could explain the putative causal relationship.
- There are no anomalous results that seem difficult to explain; anomalous results may signal the presence of confounding variables.
- Similar results are obtained under differing experimental studies; confounding variables are often less likely to be present in each of the different studies.

Example 13.6 (continued) Now, let us reexamine the IQ and breast milk results with these principles in mind. First, the study is prospective, in the sense that children received some percentage of breast milk and then were followed over time to see what IQ score they got many years later. Second, the estimated effect is reasonably large—8 IQ points is about half of a standard deviation in the population as a whole. Third, there is physiological relevance: pediatricians recommend that mothers breastfeed their babies for nutritional reasons. We have not done a careful review of the literature, however, and do not have the expertise to comment critically on this basic scientific issue.

Concerning the dose-response relationship, in the regression reported above the breast milk variable merely indicates whether or not the infant received breast milk; but the authors reported a similar regression using instead *percentage* breast milk where the regression coefficient was .09, which says that holding the same variables fixed, for every 10 % increase in breast milk the subsequent IQ would go up on average by nearly a full point. This last result is important: by removing the decision of whether or not to use breast milk as an explanatory variable, the confounding variables associated with that decision are no longer a concern.[10] Now we must shift to the question of whether some confounding variables may affect both the amount of milk a mother can express and the subsequent IQ of the child. If not, we would be regarding the percentage breast milk actually delivered as if it were a randomly-determined percentage. One possible confounding variable would be the health of the mother during pregnancy: mothers who are unable to express much milk might conceivably have been providing worse nutrition to the fetus.

As far as anomalous results are concerned, here are two possibilities: first, given the other variables, subsequent IQ decreases as social class increases, which is surprising; second, given the other variables, female babies have higher subsequent IQs. There should be explanations for these outcomes. Otherwise, they raise doubts.[11]

Overall, from the report of this study we have given here, there is clearly a substantial association between increased administration of breast milk and increased

[10] We are here assuming that the reported regression is not being driven primarily by inclusion of lots of babies with zero percent breast milk, but rather holds among the non-zero percentage babies.

[11] We do not have the full results when percentage breast milk is used, so we don’t know whether these associations diminish or change sign in that case.

IQ, when social class (measured in the way the authors did), mother's education, and days of ventilation are held fixed. However, it remains possible that some confounding variables affect breast-milk expression and IQ. As we write this, 20 years has passed since the publication of the 1992 paper. While the topic remains controversial, subsequent research has been informative. For further information see Brion et al. (2011) and the references therein. □

Example 13.5 (continued) Returning to the alcohol metabolism example, let us now consider the possibility of confounding due to the use of volunteers in the study. The chief concern is whether volunteers are different than the rest of the population with respect to alcohol metabolism. This is at least plausible, though in order to affect the study, the volunteer men and women would have to be different. For example, if the women who volunteered tended to have trouble with alcohol metabolism (perhaps they thought the study sounded interesting because they knew they had a high susceptibility to the effects of alcohol) but men just wanted the money, then the differential effect would tend to be larger in this sample than in the population. Is this kind of hypothetical scenario reasonable, or really a stretch of the imagination? Your answer to this question determines how much faith you will put in the results. □