Graduate Texts in Physics

Simon Širca

Probability for Physicists

Springer

Graduate Texts in Physics

Graduate Texts in Physics

Graduate Texts in Physics publishes core learning/teaching material for graduate- and advanced-level undergraduate courses on topics of current and emerging fields within physics, both pure and applied. These textbooks serve students at the MS- or PhD-level and their instructors as comprehensive sources of principles, definitions, derivations, experiments and applications (as relevant) for their mastery and teaching, respectively. International in scope and relevance, the textbooks correspond to course syllabi sufficiently to serve as required reading. Their didactic style, comprehensiveness and coverage of fundamental material also make them suitable as introductions or references for scientists entering, or requiring timely knowledge of, a research field.

More information about this series at http://www.springer.com/series/8431

Simon Širca

Probability for Physicists

Springer

Simon Širca
Faculty of Mathematics and Physics
University of Ljubljana
Ljubljana
Slovenia

ISSN 1868-4513 ISSN 1868-4521 (electronic)
Graduate Texts in Physics
ISBN 978-3-319-31609-3 ISBN 978-3-319-31611-6 (eBook)
DOI 10.1007/978-3-319-31611-6

Library of Congress Control Number: 2016937517

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

University-level introductory books on probability and statistics tend to be long—too long for the attention span and immediate horizon of a typical physics student who might wish to absorb the necessary topics in a swift, direct, involving manner, relying on her existing knowledge and physics intuition rather than asking to be taken through the content at a slow and perhaps over-systematic pace.

In contrast, this book attempts to deliver a concise, lively, intuitive introduction to probability and statistics for undergraduate and graduate students of physics and other natural sciences. Conceived primarily as a text for the second-year course on *Probability in Physics* at the Department of Physics, Faculty of Mathematics and Physics, University of Ljubljana, it has been designed to be as relieved of unnecessary mathematical ballast as possible, yet never to be mathematically imprecise. At the same time, it is hoped to be colorful and captivating: to this end, I have strived to avoid endless, dry prototypes with tossing coins, throwing dice and births of girls and boys, and replace them wherever possible by physics-motivated examples, always in the faith that the reader is already familiar with "at least something". The book also tries to fill a few common gaps and resurrect some content that seems to be disappearing irretrievably from the modern, Bologna-style curricula. Typical witnesses of such efforts are the sections on extreme-value distributions, linear regression by using singular-value decomposition, and the maximum-likelihood method.

The book consists of four parts. In the first part (Chaps. 1–6) we discuss the fundamentals of probability and probability distributions. The second part (Chaps. 7–10) is devoted to statistics, that is, the determination of distribution parameters based on samples. Chapters 11–14 of the third part are "applied", as they are the place to reap what has been sown in the first two parts and they invite the reader to a more concrete, computer-based engagement. As such, these chapters lack the concluding exercise sections, but incorporate extended examples in the main text. The fourth part consists of appendices. Optional contents are denoted by asterisks $\star$. Without them, the book is tailored to a compact one-semester course;

with them included, it can perhaps serve as a vantage point for a two-semester agenda.

The story-telling and the style are mine; regarding all other issues and doubts I have gladly obeyed the advice of both benevolent, though merciless reviewers, Dr. Martin Horvat and Dr. Gregor Šega. Martin is a treasure-trove of knowledge on an incredible variety of problems in mathematical physics, and in particular of *answers* to these problems. He does not terminate the discussions with the elusive "The solution exists!", but rather with a fully functional, tested and documented computer code. His ad hoc products saved me many hours of work. Gregor has shaken my conviction that a partly loose, intuitive notation could be reader-friendly. He helped to furnish the text with an appropriate measure of mathematical rigor, so that I could ultimately run with the physics hare and hunt with the mathematics hounds. I am grateful to them for reading the manuscript so attentively. I would also like to thank my student Mr. Peter Ferjančič for leading the problem-solving classes for two years and for suggesting and solving Problem 5.6.3.

I wish to express my gratitude to Professor Claus Ascheron, Senior Editor at Springer, for his effort in preparation and advancement of this book, as well as to Viradasarani Natarajan and his team for its production at Scientific Publishing Services. http://pp.books.fmf.uni-lj.si

Ljubljana Simon Širca

Contents

Part I Fundamentals of Probability and Probability Distributions

Part II Determination of Distribution Parameters

Part III Special Applications of Probability

Part I
Fundamentals of Probability and Probability Distributions

Chapter 1
Basic Terminology

Abstract The concepts of random experiment, outcomes, sample space and events are introduced, and basic combinatorics (variations, permutations, combinations) is reviewed, leading to the exposition of fundamental properties of probability. A discussion of conditional probability is offered, followed by the definition of the independence of events and the derivation of the total probability and Bayes formulas.

1.1 Random Experiments and Events

A physics experiment can be envisioned as a process that maps the initial state (input) into the final state (output). Of course we wish such an experiment to be *non-random:* during the measurement we strive to control all external conditions—input data, the measurement process itself, as well as the analysis of output data—and justly expect that each repetition of the experiment with an identical initial state and in equal circumstances will yield the same result.

In a *random experiment*, on the other hand, it *may* happen that multiple repeats of the experiment with the same input and under equal external conditions will end up in different outputs. The main feature of a random experiment is therefore our inability to uniquely predict the precise final state based on input data. We rather ask ourselves about the *frequency of occurrence* of a specific final state with respect to the number of trials. That is why this number should be as large as possible: we shall assume that, in principle, a random experiment can be repeated infinitely many times.

A specific output of a random experiments is called an *outcome*. An example of an outcome is the number of photons measured by a detector, e.g. 12. The set of all possible outcomes of a random experiment is called the *sample space*, S. In the detector example, the sample space is the set $S = \{0, 1, 2, \ldots\}$. Any subset of the sample space is called an *event*. Individual outcomes are *elementary* events. Elementary events can be joined in *compound events:* for example, the detector sees more than 10 photons (11 or 12 or 13, and so on) or sees 10 photons and less than 20 neutrons simultaneously.

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_1

The events, elementary or compound, are denoted by letters A, B, C, ... The event that occurs in all repetitions of the experiment—or can be assumed to occur in all future tries—is called a *certain* or *universal event* and is denoted by U. The event that does not occur in any repetition of the experiment is called an *impossible event*, denoted by $\varnothing$ or $\{\ \}$. The relations between events can be expressed in the language of set theory. Take two events A and B and consider the possibility that at least one of them occurs: this eventuality is called the *sum of events* and is denoted by

$$A \cup B.$$

Summing events is commutative and associative: we have $A \cup B = B \cup A$ and $(A \cup B) \cup C = A \cup (B \cup C)$. The sum of two events can be generalized: the event that at least one of the events $A_1, A_2, \ldots, A_n$ occurs, is

$$A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{k=1}^{n} A_k.$$

The event that both A and B occur simultaneously, is called the *product of events A* and B. It is written as

$$A \cap B$$

or simply

$$AB.$$

For each event A one obviously has $A\varnothing = \varnothing$ and $AU = A$. The product of events is also commutative and associative; it holds that $AB = BA$ and $(AB)C = A(BC)$. The compound event that all events $A_1, A_2, \ldots, A_n$ occur simultaneously, is

$$A_1 A_2 \ldots A_n = \bigcap_{k=1}^{n} A_k.$$

The addition and multiplication are related by the distributive rule $(A \cup B)C = AC \cup BC$. The event that A occurs but B does not, is called the *difference of events* and is denoted by

$$A - B.$$

(In general $A - B \neq B - A$.) The events A and B are *exclusive* or *incompatible* if they can not occur simultaneously, that is, if

$$AB = \varnothing.$$

The events A are B *complementary* if in each repetition of the experiment precisely one of them occurs: this implies

$$AB = \varnothing \quad \text{and} \quad A \cup B = U.$$

An event complementary to event A is denoted by $\bar{A}$. Hence, for any event A,

$$A\bar{A} = \varnothing \quad \text{and} \quad A \cup \bar{A} = U.$$

Sums of events in which individual pairs of terms are mutually exclusive, are particularly appealing. Such sums are denoted by a special sign:

$$A \cup B \overset{\text{def.}}{=} A + B \quad \Leftrightarrow \quad A \cap B = \{\,\}.$$

Event sums can be expressed as sums of incompatible terms:

$$A_1 \cup A_2 \cup \cdots \cup A_n = A_1 + \bar{A}_1 A_2 + \bar{A}_1 \bar{A}_2 A_3 + \cdots + \big(\bar{A}_1 \bar{A}_2 \ldots \bar{A}_{n-1} A_n\big). \tag{1.1}$$

The set of events

$$\{A_1, A_2, \ldots, A_n\} \tag{1.2}$$

is called the *complete set of events*, if in each repetition of the experiment precisely one of the events contained in it occurs. The events from a complete set are all possible ($A_i \neq \varnothing$), pair-wise incompatible ($A_i A_j = \varnothing$ for $i \neq j$), and their sum is a certain event: $A_1 + A_2 + \cdots + A_n = U$, where n may be infinite.

Example There are six possible outcomes in throwing a die: the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. The event A of throwing an odd number—the compound event consisting of outcomes $\{1\}$, $\{3\}$ or $\{5\}$—corresponds to $A = \{1, 3, 5\}$, while for even numbers $B = \{2, 4, 6\}$. The sum of A and B exhausts the whole sample space; $A \cup B = S = U$ implies a certain event. The event of throwing a 7 is impossible: it is not contained in the sample space at all. ◁

Example A coin is tossed twice, yielding either head (h) or tail (t) in each toss. The sample space of this random experiment is $S = \{\text{hh, ht, th, tt}\}$. Let A represent the event that in two tosses we get at least one head, $A = \{\text{hh, ht, th}\}$, and let B represent the event that the second toss results in a tail, thus $B = \{\text{ht, tt}\}$. The event that at least one of A and B occurs (i.e. A or B or both) is

$$A \cup B = \{\text{hh, ht, th, tt}\}.$$

We got $A \cup B = S$ but that does not hold in general: if, for example, one would demand event B to yield two heads, $B = \{\text{hh}\}$, one would obtain $A \cup B = \{\text{hh, ht, th}\} = A$. The event of A and B occurring simultaneously is

$$A \cap B = AB = \{\text{ht}\}.$$

This implies that A and B are *not* exclusive, otherwise we would have obtained $AB = \{\,\} = \varnothing$. The event that A occurs but B does *not* occur is

$$A - B = A \cap \bar{B} = \{\text{hh}, \text{ht}, \text{th}\} \cap \{\text{hh}, \text{th}\} = \{\text{hh}, \text{th}\}.$$

The complementary event to A is $\bar{A} = S - A = \{\text{tt}\}$. ◁

The sample spaces in the above examples are discrete. An illustration of a continuous one can be found in thermal implantation of ions into quartz (SiO_2) in the fabrication of chips. The motion of ions in the crystal is diffusive and the ions penetrate to different depths: the sample space for the depths over which a certain concentration profile builds up is, say, the interval $S = [0, 1]\,\mu\text{m}$.

1.2 Basic Combinatorics

1.2.1 Variations and Permutations

We perform m experiments, of which the first has n_1 possible outcomes, the second has n_2 outcomes for each outcome of the first, the third has n_3 outcomes for each outcome of the first two, and so on. The number of possible outcomes of all m experiments is

$$n_1 n_2 n_3 \ldots n_m.$$

If $n_i = n$ for all i, the number of all possible outcomes is simply

$$n^m.$$

Example A questionnaire contains five questions with three possible answers each, and ten questions with five possible answers each. In how many ways the questionnaire can be filled out if exactly one answer is allowed for each question? By the above formulas, in no less than $3^5 5^{10} = 2373046875$ ways. ◁

What if we have n different objects and are interested in how many ways (that is, *variations*) m objects from this set can be reshuffled, paying attention to their *ordering*? The first object can be chosen in n ways. Now, the second one can only be chosen from the reduced set of $n - 1$ objects, ..., and the last object from the remaining $n - m + 1$. The number of variations is then

$$n(n-1)\cdots(n-m+1) = \frac{n!}{(n-m)!} = {}_nV_m = (n)_m. \tag{1.3}$$

The symbol on the right is known as the Pochammer symbol.

Example The letters A, B, C and D ($n = 4$) can be assembled in groups of two ($m = 2$) in $4!/2! = 12$ ways: {AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, DC}. Note that in this procedure, ordering is crucial: AB does not equal BA. ◁

A special case of (1.3) is $m = n$ when variations are called *permutations:* the number of permutations of n objects is

$$n(n-1)(n-2)\cdots 3{\cdot}2{\cdot}1 = n! = P_n.$$

Speaking in reverse, $n!$ is the number of all permutations of n objects, while (1.3) is the number of *ordered* sub-sequences of length m from these n objects.

Example We would like to arrange ten books (four physics, three mathematics, two chemistry books and a dictionary) on a shelf such that the books from the same field remain together. For each possible arrangement of the fields we have $4!\,3!\,2!\,1!$ options, while the fields themselves can be arranged in $4!$ ways, hence there are a total of $1!\,2!\,3!\,4!\,4! = 6912$ possibilities. ◁

We are often interested in the permutations of n objects, n_1 of which are of one kind and indistinguishable, n_2 of another kind ..., n_m of the mth kind, while $n = n_1 + n_2 + \cdots + n_m$. From all $n!$ permutations the indistinguishable ones $n_1!$, $n_2!$ ... must be removed, hence the required number of permutations is $n!/(n_1!\,n_2!\cdots n_m!)$ and is denoted by the *multinomial symbol:*

$$\frac{n!}{n_1!\,n_2!\ldots n_m!} = {}_nP_{n_1,n_2,\ldots,n_m} = \binom{n}{n_1, n_2, \ldots, n_m}. \tag{1.4}$$

1.2.2 Combinations Without Repetition

In how many ways can we arrange n objects into different groups of m objects if the ordering is irrelevant? (For example, the letters A, B, C, D and E in groups of three.) Based on previous considerations leading to (1.3) we would expect $n(n-1)\cdots(n-m+1)$ variations. But in doing this, equal groups would be counted multiple ($m!$) times: the letters A, B and C, for example, would form $m! = 3! = 6$ groups ABC, ACB, BAC, BCA, CAB and CBA, in which the letters are just mixed. Thus the desired number of arrangements—in this case called *combinations of mth order among n elements without repetition*—is

$$\frac{n(n-1)\cdots(n-m+1)}{m!} = \frac{n!}{(n-m)!\,m!} = {}_nC_m = \frac{{}_nV_m}{P_m} = \binom{n}{m}. \tag{1.5}$$

The symbol at the extreme right is called the *binomial symbol*. It can not hurt to recall its parade discipline, the *binomial formula*

$$(x+y)^n = \sum_{m=0}^{n} \binom{n}{m} x^{n-m} y^m. \tag{1.6}$$

1.2.3 Combinations with Repetition

In combinations *with repetition* we allow the elements to appear multiple times, for example, in combining four letters (A, B, C and D) into groups of three, where not only triplets with different elements like ABC or ABD, but also the options AAA, AAB and so on should be counted. The following combinations are allowed:

AAA, AAB, AAC, AAD, ABB, ABC, ABD, ACC, ACD, ADD,
BBB, BBC, BBD, BCC, BCD, BDD, CCC, CCD, CDD, DDD.

In general the *number of combinations of* m*th order among* n *elements with repetition* is

$$\frac{(n+m-1)!}{(n-1)!\,m!} = \binom{n+m-1}{m}. \tag{1.7}$$

In the example above ($n = 4$, $m = 3$) one indeed has $6!/(3!\,3!) = 20$.

1.3 Properties of Probability

A random experiment always leaves us in doubt whether an event will occur or not. A measure of probability with which an event may be expected to occur is its relative frequency. It can be calculated by applying "common sense", i.e. by dividing the number of chosen ("good") events A to occur, by the number of all encountered events: in throwing a die there are six possible outcomes, three of which yield odd numbers, so the relative frequency of the event A = "odd number of points" should be $P(A) = \text{good}/\text{all} = 3/6 = 0.5$. One may also proceed pragmatically: throw the die a thousand times and count, say, 513 odd and 487 even outcomes. The empirical relative frequency of the odd result is therefore $513/1000 = 0.513$. Of course this value will fluctuate if a die is thrown a thousand times again, and yet again—to 0.505, 0.477, 0.498 and so on. But we have reason to believe that after many, many trials the value will stabilize at the previously established value of 0.5.

We therefore define the probability $P(A)$ of event A in a random experiment as the value at which the relative frequency of A usually stabilizes after the experiment has been repeated many times[1] (see also Appendix A). Obviously

$$0 \le P(A) \le 1.$$

The probability of a certain event is one, $P(U) = 1$. For any event A we have

$$P(A) + P(\bar{A}) = 1,$$

hence also $P(\emptyset) = 1 - P(U) = 0$: the probability of an impossible event is zero. For arbitrary events A and B the following relation holds:

$$P(A \cup B) = P(A) + P(B) - P(AB). \tag{1.8}$$

For exclusive events, $AB = \emptyset$ and the equation above reduces to

$$P(A + B) = P(A) + P(B),$$

which can be generalized for pair-wise exclusive events as

$$P(A \cup B \cup C \cup \cdots) = P(A) + P(B) + P(C) + \cdots .$$

To generalize (1.8) to multiple events one only needs to throw a glance at (1.1): for example, with three events A, B and C we read off

$$\begin{aligned} A \cup B \cup C &= A + \bar{A}B + \bar{A}\bar{B}C \\ &= A + (U - A)B + (U - A)(U - B)C \\ &= A + B + C - AB - AC - BC + ABC, \end{aligned}$$

therefore also

$$\begin{aligned} &P(A \cup B \cup C) \\ &\quad = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC). \end{aligned} \tag{1.9}$$

Example (Adapted from [3].) In the semiconductor wafer production impurities populate the upper layers of the substrate. In the analysis of 1000 samples one finds a large concentration of impurities in 113 wafers that were near the ion source during the process, and in 294 wafers that were at a greater distance from it. A low concentration is found in 520 samples from near the source and 73 samples that were farther away. What is the probability that a randomly selected wafer was near the source during the production (event N), or that it contains a large concentration of impurities (event L), or both?

[1] This is the so-called *frequentist approach* to probability, in contrast to the *Bayesian approach:* an introduction to the latter is offered by [2].

We can answer the question by carefully counting the measurements satisfying the condition: $P(N \cup L) = (113 + 294 + 520)/1000 = 0.927$. Of course, (1.8) leads to the same conclusion: the probability of N is $P(N) = (113 + 520)/1000 = 0.633$, the probability of L is $P(L) = (113 + 294)/1000 = 0.407$, while the probability of N and L occurring simultaneously—they are not exclusive!—is $P(NL) = 113/1000 = 0.113$, hence

$$P(N \cup L) = P(N) + P(L) - P(NL) = 0.633 + 0.407 - 0.113 = 0.927.$$

Ignoring the last term, $P(NL)$, is a frequent mistake which, however, is easily caught as it leads to probability being greater than one. ◁

Example (Adapted from [4].) A detector of cosmic rays consists of nine smaller independent sub-detectors all pointing in the same direction of the sky. Suppose that the probability for the detection of a cosmic ray shower (event E) by the individual sub-detector—the so-called detection efficiency—is $P(E) = \varepsilon = 90\%$. If we require that the shower is seen by all sub-detectors simultaneously (nine-fold coincidence, Fig. 1.1 (left)), the probability to detect the shower (event X) is

$$P(X) = \bigl(P(E)\bigr)^9 \approx 0.387.$$

The sub-detectors can also be wired in three triplets, where a favorable outcome is defined by at least one sub-detector in the triplet observing the shower. Only then a

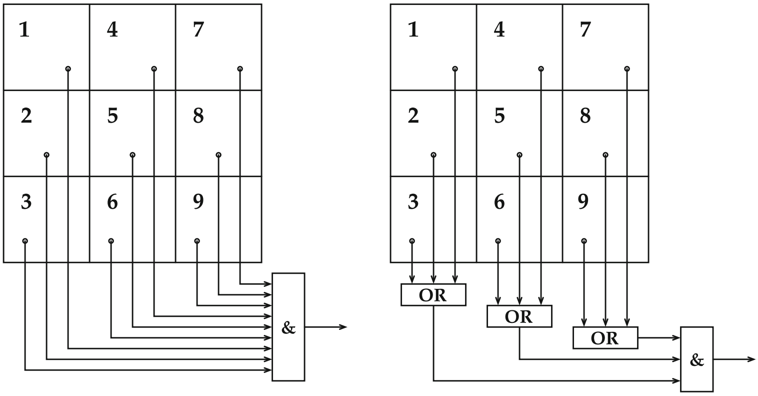


Fig. 1.1 Detector of cosmic rays. [Left] Sub-detectors wired in a nine-fold coincidence. [Right] Triplets of sub-detectors wired in a three-fold coincidence

triple coincidence is formed from the three resulting signals (Fig. 1.1 (right)). In this case the total shower detection probability is

$$P(X) = \big(P(E_1 \cup E_2 \cup E_3)\big)^3 = \big(3\varepsilon - 3\varepsilon^2 + \varepsilon^3\big)^3 \approx 0.997,$$

where we have used (1.9). ◁

1.4 Conditional Probability

Let A be an event in a random experiment (call it 'first') running under a certain set of conditions, and $P(A)$ its probability. Imagine another event B that may occur in this or another experiment. What is the probability $P'(A)$ of event A if B is interpreted as an additional condition for the first experiment? Because event B modifies the set of conditions, we are now actually performing a new experiment differing from the first one, thus we generally expect $P(A) \neq P'(A)$. The probability $P'(A)$ is called the *conditional probability* of event A *under the condition* B or *given event* B, and we appropriately denote it by $P(A|B)$. This probability is easy to compute: in n repetitions of the experiment with the augmented set of conditions B occurs n_B times, while $A \cap B$ occurs n_{AB} times, therefore

$$P(A|B) = \lim_{n\to\infty} \frac{n_{AB}/n}{n_B/n} = \frac{P(AB)}{P(B)}.$$

The conditional probability for A given B ($P(B) \neq 0$) is therefore computed by dividing the probability of the simultaneous event, $A \cap B$, by $P(B)$. Obviously, the reverse is also true:

$$P(B|A) = \frac{P(AB)}{P(A)}.$$

Both relations can be merged into a single statement known as the *theorem on the probability of the product of events* or simply the *product formula:*

$$P(AB) = P(B|A)P(A) = P(A|B)P(B). \tag{1.10}$$

The first part of the equation can be verbalized as follows: the probability that A and B occur simultaneously equals the product of probabilities that A occurs first, and the probability that B occurs, given that A has already occurred. (The second part proceeds analogously.)

The theorem can be generalized to multiple events. Let $A_1, A_2, \ldots, A_n$ be arbitrary events and let $P(A_1A_2\ldots A_n) > 0$. Then

$$P(A_1A_2\ldots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\ldots P(A_n|A_1A_2\ldots A_{n-1}). \tag{1.11}$$

Perhaps the essence becomes even clearer if we reverse the ordering of the factors and digest the formula from right to left:

$$P(A_n \ldots A_2 A_1) = P(A_n | A_{n-1} \ldots A_2 A_1) \ldots P(A_3 | A_2 A_1) P(A_2 | A_1) P(A_1).$$

Example What is the probability that throwing a die yields a number of spots which is less than four *given that* the number is odd? Let A mean "odd number of spots" ($P(A) = 1/2$), and B "the number of spots less than four" ($P(B) = 1/2$). If A and B occur *simultaneously*, the probability of the compound event can be inferred from the intersection of sets A and B in Fig. 1.2 (left): it is

$$P(AB) = \tfrac{2}{6} = \tfrac{1}{3},$$

since only elements $\{1, 3\}$ inhabit the intersection, while the complete sample space is $\{1, 2, 3, 4, 5, 6\}$. But this is not yet the answer to our question! We are interested in the probability of B once A ("the condition") has already occurred: this implies that the sample space has shrunk to $\{1, 3, 5\}$ as shown in Fig. 1.2 (right). From this reduced space we need to pick the elements that fulfill the requirement B: they are $\{1, 3\}$ and therefore

$$P(B|A) = \tfrac{2}{3}.$$

Equation (1.10) says the same: $P(B|A) = P(AB)/P(A) = \frac{1}{3}\big/\frac{1}{2} = \frac{2}{3}$. We can imagine that the unconditional probability $P(B) = 1/2$ has increased to $P(B|A) = 2/3$ by the additional *information* that the throw yields an odd number. ◁

Example A box in our cellar holds 32 bottles of wine, eight of which are spoiled. We randomly select four bottles from the box for today's dinner. What is the probability that not a single one will be spoiled?

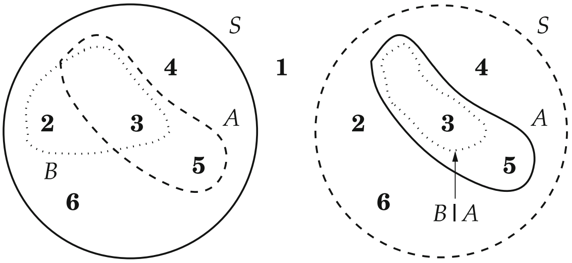


Fig. 1.2 The conditional probability in throwing a die. [Left] The probability of events A and B occurring *simultaneously* corresponds to the intersection of the sets $\{1, 3, 5\}$ and $\{1, 2, 3\}$ within the complete sample space S. [Right] The condition A first isolates the set $\{1, 3, 5\}$ from the complete S. The conditional probability of B given A corresponds to the fraction of the elements in this set that also fulfill the requirement B

This can be solved in two ways. The first method is to apply the product formula by considering that with each new bottle fetched from the box, both the total number of bottles and the number of spoiled bottles in it are reduced by one. Let A_i denote the event that the ith chosen bottle is good, and A the event that all four bottles are fine. The probability of the first bottle being good is $P(A_1) = 24/32$. This leaves 31 bottles in the box, 23 of which are good, hence the probability of the second bottle being intact is $P(A_2|A_1) = 23/31$. Analogously $P(A_3|A_1A_2) = 22/30$ and $P(A_4|A_1A_2A_3) = 21/29$ for the third and fourth bottle, respectively. Formula (1.11) then gives

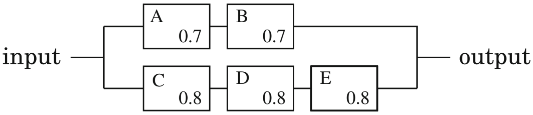
$$\begin{aligned} P(A) = P(A_1A_2A_3A_4) &= P(A_1)P(A_2|A_1)P(A_3|A_1A_2)P(A_4|A_1A_2A_3) \\ &= \frac{24}{32}\frac{23}{31}\frac{22}{30}\frac{21}{29} \approx 0.2955. \end{aligned}$$

The second option is to count the number of ways in which 24 *good* bottles can be arranged in four places: it is equal to $24!/(4!\,20!)$ (see (1.5)). But this number must be divided by the number of *all* possible combinations of bottles in four places, which is $32!/(4!\,28!)$. The probability of four bottles being good is then

$$P(A) = \frac{24!}{4!\,20!}\frac{4!\,28!}{32!} \approx 0.2955.$$

◁

Example An electric circuit has five independent elements with various degrees of reliability—probabilities that an element functions—shown in the figure.



What is the probability that the circuit works (transmits signals from input to output) and the probability that A does not work, given that the circuit works?

Let us denote the event "element A works" by A (and analogously for the elements B, C, D and E). The circuit works (event V) when the elements A and B work *or* the elements C, D and E work *or* all five of them work, hence

$$\begin{aligned} P(V) = P(AB \cup CDE) &= P(AB) + P(CDE) - P(ABCDE) \\ &= P(A)P(B) + P(C)P(D)P(E) \\ &\quad -P(A)P(B)P(C)P(D)P(E) \\ &= (0.7)^2 + (0.8)^3 - (0.7)^2(0.8)^3 = 0.75112, \end{aligned}$$

where we have used (1.8). The probability that A has failed (event $\bar{A}$), given that the circuit works, is obtained by the following consideration, noting that $XY = X \cap Y$. We first calculate the probability that A does not work while the circuit as a whole works. If A has failed, then the bottom branch of the circuit *must* work. But even if A is inoperational, *two* options remain for B: it either works or it does not. Thus $\bar{A} \cap V = \left[\left(\bar{A} \cap B\right) \cup \left(\bar{A} \cap \bar{B}\right)\right] \cap (C \cap D \cap E)$. Thus the conditional probability we have been looking for is

$$P\left(\bar{A}|V\right) = \frac{P\left(\bar{A}V\right)}{P(V)} = \frac{P\left[\left(\bar{A}B \cup \bar{A}\bar{B}\right)\left(CDE\right)\right]}{P(V)} = \frac{\left[P\left(\bar{A}B\right) + P\left(\bar{A}\bar{B}\right)\right] \cdot P(CDE)}{P(V)}$$
$$= \frac{[(1-0.7)0.7 + (1-0.7)^2] \cdot (0.8)^3}{0.75112} = 0.2045,$$

where we have used $\bar{A} \cap B \cap \bar{A} \cap \bar{B} = \{\,\}$, since $A \cap \bar{A} = B \cap \bar{B} = \{\,\}$. ◃

1.4.1 Independent Events

If events A and B are *independent*, the probability that A occurs (or does not occur) is independent of whether we have any information on B (and vice-versa), hence

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B).$$

According to (1.10), the probability that such events occur simultaneously equals the product of probabilities of them occurring individually:

$$P(AB) = P(A)P(B). \tag{1.12}$$

When more than two events are involved, independence must be defined more carefully. The events in the set

$$\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$$

are *mutually* or *completely independent* if, for every combination $(i_1, i_2, \ldots, i_k)$ of kth order without repetition ($k = 2, 3, \ldots, n$) among the numbers $1, 2, \ldots, n$, it holds that

$$P\left(A_{i_1}A_{i_2}\ldots A_{i_k}\right) = P\left(A_{i_1}\right)P\left(A_{i_2}\right)\ldots P\left(A_{i_k}\right). \tag{1.13}$$

When $k = n$ this system of equations has the form

$$P(A_1A_2\ldots A_n) = P(A_1)P(A_2)\ldots P(A_n),$$

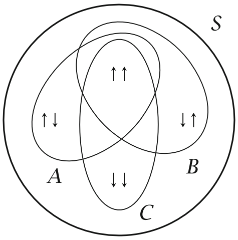
which is a special case of (1.11); when $k = 2$, the leftover of (1.13) is simply

$$P(A_iA_j) = P(A_i)P(A_j).$$

If (1.12) applies to any pair of events in $\mathcal{A}$, we say that such events are *pair-wise independent*, but this is still a far cry from *mutual* (complete) independence! There are 2^n combinations without repetition among n elements (see (1.6) with $x = y = 1$). One of them corresponds to the empty set, while there are n combinations of the first order, as we learn from (1.5). The system above therefore imposes $2^n - n - 1$ conditions that must be fulfilled by the events from $\mathcal{A}$ in order for them to be mutually independent. In the special case $n = 3$ there are four such conditions:

$$\begin{aligned} P(A_1A_2) &= P(A_1)P(A_2), \\ P(A_1A_3) &= P(A_1)P(A_3), \\ P(A_2A_3) &= P(A_2)P(A_3), \\ P(A_1A_2A_3) &= P(A_1)P(A_2)P(A_3). \end{aligned}$$

This important distinction between pair-wise and mutual independence is discussed in the following Example.



Example The spin in a quantum system can have two projections: $+\frac{1}{2}$ (spin "up", ↑) or $-\frac{1}{2}$ (spin "down", ↓). The orientation of the spin is measured twice in a row. We make the following event assignments: event A means "spin ↑ in the first measurement", event B is "spin ↑ in the second measurement", and event C is "both measurements show the same projection". The sample space for the measured pairs of orientations is $S = \{\uparrow\uparrow, \uparrow\downarrow, \downarrow\uparrow, \downarrow\downarrow\}$, while the chosen three events correspond to its subsets $A = \{\uparrow\uparrow, \uparrow\downarrow\}$, $B = \{\uparrow\uparrow, \downarrow\uparrow\}$ and $C = \{\uparrow\uparrow, \downarrow\downarrow\}$ shown in the Figure. We immediately obtain the probabilities

$$P(A) = P(B) = P(C) = \tfrac{2}{4} = \tfrac{1}{2},$$

as well as

$$P(AB) = P(AC) = P(BC) = \tfrac{1}{4} \quad \text{and} \quad P(ABC) = \tfrac{1}{4}.$$

Since

$$P(AB) = P(A)P(B) = P(AC) = P(A)P(C) = P(BC) = P(B)P(C),$$

events A, B and C are pair-wise independent. On the other hand,

$$P(ABC) = \tfrac{1}{4} \neq \tfrac{1}{8} = P(A)P(B)P(C),$$

so the events are *not* mutually independent. ◁

1.4.2 Bayes Formula

When an event A occurs under different, mutually exclusive conditions, and we know the conditional probabilities of A given all these conditions, we can also calculate the unconditional probability of A. The two-condition case is illustrated by the following classic insurance-company example.

Example An insurance company classifies the drivers into those deemed less (85%) and those more accident-prone (15%). These are two mutually exclusive 'conditions'—call them B and $\bar{B}$—that exhaust all options, as there is no third class, thus $P(B) = 0.85$, $P(\bar{B}) = 0.15$. On average, a first-class driver causes a crash every 10 years, and the second-class driver once in 5 years. Let A denote the event of an accident, regardless of its cause. The probability for a first-tier driver to cause a crash within a year is $P(A|B) = 1/10$, while it is $P(A|\bar{B}) = 1/5$ for the second-tier driver. What is the probability that a new customer will cause an accident within the first year? Since for any A and B, $A = (A \cap B) \cup (A \cap \bar{B})$, we also have

$$P(A) = P(AB) + P(A\bar{B}),$$

while from (1.10) it follows that

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}). \tag{1.14}$$

Statistically, the company may therefore expect the probability of

$$P(A) = 0.1 \cdot 0.85 + 0.2 \cdot 0.15 = 0.115$$

for a newly insured driver to cause an accident within a year. ◁

Equation (1.14) is a sort of weighted average over both driver classes, where the weights depend on conditions B and $\bar{B}$. Suppose that there are more such mutually exclusive conditions: we then prefer to call them *assumptions* or *hypotheses* and denote them by H_i: we have H_1 or H_2 … or H_n, exhausting all possibilities. The set of all H_i constitutes a complete set defined by (1.2), hence

$$P(A) = P(AH_1) + P(AH_2) + \cdots + P(AH_n).$$

Applying the left-hand side of (1.10) to each term separately yields the so-called *total probability formula*

$$P(A) = P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \cdots + P(A|H_n)P(H_n), \tag{1.15}$$

illustrated in Fig. 1.3.

Let us recall (1.10) once more, this time in its second part, whence one reads off $P(H_i|A)P(A) = P(A|H_i)P(H_i)$ or

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{P(A)}.$$

The denominator of this expression is given by (1.15) and the final result is the famous Bayes formula [5]

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{P(A|H_1)P(H_1) + \cdots + P(A|H_n)P(H_n)}, \quad i = 1, 2, \ldots, n. \tag{1.16}$$

A random experiment may repeatedly yield events A, but the events H_i conditioning A—with corresponding probabilities $P(H_i)$—occurred *prior to* A. The quantities $P(H_i)$ are therefore called *prior probabilities* since they are, in principle, known in advance. In contrast, the left side of the Bayes formula gives the probability that the hypothesis H_i is valid with respect to the *later* outcome A. The conditional probability $P(H_i|A)$ is called *posterior*, since it uses the present outcome A to specify the probability of H_i occurring prior to A. This is why the Bayes formula is also known as the *theorem on probability of hypotheses.*

Example A company decides to manufacture cell-phones by using processor chips of different suppliers. The first type of chip is built into 70%, the second into 20%, and the third into 10% of devices. A randomly chosen device contains chip i (event C_i) with probability $P(C_i)$, where $P(C_1) = 0.7$, $P(C_2) = 0.2$ and $P(C_3) = 0.1$: these are the known prior probabilities. Some chips are unreliable, causing the devices to malfunction. The probability that a cell-phone breaks down (event A), given

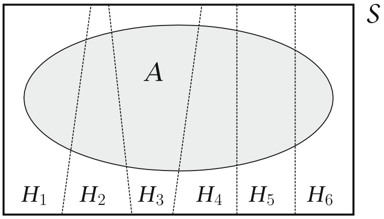


Fig. 1.3 Illustration of the total probability formula. Mutually exclusive conditions or hypotheses H_i are disjoint sets that partition the sample space S and therefore also an arbitrary event A

that it contains chip i, is $P(A|C_i)$. The manufacturer establishes $P(A|C_1) = 0.01$, $P(A|C_2) = 0.03$ and $P(A|C_3) = 0.05$.

We go to a store and buy a cell-phone made by this company. It breaks down immediately (event A at this very moment). What is the probability that it was manufactured (event C_i in the past) in the factory installing type-i chips ($i = 1, 2, 3$)? We are looking for the posterior probabilities $P(C_i|A)$ given by the Bayes formula. Its denominator contains $P(A) = \sum_{i=1}^{3} P(A|C_i)P(C_i) = 0.01 \cdot 0.7 + 0.03 \cdot 0.2 + 0.05 \cdot 0.1 = 0.018$, which is common to all three cases—and this is the probability that the cell-phone breaks down. This leads to

$$\begin{aligned} P(C_1|A) &= \frac{P(A|C_1)P(C_1)}{P(A)} = \frac{0.01 \cdot 0.7}{0.018} \approx 38.9\%, \\ P(C_2|A) &= \frac{P(A|C_2)P(C_2)}{P(A)} = \frac{0.03 \cdot 0.2}{0.018} \approx 33.3\%, \\ P(C_3|A) &= \frac{P(A|C_3)P(C_3)}{P(A)} = \frac{0.05 \cdot 0.1}{0.018} \approx 27.8\%. \end{aligned}$$

Of course we also have $P(C_1|A) + P(C_2|A) + P(C_3|A) = 1$. ◁

1.5 Problems

1.5.1 Boltzmann, Bose–Einstein and Fermi–Dirac Distributions

(Adapted from [1].) Imagine a system of n particles in which the state of each particle is described by p values (components of the position vector or linear momentum, spin quantum number, and so on). Each particle state can be represented by such a p-plet, which is a point in p-dimensional space. The state of the whole system is uniquely specified by a n-plet of such points.

Let us divide the phase space into N ($N \geq n$) cells. The state of the system is described by specifying the distribution of states among the cells. We are interested in the probability that a given cell is occupied by the prescribed number of particles. Consider three options: ① The particles are distinguishable, each cell can be occupied by an arbitrary number of particles, and all such distributions are equally probable. We then say that the particles "obey" Boltzmann statistics: an example of such a system are gas molecules. ② The particles are *indistinguishable*, but the cells may still be occupied by arbitrary many particles and all such distributions are equally probable. This is the foundation of Bose–Einstein statistics obeyed by particles with integer spins (bosons), e.g. photons. ③ The particles are indistinguishable, each cell may accommodate *at most one particle* due to the Pauli principle [6]. All distributions are equally probable. This case refers to the Fermi–Dirac statistics applicable to particles with half-integer spins (fermions), e.g. electrons, protons and neutrons.

✎ Let A_k be the event that there are precisely k particles ($0 \le k \le n$) in a certain cell, regardless of their distribution in other cells. ① Each of the n particles can be put into any of the N cells, even if other particles are already sitting there. All particles can therefore be arranged in N^n ways and this is the number of all possible outcomes. How many correspond to event A_k? Into the chosen cell one can pour k particles in $\binom{n}{k}$ ways, while the remaining $n-k$ particles can be arranged into the other $N-1$ cells in $(N-1)^{n-k}$ ways. Event A_k therefore accommodates $\binom{n}{k}(N-1)^{n-k}$ outcomes, thus

$$P(A_k) = \binom{n}{k}(N-1)^{n-k}\frac{1}{N^n} = \binom{n}{k}\left(\frac{1}{N}\right)^k\left(1-\frac{1}{N}\right)^{n-k}.$$

② Since particles are indistinguishable and each cell is allowed to swallow an arbitrary number of particles, the number of all possible distributions equals the number of combinations of nth order among N elements with repetition (1.7), i.e. $\binom{N+n-1}{n}$. How many are acceptable for A_k? Event A_k occurs precisely when k particles are selected for a given cell—since they are indistinguishable, this can be accomplished in one way only—while the remaining $n-k$ are distributed among $N-1$ cells, which amounts to combinations of order $n-k$ among $N-1$ elements with repetition, i.e. $\binom{N+n-k-2}{n-k}$. It follows that

$$P(A_k) = \binom{N+n-k-2}{n-k}\bigg/\binom{N+n-1}{n}.$$

③ Since at most one particle is allowed to occupy any single cell, all possible distributions can be counted by choosing n cells out of N and putting one particle into every one of them: this can be accomplished in $\binom{N}{n}$ ways. How many of them correspond to event A_k? For $k > 1$ there are none, while for $k = 0$ or $k = 1$ there are as many ways as one can arrange $n-k$ particles over $N-1$ cells, which is $\binom{N-1}{n-k}$. Therefore

$$P(A_k) = \binom{N-1}{n-k}\bigg/\binom{N}{n} = \begin{cases} 1-\dfrac{n}{N} \ ; & k=0, \\ \dfrac{n}{N} \ ; & k=1, \end{cases}$$

while $P(A_k) = 0$ for $k > 1$. Figure 1.4 (left) shows the probabilities $P(A_k)$ for all three distributions in the case $N = 15$, $n = 5$, while Fig. 1.4 (right) shows the Boltzmann and the Bose–Einstein distribution in the case $N = 100$, $n = 10$.

1.5.2 Blood Types

The fractions of blood types O, A, B and AB in the whole population are

$$\text{O}: \ 44\%, \quad \text{A}: \ 42\%, \quad \text{B}: \ 10\%, \quad \text{AB}: \ 4\%.$$

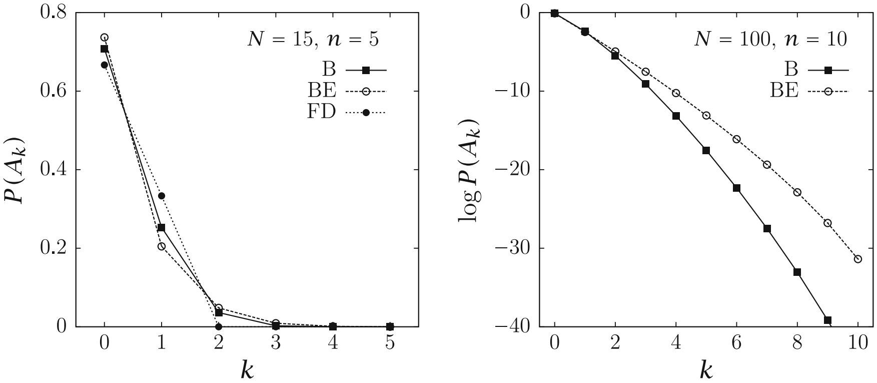


Fig. 1.4 The probability of finding k particles in any chosen cell of a N-cell phase space flooded with n particles, in the case of Boltzmann (B), Bose–Einstein (BE) and Fermi–Dirac (FD) statistics. [Left] $N = 15, n = 5$. The sum of all probabilities within a given distribution of course equals 1, as it is obvious e.g. in the case of the Fermi–Dirac distribution: $P(A_0) = 1 - \frac{5}{15} = \frac{2}{3}, P(A_1) = \frac{5}{15} = \frac{1}{3}$. [Right] $N = 100, n = 10$

① Two persons are picked at random from the population. What is the probability of their having the same blood type, and what is the probability that their types differ? ② We pick four people from the same population. What is the probability that precisely k ($k = 1, 2, 3, 4$) blood types will be found among them?

✎ Let us replace the letter notation O, A, B, AB by indices 1, 2, 3, 4, and let P_i denote the probability that a person has blood type i ($i = 1, 2, 3, 4$). ① All possible pairs are $\{i, i\}$, $i = 1, 2, 3, 4$, each having probability P_i^2, therefore $P = P_1^2 + P_2^2 + P_3^2 + P_4^2 = 0.44^2 + 0.42^2 + 0.1^2 + 0.04^2 = 0.3816$. The complementary event has probability $1 - P = 0.6184$ which, in a more arduous manner, can be computed as:

$$\begin{aligned}1 - P &= P_1(P_2 + P_3 + P_4) + P_2(P_1 + P_3 + P_4) + P_3(P_1 + P_2 + P_4)\\ &\quad + P_4(P_1 + P_2 + P_3)\\ &= 2\big[P_1(P_2 + P_3 + P_4) + P_2(P_3 + P_4) + P_3P_4\big] = 0.6184.\end{aligned}$$

② Let $P(k)$ denote the probability that precisely k blood types will be found in the chosen four. For $k = 1$ the quartets are $\{i, i, i, i\}$, $i = 1, 2, 3, 4$, hence $P(1) = \sum_i P_i^4 = 0.44^4 + 0.42^4 + 0.1^4 + 0.04^4 = 0.0687$. For $k = 2$ we use (1.4) to obtain the number of possible combinations in samples of the form $\{i, j, j, j\}$ ($i \neq j$), which is $N_{13} = 4!/(1!\,3!) = 4$, and the number of combinations in samples of the form $\{i, i, j, j\}$ ($i \neq j$), which is $N_{22} = 4!/(2!\,2!) = 6$. We get

$$P(2) = N_{13} \sum_{i \neq j} P_i^1 P_j^3 + N_{22} \sum_{i<j} P_i^2 P_j^2 = 0.3665 + 0.2308 = 0.5973.$$

The calculation for $k = 3$ is tedious and is best avoided by calculating the probability for $k = 4$, which is $P(4) = 4! \cdot P_1 P_2 P_3 P_4 = 0.0177$, and accumulating all previously computed $P(k)$ into the complementary event: $P(3) = 1 - P(1) - P(2) - P(4) = 0.3163$.

1.5.3 Independence of Events in Particle Detection

Two detectors are used to detect charged particles with different parities (mirror symmetries of their wave-functions): pions (π^+ and π^-) and kaons (K^+ and K^-), all possessing negative parity, as well as protons (p), deuterons (d) and ^{3}He and ^{4}He nuclei, all of which have positive parities. Assume that all particles appear with equal frequencies and assign indices $\{1, 2, 3, 4, 5, 6, 7, 8\}$ to types $\{\pi^+, \pi^-, K^+, K^-, p, d, {}^3\mathrm{He}, {}^4\mathrm{He}\}$. Let A denote the event that the first detector has seen a negative-parity particle. Let B denote the event that the second detector has detected a positive-parity particle, and suppose that

$$P(A) = P(A|B) = \tfrac{4}{8} = \tfrac{1}{2},$$
$$P(B) = P(B|A) = \tfrac{4}{8} = \tfrac{1}{2}.$$

Let C denote the event that both detectors observe particles with equal parities. Are events A, B and C (pair-wise or mutually) independent?

✎ There are 64 equally probable outcomes (i, j) in an experiment where the first and second detector detect particles i and j, respectively; 16 of them are pion-kaon combinations fulfilling condition C:

$$(1,1),\ (1,2),\ (1,3),\ (1,4),\ (2,1),\ (2,2),\ (2,3),\ (2,4),$$
$$(3,1),\ (3,2),\ (3,3),\ (3,4),\ (4,1),\ (4,2),\ (4,3),\ (4,4),$$

as well as 16 combinations of atomic nuclei,

$$(5,5),\ (5,6),\ (5,7),\ (5,8),\ (6,5),\ (6,6),\ (6,7),\ (6,8),$$
$$(7,5),\ (7,6),\ (7,7),\ (7,8),\ (8,5),\ (8,6),\ (8,7),\ (8,8),$$

thus $P(C) = (16 + 16)/64 = \frac{1}{2}$. Suppose that the first detector has seen a negative-parity particle and has thereby imposed condition A: then C occurs if the second detector also reports a negative-parity particle (probability $1/2$), implying $P(C|A) = 1/2$. Analogously we conclude $P(C|B) = 1/2$, and finally

$$P(C) = P(C|A) = P(C|B) = \tfrac{1}{2}.$$

We conclude that A, B and C are pair-wise but not mutually independent since $P(ABC) = P(A)P(B)P(C)$ does not hold true. Our calculation shows that $P(A)P(B)P(C) = \frac{1}{8}$, while $A \cap B \cap C$ is an impossible event: if there is a negative-parity particle in the first detector and a positive-parity particle in the second one, we can not have the same parity in both detectors, thus $P(ABC) = 0$.

How do these considerations change if the detectors are inefficient in detecting heavier nuclei (^{3}He and ^{4}He)? Do events A, B and C remain independent? How does the result change in physically more sensible circumstances in which the number of pions exceeds the number of kaons by a factor of 100?

1.5.4 Searching for the Lost Plane

The authorities believe that an airliner has been lost in one of the three regions R_i $(i = 1, 2, 3)$ in which the crash has occurred with equal probability, $P(R_i) = 1/3$. Let P_i denote the probability that the plane search in region i will locate the plane that actually does lie in i. Calculate the conditional probability that the plane crashed in region i, given that the search in region 1 was unsuccessful!

✎ Let R_i $(i = 1, 2, 3)$ denote the event that the plane went down in region i, and N the event that the search in region 1 was unsuccessful. Bayes formula for $i = 1$ gives

$$\begin{aligned} P(R_1|N) &= \frac{P(NR_1)}{P(N)} = \frac{P(N|R_1)P(R_1)}{\sum_{i=1}^{3} P(N|R_i)P(R_i)} \\ &= \frac{(1-P_1)\frac{1}{3}}{(1-P_1)\frac{1}{3} + 1\cdot\frac{1}{3} + 1\cdot\frac{1}{3}} = \frac{1-P_1}{3-P_1}, \end{aligned}$$

while for $i = 2$ and $i = 3$ one gets

$$P(R_i|N) = \frac{P(N|R_i)P(R_i)}{P(N)} = \frac{1\cdot\frac{1}{3}}{(1-P_1)\frac{1}{3} + 1\cdot\frac{1}{3} + 1\cdot\frac{1}{3}} = \frac{1}{3-P_1}, \quad i = 2, 3,$$

where we have exploited the fact that the search in region 1 *must* be unsuccessful if the plane lies in region 2 or 3, hence $P(N|R_2) = P(N|R_3) = 1$. For example, if $P_1 = 0.7$, the probability that the plane is in region 1—given that it has *not* been found in it—is $0.3/2.3 \approx 13\%$. Note that $\sum_i P(R_i|N) = 1$.

1.5.5 The Monty Hall Problem ⋆

In the Monty Hall TV show with three boxes (adapted from [7, 8]) one box contains the car keys while the remaining boxes are empty. When the contestant picks one of the boxes (e.g. box 1), Monty Hall (MH) tells her: “I’ll make you a favor and open one of the remaining boxes that *does not contain the keys* (e.g. 2). Thus the keys are either in your chosen box or in box 3, so the probability of your winning the car has increased from 1/3 to 1/2.” The contestant (C) responds: “I’ve changed my mind. I prefer to pick box 3 instead of box 1.”

① Is Monty’s claim correct? What is the probability of the contestant winning the car if she changes her mind following Monty’s disclosure, and what is her chance of winning if she insists on her initial choice? ② Suppose that the contestant has been playing this game for a long time and knows that different boxes have different probabilities of containing the keys, e.g. 50, 40 and 10% for boxes 1, 2 and 3. What is the most promising strategy in this case?

✎ Two observations are crucial: MH *knows* which box contains the keys and obviously does not wish to reveal it; he opens one of the two remaining boxes at random and with equal probability. The answer to ① can then be obtained by simple counting of possible outcomes shown in Table 1.1: ‘W’ means that the contestant ‘wins’, ‘L’ means ‘loses’. (All information is contained in the first three rows since the rest consists just of cyclic permutations.) The probability of C winning the car when insisting on the initial choice is 1/3. The probability of winning the car after having changed her mind is 2/3. Consequently, Monty’s claim is false.

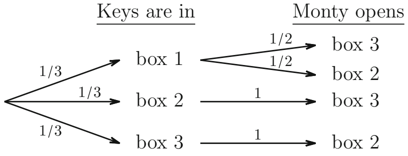
The problem can be approached from another, more intuitive viewpoint [9]. Suppose C decides to *always* switch. If she chooses an empty box, she can not lose: MH is then obliged to open the other empty box, so, by switching, C gets the only remaining box—the one containing the keys. C loses only if she initially chooses the box with the keys. Whether this strategy of “perpetual switching” works depends only on the initial choice of the empty box (probability 2/3) or the box containing the keys (probability 1/3).

Table 1.1 Possible outcomes in the Monty Hall contest

Keys are in	C picks	MH opens	Outcome	C switches	Outcome
1	1	2 or 3	W	1 for 3 or 2	L
1	2	3	L	2 for 1	W
1	3	2	L	3 for 1	W
2	1	3	L	1 for 2	W
2	2	1 or 3	W	2 for 3 or 1	L
2	3	1	L	3 for 2	W
3	1	2	L	1 for 3	W
3	2	1	L	2 for 3	W
3	3	1 or 2	W	3 for 2 or 1	L

Both contestant’s strategies are shown: “C picks” means the one and only choice of the box, while “C switches” means that the contestant selects a different box after Monty’s disclosure

Conditional probability offers yet another vantage point. Suppose that C chooses box 1 while the keys are in box 2 (event A, $P(A) = 1/3$). MH opens box 3 (event B). The graph



then tells us that $P(B|A) = 1$ and $P(B) = \frac{1}{3}\frac{1}{2} + \frac{1}{3}1 = \frac{1}{2}$, hence, by Bayes formula,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

In two thirds of the cases the keys are in the remaining box, so C doubles her $1/3$ chance of winning by switching. The same conclusion can be reached by analyzing the sample space in which the events are not equally probable. Denote all possible outcomes by (i, j), where i is the box containing the keys, j is the box opened by MH, and let P_{ij} denote the corresponding probability for such an outcome. When we shall later become familiar with the concept of random variables, all these values will be merged into the expression

$$X \sim \begin{pmatrix} (i,j) & \cdots \\ P_{ij} & \cdots \end{pmatrix} = \begin{pmatrix} (1,3) & (2,3) & (1,2) & (3,2) \\ 1/6 & 1/3 & 1/6 & 1/3 \end{pmatrix} \tag{1.17}$$

which we shall read as: “The discrete variable X is distributed such that the probability of outcome $(1, 3)$ is $P_{13} = 1/6$, the probability of outcome $(2, 3)$ is $P_{23} = 1/3$, and so on.” In a compact manner, however, we can write down the sample space with the probability values attached as subscripts:

$$S = \big\{(1,3)_{1/6},\ (2,3)_{1/3},\ (1,2)_{1/6},\ (3,2)_{1/3}\big\}.$$

In this notation, $A = \big\{(2,3)_{1/3}\big\}$ and $B = \big\{(1,3)_{1/6}, (2,3)_{1/3}\big\}$. Since $A \subset B$, we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{6}} = \frac{2}{3}.$$

Monty’s tempting offer to increase the contestant’s probability of winning to $1/2$ is based on the wrong assumption that the remaining two possible events $(1, 3)$ and $(2, 3)$ are equally probable—i.e. that the sample space after the condition A has been imposed is $\big\{(1,3)_{1/2},\ (2,3)_{1/2}\big\}$—leading to the wrong result $P(A|B) = 1/2$.

The best strategy for ② is: C should choose the *least* probable box (box 3); when MH reveals an empty box, C should switch. In this case C will win 90% of the time.

Table 1.2 Conditional probabilities in a diagnostic test that can be negative when the disease is absent (specificity $\mathcal{R}$), negative in spite of the disease (false negative), positive with no disease (false positive) or positive with the disease present (sensitivity $\mathcal{O}$)

	Disease absent	Disease present
Negative test	$P(L\|\bar{D}) = \mathcal{R}$	$P(L\|D) = 1 - \mathcal{O}$
Positive test	$P(H\|\bar{D}) = 1 - \mathcal{R}$	$P(H\|D) = \mathcal{O}$

The limiting case that box 3 *never* holds the keys is also covered: MH reveals the other empty box so, by switching, C always wins.

1.5.6 Bayes Formula in Medical Diagnostics

We have fallen ill with fever and visit a doctor. Recently he has read some news on the west Nile virus that, on average, infects one person per million. He draws a blood sample for a test that has a positive outcome in $\mathcal{O} = 99\%$ of the cases where the disease is actually present (the so-called *sensitivity* of the test), and a negative outcome in $\mathcal{R} = 95\%$ of the cases where the disease is not present (the so-called *specificity* of the test). The test of our blood comes out positive. ① What is the probability that we are actually infected by the virus? ② Analyze the more general case of a disease probed by a larger number of tests or exhibiting multiple symptoms.

✎ Let us denote the positive outcome of the test by H (“high titer”) and negative by L (“low titer”) and write the corresponding conditional probabilities in Table 1.2. Now just read it carefully. ① The probability that the test is positive and the disease (D) is in fact present, is indeed $P(\text{high titer}|\text{infected}) = P(H|D) = \mathcal{O} = 99\%$. But the probability that we are actually infected by the virus, given the test was positive, is $P(D|H)$, and can be computed by using the Bayes formula (1.16):

$$P(D|H) = \frac{P(H|D)P(D)}{P(H|D)P(D) + P(H|\bar{D})P(\bar{D})} = \frac{\mathcal{O}P(D)}{\mathcal{O}P(D) + (1-\mathcal{R})P(\bar{D})},$$

where we have used $P(\text{low titer}|\text{not infected}) = P(L|\bar{D}) = \mathcal{R}$ and thus, due to the complementarity of H and L, $P(H|\bar{D}) = 1 - P(L|\bar{D}) = 1 - \mathcal{R}$. But the numerator also contains the prior probability that, as a random member of the population, we catch the disease at all, which is $P(D) = 10^{-6}$. This results in a very small probability

$$\begin{aligned} P(D|H) &= \frac{0.99 \times 10^{-6}}{0.99 \times 10^{-6} + 0.05\,(1 - 10^{-6})} \\ &= \frac{9.9 \times 10^{-7}}{9.9 \times 10^{-7} + 0.04999995} \approx 1.98 \times 10^{-5}. \end{aligned}$$

② When a disease manifests itself in several symptoms or tests ($S = S_1 \cap S_2 \cap \cdots \cap S_m$), the posterior probability for the disease is still given by the Bayes formula

$$P(D|S) = \frac{P(S|D)P(D)}{P(S)} = \frac{P(S_1S_2\ldots S_m|D)P(D)}{P(S_1S_2\ldots S_m)},$$

but it becomes useless in practical cases. Namely, for a specific disease D_j from a set of n diseases and a single symptom S one would have the expression

$$P(D_j|S) = \frac{P(S|D_j)P(D_j)}{\sum_{k=1}^{n} P(S|D_k)P(D_k)},$$

which becomes much more complex by adding new symptoms. With each new symptom S_m added to the previous set of symptoms $S_1, S_2, \ldots, S_{m-1}$, one would have to compute

$$P\big(D_j\big|S_1S_2\ldots S_m\big) = \frac{P\big(S_m\big|D_jS_1S_2\ldots S_{m-1}\big)\,P\big(D_j\big|S_1S_2\ldots S_{m-1}\big)}{\sum_{k=1}^{n} P\big(S_m\big|D_kS_1S_2\ldots S_{m-1}\big)\,P\big(D_k|S_1S_2\ldots S_{m-1}\big)}.$$

For a diagnostic system incorporating, say, 50 diseases and 500 symptoms that may occur individually or collectively in any of these diseases, we would require the data on $n \cdot 2^m = 50 \cdot 2^{500} \approx 10^{152}$ conditional probabilities. In the so-called *naive Bayes approach* one therefore frequently assumes that the symptoms are independent in the sense that

$$P(S_i|S_j) = P(S_i), \quad P(S_i|DS_j) = P(S_i|D).$$

The first equation states that the probability for S_i to appear in a part of the population that also exhibits symptom S_j is equal to the probability of S_i appearing in the whole population. The second approximation says that the probability of S_i appearing in a part of the population that has the disease D and some other symptom S_j, is equal to the probability of S_i appearing in *all* persons having the disease D. These simplifications allow us to operate with far fewer conditional probabilities $P(S_i|D_j)$—only $m \cdot n = 50 \cdot 500 = 25{,}000$ in the example above—expressing the probability of S_i given the presence of the disease D_j [10]:

$$P\big(D_j\big|S_1S_2\ldots S_m\big) \approx \frac{\prod_{i=1}^{m} P(S_i|D_j)P(D_j)}{\prod_{i=1}^{m}\sum_{k=1}^{n} P(S_i|D_k)P(D_k)}.$$

Inevitably, the assumption of symptom independence is quite coarse: given the presence of the disease, the probability of two symptoms appearing simultaneously is larger than the product of probabilities of individual symptoms. (If we have a headache and know it was caused by the flu, we will most likely develop a sore throat as well.)

1.5.7 One-Dimensional Random Walk ⋆

(Adapted from [1].) A particle moves along the real axis, starting at the origin ($x = 0$). Consecutive random collisions uniformly spaced in time send it one step to the left (-1) or to the right ($+1$) with probabilities $1/2$ either way. ① What is the probability that after $2n$ collisions the particle will return to $x = 0$ without ever meandering into the $x < 0$ region? Five random walks are shown for illustration in Fig. 1.5 (left). For example, walk number 3 that has always remained at $x \ge 0$ and has terminated at $x = 0$ after 100 collisions is "acceptable". ② Verify your result by a computer simulation. (Random walks will be discussed more generally in Sects. 6.7 and 6.8.)

✎ Each random walk is a consequence of $2n$ collisions. Each collision shifts the particle to the left ($x \mapsto x - 1$) or to the right ($x \mapsto x + 1$), thus the number of all possible walks is 2^{2n}. Let A be the event that the particle returns to the origin after $2n$ collisions, and B the event that the particle does not wander to $x < 0$ during $2n$ collisions. We are looking for the probability $P(AB)$, where $P(AB) = P(B|A)P(A)$.

① Let us first determine $P(A)$. From 2^{2n} possible and equally probable walks only those are acceptable for event A that end up at $(2n, 0)$, like the walk in Fig. 1.6 denoted by the full line. In all of them the particle has experienced n unit kicks to the left and n unit kicks to the right. The number of all such walks can be calculated by counting all possible ways of choosing n collisions that result in a left (or right) shift, from the total $2n$ collisions. There are $\binom{2n}{n}$ such ways, therefore

$$P(A) = \frac{1}{2^{2n}}\binom{2n}{n}.$$

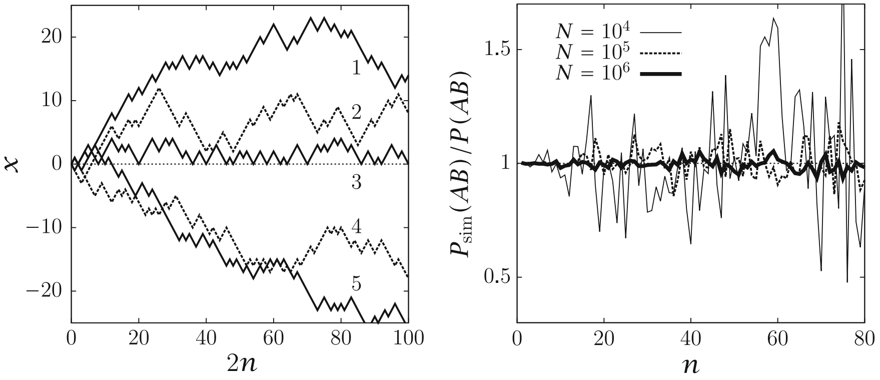


Fig. 1.5 [Left] Five one-dimensional random walks with 100 time steps. We are looking for the fraction of the walks that terminate at the origin (event A) and never blunder to $x < 0$ (condition B), as in walk number 3 shown here. [Right] The ratio between the simulated and theoretical expectation value for event AB

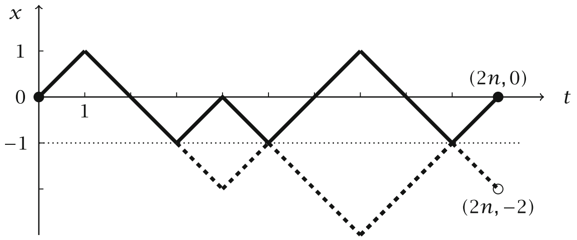


Fig. 1.6 A random walk that enters the region $x < 0$ after a certain time—still returning to the origin after $2n$ steps—and its mirror image from that moment on

From the walks ending up at $(2n, 0)$ and thereby fulfilling condition A, we should disregard those that fluctuate to $x < 0$ if we wish to satisfy condition B. How do we count such occurrences? For each such walk (from the very moment it has crossed the boundary and reached the point $x = -1$) we imagine a new walk, which is the mirror image of the remainder of the previous walk across the $x = -1$ axis (dashed line in Fig. 1.6). The new walk certainly terminates at $(2n, -2)$ and is therefore composed of $n - 1$ right and $n + 1$ left shifts. Hence, under condition A, $\binom{2n}{n+1}$ do not fulfill B, while $\binom{2n}{n} - \binom{2n}{n+1}$ do. This implies that

$$P(B|A) = \frac{\binom{2n}{n} - \binom{2n}{n+1}}{\binom{2n}{n}}.$$

The probability we have been looking for is therefore

$$P(AB) = P(B|A)P(A) = \frac{1}{2^{2n}}\left[\binom{2n}{n} - \binom{2n}{n+1}\right] = \frac{1}{2^{2n}(n+1)}\binom{2n}{n}. \tag{1.18}$$

② You do not trust this calculation? Let us try to check it by a simple computer simulation. For each n chosen in advance, start with a particle at the origin, then randomly add $+1$ or -1 to its current position and write down its final coordinate after $2n$ steps. A walk that ends up at $x = 0$ and has never erred into $x < 0$ is counted as "good". If for each n we repeat N walks, we may expect that the ratio of the good walks and all attempted walks will approach the calculated probability (1.18) in the limit $N \to \infty$. Let us denote this simulated probability by $P_{\text{sim}}(AB)$. Figure 1.5 (right) shows the ratio between $P_{\text{sim}}(AB)$ and the theoretical $P(AB)$ as a function of the walk duration n for three different numbers N of how many times the simulation was re-run. Apparently our calculation was correct: with increasing N the ratio does stabilize near 1. The thick line in the figure still looks wiggly? It is! Recall that for $n = 80$ there are $2^{160} \approx 10^{48}$ all possible walks, while we have performed only a million of them at each n.

References

1. R. Jamnik, *Verjetnostni račun* (Mladinska knjiga, Ljubljana, 1971)
2. P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences* (Cambridge University Press, Cambridge, 2005)
3. D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, 5th edn. (Wiley, New York, 2010)
4. A.G. Frodesen, O. Skjeggestad, H. Tøfte, *Probability and Statistics in Particle Physics* (Universitetsforlaget, Bergen, 1979)
5. T. Bayes, An essay toward solving a problem in the doctrine of chances. Philos. Trans. **53**, 370 (1763)
6. J.J. Brehm, W.J. Mullin, *Introduction to the Structure of Matter* (Wiley, New York, 1989)
7. S. Selvin, A problem in probability. Am. Stat. **29**, 67 (1975)
8. S. Selvin, On the Monty Hall problem. Am. Stat. **29**, 134 (1975)
9. M.A. Carlton, Pedigrees, prizes, and prisoners: the misuse of conditional probability. J. Stat. Educ. **13**(2) (2005)
10. S. Schwartz, J. Baron, J. Clarke, A casual Bayesian model for the diagnosis of appendicitis, *Conference on Uncertainty in Artificial Intelligence (UAI-86)* (Elsevier Science Publishers, Amsterdam, 1986), p. 423

Chapter 2
Probability Distributions

Abstract Starting with the examples of distributions in general, the Dirac delta and the Heaviside unit functions are presented, followed by the definition of continuous and discrete random variables and their corresponding probability distributions. Probability functions, probability densities and (cumulative) distribution functions are introduced. Transformations of random variables are discussed, with particular attention given to the cases where the inverse of the mapping is not unique. Two-dimensional cases are treated separately, defining joint and marginal distributions, as well as explaining the variable transformation rules in multiple dimensions.

Having become acquainted with the basic properties of probability, we shall devote this chapter to the question of how probability can be related to the all-pervading concept of *distribution*. We introduce two general-purpose tools, the so-called Dirac delta "function" and the Heaviside step function, then move on to random variables and their discrete and continuous probability distributions.

2.1 Dirac Delta

The value of a real function f of a real variable x at $x = 0$ can be calculated, of course, by evaluating $f(0)$. But we would like to possess a mathematical tool—denote it by δ—that supplies $f(0)$ as the result of *integrating* f over the whole real axis,

$$\int_{-\infty}^{\infty} f(x)\delta(x)\,\mathrm{d}x = f(0), \qquad f : \mathbb{R} \to \mathbb{R}. \tag{2.1}$$

Physicists call this tool the "Dirac delta". It is a sort of functional, since it maps from a function space to the range of f—for purposes of our discussion, let this be simply $\mathbb{R}$. It seems to operate as a multiplication of f by a very narrow spike (Fig. 2.1 (left)), resulting in the value of f at the origin. This is the reason one often identifies this tool as a genuine δ "function". Due to its property

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_2

$$\int_{-\infty}^{\infty} \delta(x)\,\mathrm{d}x = 1, \tag{2.2}$$

which is nothing but (2.1) in the special case $f(x) = 1$, one often hears even the—completely nonsensical—claim that the δ "function" is normalized. From the strict mathematical point of view, the Dirac delta is neither a function nor a functional, but a measure (see [1] and Appendix A).

Obviously the δ "function" must possess a unit inverse to the unit of x, since $f(x)$ and $f(0)$ must have the same units. If x measures distance (unit [m]), then $\delta(x)$ must have unit [m^{-1}]. In atomic physics, for example, a very narrow and deep square-well potential $V(x)$ with depth $-V_0$ (in [eV]) and width a (in [nm]) around the origin at $x = 0$ (Fig. 2.1 (right)) can be written as

$$V(x) = -aV_0\delta(x).$$

By writing V in this manner we wish to say that in the limits $a \to 0$ and $V_0 \to \infty$, such that the product aV_0 [nm eV] remains constant, the precise shape of $V(x)$ is irrelevant: in computing the expectation values with such a potential only the value of the integrand f at the origin matters. Obviously we have acquired a tool that allows us to represent any point, point-like or very compact quantity in physics, for example, a point electric charge or a tiny mass.

The box-like picture does not appear fancy enough? Two very popular ways to convey the essence of the Dirac delta are the limit (in the sense of functions) of the Gaussian, which we will discuss later, and the Fourier integral representation:

$$\delta(x) \sim \lim_{a\to 0} \frac{1}{\sqrt{\pi}a}\,\mathrm{e}^{-x^2/a^2}, \quad \delta(x) \sim \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}kx}\mathrm{d}k.$$

We proceed analogously in multiple dimensions. The one-dimensional Dirac delta can be generalized to $\mathbb{R}^3$ as

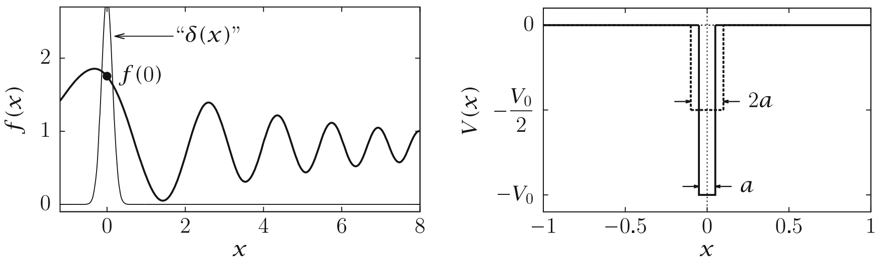


Fig. 2.1 [Left] Illustration of definition (2.1). The integral of the product of a continuous function f and the spike "function" $\delta(x)$ yields the value of f at the origin. [Right] A narrow and deep square potential $V(x)$, for which only the product of its linear dimension and depth is relevant, can be approximated by $V(x) = -aV_0\delta(x)$

$$\int_{\mathbb{R}^3} f(\boldsymbol{r})\delta(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r} = f(\boldsymbol{0}).$$

In this case $\delta(\boldsymbol{r}) = \delta(x)\delta(y)\delta(z)$ must have units of [m^{-3}] if x, y and z are in [m].

2.1.1 Composition of the Dirac Delta with a Function

How does the Dirac delta behave when its argument is a function, as in $\delta\,(g(x))$? If the function $g : \mathbb{R} \to \mathbb{R}$ has precisely one zero at $x_0 \in \mathbb{R}$, i.e. $g(x_0) = 0$, and satisfies the condition $g'(x) \neq 0$, $\forall x \in \mathbb{R}$, then [1]

$$\delta\,(g(x)) = \frac{\delta(x - x_0)}{|g'(x_0)|}. \tag{2.3}$$

The simplest case is $g(x) = x - y$, hence $x_0 = y$ and $|g'(x_0)| = 1$. It follows from (2.3) that by an additive change of the variable, $x \mapsto x - y$, the Dirac delta yields the functional value corresponding to a translation along the abscissa:

$$\int_{-\infty}^{\infty} f(x)\delta(x - y)\,\mathrm{d}x = f(y), \qquad \int_{-\infty}^{\infty} f(x - y)\delta(x)\,\mathrm{d}x = f(-y).$$

We imagine that the Dirac delta "combs" the real axis and thereby "samples" the function f at $x = y$. When g has several simple zeros $\{x_0, x_1, \ldots, x_n\}$, (2.3) must be considered in the vicinity of each zero separately:

$$\int_{-\infty}^{\infty} f(x)\delta\,(g(x))\,\mathrm{d}x = \sum_{i=0}^{n} \frac{f(x_i)}{|g'(x_i)|}. \tag{2.4}$$

In such a case (2.3) is replaced by

$$\delta\big(g(x)\big) = \sum_{i=0}^{n} \frac{\delta(x - x_i)}{|g'(x_i)|}.$$

The zeros x_i *must* be simple (single). The formulas listed above may not be used in the case of multiple zeros, $\int_{-\infty}^{\infty} \delta(x^2)\,\mathrm{d}x =$ "∞". One can use the same tool to deal with the case $g(x) = ax$, $a \in \mathbb{R}$, $a \neq 0$, corresponding to $x_0 = 0$ and $|g'(x_0)| = |a|$. Rescaling the argument x by a non-zero a then yields $\int_{-\infty}^{\infty} \delta(ax)\,\mathrm{d}x = 1/|a|$, which we write symbolically as

$$\delta(ax) = \frac{1}{|a|}\,\delta(x), \qquad a \neq 0, \tag{2.5}$$

or, in three dimensions, as $\delta(a\boldsymbol{x}) = \delta(\boldsymbol{x})/|a|^3$. Let us generalize this to the case that the vector $\boldsymbol{x}$ is rescaled by a matrix A instead of the scalar a! We have

$$\int \delta\big(\underbrace{A\boldsymbol{x}}_{\boldsymbol{y}}\big)\,\mathrm{d}V(\boldsymbol{x}) = \int \delta(\boldsymbol{y})\,\mathrm{d}V\big(A^{-1}\boldsymbol{y}\big) = \big|\det A^{-1}\big| \int \delta(\boldsymbol{y})\,\mathrm{d}V\big(\boldsymbol{y}\big) = \frac{1}{|\det A|},$$

where $\mathrm{d}V$ is the appropriate volume element. Formula (2.5) is then replaced by

$$\delta(A\boldsymbol{x}) = \frac{1}{|\det A|}\,\delta(\boldsymbol{x}), \qquad \det A \neq 0.$$

Example To calculate the effect of the Dirac delta when its argument is the function $g(x) = x^2 - a^2$ possessing two real simple zeros, $x_0 = a$ and $x_1 = -a$, (2.4) must be applied with $g'(x) = 2x$. We obtain

$$\delta(x^2 - a^2) = \frac{\delta(x+a)}{|g'(-a)|} + \frac{\delta(x-a)}{|g'(a)|} = \frac{1}{2|a|}\,[\delta(x+a) + \delta(x-a)]\,.$$

Such a form can be used, for instance, to describe two very narrow and very high potential layers (upward-facing square-well potentials) centered at $x = -a$ and $x = a$. ◁

It is worth mentioning that the Dirac delta is part of the tool used primarily in quantum mechanics to evaluate the integrals of the form

$$\int_{-\infty}^{\infty} \mathrm{d}E \int_0^{\infty} f(E)\,\mathrm{e}^{-\mathrm{i}\,Et}\mathrm{d}t, \tag{2.6}$$

where E and T denote energy and time, respectively. This tool is based on the Sokhotsky–Plemelj theorem dealing with a particular family of Cauchy integrals along closed curves C in the complex plane, providing the limit values of the integral from both sides of C. The version of the theorem on the real axis states that

$$\frac{1}{x \pm \mathrm{i}\,\varepsilon} = \mp\,\mathrm{i}\,\pi\delta(x) + \mathcal{P}\frac{1}{x}.$$

Here $\mathcal{P}$ denotes the principal (generalized) value of the integral, so this compact notation must actually be read as

$$\lim_{\varepsilon \searrow 0} \int_{-\infty}^{\infty} \frac{f(x)}{x \pm \mathrm{i}\,\varepsilon}\,\mathrm{d}x = \mp\,\mathrm{i}\,\pi f(0) + \lim_{\varepsilon \searrow 0} \int_{|x|>\varepsilon} \frac{f(x)}{x}\,\mathrm{d}x, \tag{2.7}$$

where $f : \mathbb{R} \to \mathbb{C}$. We add an infinitesimally small negative real term to the purely imaginary argument of the exponential function in (2.6), integrate over time, and apply (2.7):

$$\lim_{\varepsilon\searrow 0}\int\limits_{-\infty}^{\infty}\mathrm{d}E f(E)\int\limits_{0}^{\infty}\mathrm{e}^{-\mathrm{i}Et-\varepsilon t}\mathrm{d}t = -\mathrm{i}\lim_{\varepsilon\searrow 0}\int\limits_{-\infty}^{\infty}\frac{f(E)}{E-\mathrm{i}\varepsilon}\,\mathrm{d}E = \pi f(0) - \mathrm{i}\lim_{\varepsilon\searrow 0}\int\limits_{|E|>\varepsilon}\frac{f(E)}{E}\,\mathrm{d}E.$$

2.2 Heaviside Function

The Heaviside function H is defined as

$$H(x) = \int_{-\infty}^{x}\delta(t)\,\mathrm{d}t. \tag{2.8}$$

From $x = -\infty$ to just slightly below $x = 0$ the integral yields zero; but as soon as we cross $x = 0$, the value of the integral jumps to 1 according to (2.2) and stays there until $x = \infty$. The function H is therefore also known as the *step function* (Fig. 2.2 (left)). Two handy analytic approximations of H (Fig. 2.2 (right)) are

$$H(x) \sim \lim_{k\to\infty}\left[\frac{1}{2}+\frac{1}{\pi}\arctan kx\right], \tag{2.9}$$

$$H(x) \sim \lim_{k\to\infty}\left[\frac{1}{1+\mathrm{e}^{-2kx}}\right]. \tag{2.10}$$

In literature one occasionally encounters a non-standard definition of the step function, where $H(x<0)=0$, $H(x=0)=1/2$, and $H(x>0)=1$. Its symmetry about the origin does establish a neat resemblance to these analytic approximations, but one should use it carefully.

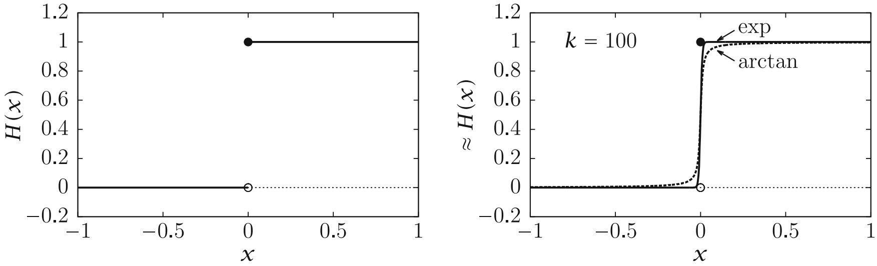


Fig. 2.2 [Left] Heaviside function, known as the *unit step*. We define it to be continuous *from the right*. (Alternative definition with continuity from the left is also possible.) [Right] Analytic approximations (2.9) and (2.10) of the Heaviside function

2.3 Discrete and Continuous Distributions

Before we try to understand probability distributions, consider a distribution of some well-known physical quantity like, for example, mass. What is the spatial distribution of mass in the globular cluster NGC 7006 shown in Fig. 2.3?

From the viewpoint of gravity individual stars can be treated as point bodies, since the stars in the cluster do not overlap and gravity acts as if they were compacted to single points, their respective centers of mass. The spatial dependence of mass density within such a cluster—and in any set of point masses—can be described by the formula

$$\rho(\boldsymbol{r}) = \sum_i m_i \, \delta(\boldsymbol{r} - \boldsymbol{r}_i), \tag{2.11}$$

where m_i is the mass of the individual body and r_i is its position vector. Only at distances smaller than the star radii this description becomes inadequate and forces us to abandon the discrete picture and switch to the continuum. Within an individual star, of course, the distribution of mass is given by the density

$$\rho(\boldsymbol{r}) = \frac{\mathrm{d}m}{\mathrm{d}V},$$

which makes physical sense in the limit $\mathrm{d}V \to 0$. But even this limit must be taken with a grain of salt: descending the order-of-magnitude ladder to ever smaller volumes and into the realm of molecules and atoms, the continuous description again becomes inappropriate and must be replaced by discrete distributions.

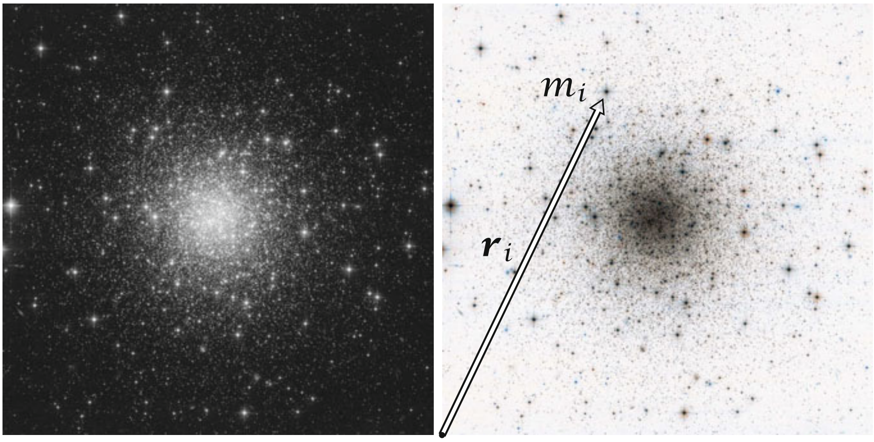


Fig. 2.3 The globular cluster NGC 7006 at a distance of approximately 135,000 light years from the Earth contains hundreds of thousands of stars. [Left] Photograph taken by the Hubble Space Telescope. [Right] The spatial distribution of mass density within the cluster (and in any set of point masses) can be described by (2.11)

2.4 Random Variables

Each outcome of a random experiment is specified by the value of one or more *random* or *stochastic variables*. Random variables are *functions*, defined on the sample space S. Their role is to assign a number to each possible outcome in S; in addition, the frequency with which a certain number occurs, is associated with the corresponding probability. For example, if throwing a die is considered to be a random process with sample space

$$S = \left\{ \text{⚀}, \text{⚁}, \text{⚂}, \text{⚃}, \text{⚄}, \text{⚅} \right\}, \tag{2.12}$$

the value of a random variable X "communicates" the outcome:

$$X\left(\text{⚂}\right) = x_3 = 3\,.$$

We denote random variables by upper-case and their values by lower-case letters. An individual outcome is called the *realization of a random variable* or *draw*. The probability for any outcome in (2.12) is 1/6, hence we can write, as in (1.17):

$$X \sim \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

2.5 One-Dimensional Discrete Distributions

A discrete random variable is a random variable that can assume a finite number of different values $x_i \in \mathbb{R}$ $(i = 1, 2, \ldots, n)$. Let the points x_i on the real axis be arranged such that $x_1 < x_2 < \cdots < x_n$. The probability that in a particular repetition of the experiment X acquires the value x_i, is written as

$$P_i = P(X = x_i) = f_X(x_i), \qquad i = 1, 2, \ldots, n. \tag{2.13}$$

The function f_X is called the *probability [mass] function* that corresponds to a *discrete probability distribution*. Probability is a non-negative quantity, therefore

$$f_X(x) \ge 0.$$

The outcomes of an experiment $(X = x_1, X = x_2, \ldots)$ constitute a complete set of events (1.2), hence the sum of the probabilities of individual outcomes is one:

$$\sum_{i=1}^{n} f_X(x_i) = 1. \tag{2.14}$$

This equation states that the probability distribution is *normalized*.

It makes sense to define the probability that X assumes a value smaller than or equal to some value x. For example, in throwing a die we are interested in the probability that the observed number of spots (random variable X) is less than four ($X < 4$) or that the number of spots is $x_1 = 1$, $x_2 = 2$ or $x_3 = 3$. The sum of probabilities must therefore collect ("accumulate") the values P_1, P_2 and P_3. In other words, the sum in (2.14) should not be pulled all the way to n but only up to $i = 3$. This sum is given by the *[cumulative] distribution function*

$$F_X(x) = P(X \leq x), \qquad -\infty < x < \infty.$$

Since P_i are non-negative, F_X is a non-decreasing function.

$$x \leq y \quad \Longrightarrow \quad F_X(x) \leq F_X(y).$$

The definition domain of F_X formally ranges from $-\infty$ to ∞, so F_X certainly vanishes from the left extreme of the real axis until just below the point x_1, while it is equal to one from the point x_n upwards, since by that point all possible P_i have been collected in the sum:

$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to \infty} F_X(x) = 1.$$

When moving along the x axis, "continuity from the right" applies to F_X:

$$\lim_{\varepsilon \searrow 0} F_X(x + \varepsilon) = F_X(x) \quad \forall x.$$

Hence, for any value x_i encountered while combing x in the positive sense, F_X jumps to a value which is P_i higher than the previous one.

Example A fair die is thrown twice. What is the expected distribution of the sum of spots from both throws after many trials? Let the variable X measure the sum of spots, which can be $x_1 = 2, x_2 = 3, \ldots, x_{11} = 12$. There are $6 \cdot 6 = 36$ possible outcomes, all equally probable (1/36). But different sums are *not* equally probable. The sum of 2 can be obtained in a single way, namely by one spot appearing on the first die and one on the second, hence $P(X = x_1) = f_X(x_1) = 1/36$. The sum of 3 can be realized by $1 + 2$ or $2 + 1$, thus $P(X = x_2) = f_X(x_2) = 2/36$. The sum of 4 appears in three cases: $1 + 3$, $2 + 2$ or $3 + 1$, thus $P(X = x_3) = f_X(x_3) = 3/36$, and so on, up to $P(X = x_{11}) = f_X(x_{11}) = 1/36$. Hence X can be assigned the probability distribution shown in Fig. 2.4 (top). It is non-zero only at eleven points x_i (values denoted by circles), and zero elsewhere.

The distribution function F_X is shown in Fig. 2.4 (bottom). It vanishes from $-\infty$ to $x_1 = 2$, where it jumps to the value $f_X(x_1) = 1/36$. With increasing x, each x_i adds a value of $f_X(x_i)$ to F_X, where it remains until it bumps into the next point, x_{i+1}. When the last point ($x_{11} = 12$) has been accounted for, we have exhausted all possibilities: henceforth, up to $+\infty$, the value $F_X = 1$ stays fixed.

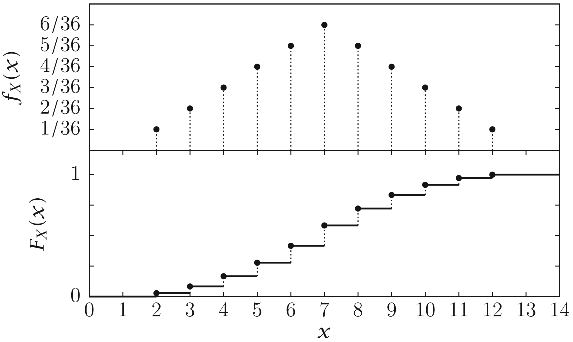


Fig. 2.4 Discrete probability distribution of individual outcomes (sum of spots) in two throws of a die. [Top] The probability distribution of the sum of spots, given by the values of the probability function $f_X(x_i)$. [Bottom] Cumulative distribution function F_X

Let us calculate the probability that the sum of spots is at most $x_5 = 6$:

$$P(X \le 6) = F_X(6) = P_1 + P_2 + P_3 + P_4 + P_5 = \tfrac{1}{36} + \tfrac{2}{36} + \tfrac{3}{36} + \tfrac{4}{36} + \tfrac{5}{36} = \tfrac{5}{12}.$$

What is the probability that it is more than 6? We should not plunge blindly into the calculation. Indeed $P(X > 6) = P_6 + P_7 + P_8 + P_9 + P_{10} + P_{11}$, but one also sees

$$P(X > 6) = F_X(12) - F_X(6) = 1 - P(X \le 6) = \tfrac{7}{12}.$$

The probability that we encounter a sum of spots less than 1 is zero, of course: $P(X \le 1) = F_X(1) = 0$. By common sense or by looking at the figure we also realize that $P(X \le 12, 13, \ldots) = 1$ and $P(X > 12, 13, \ldots) = 0$. ◁

2.6 One-Dimensional Continuous Distributions

In continuous probability distributions we can never speak of "a probability that a continuous variable X assumes a value x". This would be just as inappropriate as claiming that a certain *point* along a thin wire with linear mass density ρ has a finite mass. A point is a mathematical abstraction with dimension zero and can not contain a finite mass. We can only refer to probabilities that "1 m of a thin wire has a mass of 1 mg", "a random variable X has a value between a and b", "value between x and $x + \Delta x$", and so on. Analogously to the mass distribution a continuous probability distribution can be assigned a *probability density function*, which is non-negative and normalized:

$$f_X(x) \ge 0, \qquad \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1.$$

An example of a continuous probability density is shown in Fig. 2.5. The corresponding distribution function F_X can be obtained by integrating the density f_X from $-\infty$ (or the extreme left of its definition domain) up to the current x. From "tiny bits of

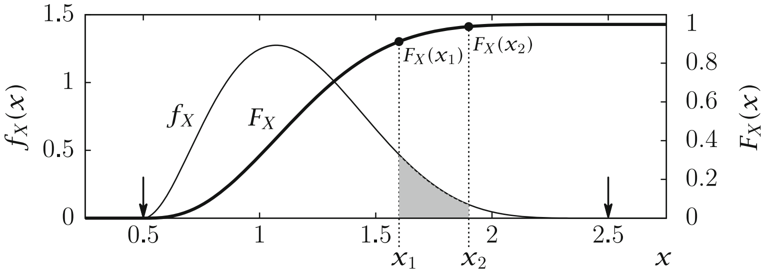


Fig. 2.5 An example of a normalized probability density f_X (*thin curve*, left ordinate) for a continuous probability distribution which differs from zero only on the interval [0.5, 2.5] (*arrows*), and the corresponding distribution function F_X (*thick curve*, right ordinate). The probability $P(x_1 \le X \le x_2)$ is equal to the area of the shaded region (integral of f_X from x_1 to x_2) and also equal to $F_X(x_2) - F_X(x_1)$

probability" $\mathrm{d}F_X = f_X(t)\,\mathrm{d}t$ one obtains the probability that $X \le x$:

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t.$$

If these relations are valid, we say that "variable X is distributed according to the distribution F_X" and denote this as $X \sim F_X$ or $X \sim$ *name of distribution*. (The same convention applies for discrete distributions.) If the variables X and Y are distributed according to the same distribution, we write $X \sim Y$. The obvious fact that

$$f_X(x) = \frac{\mathrm{d}F_X}{\mathrm{d}x} = F_X'(x)$$

will serve us well later. Just as before, the cumulative distribution is a non-decreasing function: from $x = -\infty$ up to the leftmost edge of the interval where $f_X \neq 0$ ($x = 0.5$ in Fig. 2.5), F_X vanishes. With increasing x, an ever larger portion of probability is integrated into the cumulative distribution from this point upwards, until the rightmost edge of the domain of f_X is reached ($x = 2.5$). Here F_X becomes equal to 1 and remains so all the way to $x = +\infty$.

The probability that X assumes a value on the interval $[x_1, x_2]$ is given by the definite integral of the probability density over this range,

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f_X(x)\,\mathrm{d}x.$$

The integral can be written as a difference of integrals

$$\int_{x_1}^{x_2} = \int_{-\infty}^{x_2} - \int_{-\infty}^{x_1},$$

therefore also

$$P(x_1 \le X \le x_2) = F_X(x_2) - F_X(x_1).$$

The shaded region in Fig. 2.5 shows the area under the graph of f_X on the interval $[x_1, x_2] = [1.6, 1.9]$, which equals the probability $P(x_1 \le X \le x_2) \approx 0.0772$. The same result is obtained by subtracting $F_X(x_2) - F_X(x_1) \approx 0.9887 - 0.9115$.

Example Let X be distributed according to the density $f_X(x) = C/(1 + x^2)$, where C is a constant and $-\infty < x < \infty$ (Cauchy distribution). What is the probability that the value of X^2 lies between $\frac{1}{3}$ and 1? We first determine C:

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{C}{1+x^2}\,\mathrm{d}x = C\ \arctan x\Big|_{-\infty}^{\infty} = C\left(\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right) = C\pi = 1,$$

whence $C = 1/\pi$. The condition $1/3 \le X^2 \le 1$ is fulfilled on two intervals, as one can have either $1/\sqrt{3} \le X \le 1$ or $-1 \le X \le -1/\sqrt{3}$. These "events" are mutually exclusive, so the corresponding probabilities should be summed:

$$\begin{aligned} P\left(\tfrac{1}{3} \le X^2 \le 1\right) &= P\left(1/\sqrt{3} \le X \le 1\right) + P\left(-1 \le X \le -1/\sqrt{3}\right) \\ &= \frac{1}{\pi}\int_{1/\sqrt{3}}^{1} \frac{\mathrm{d}x}{1+x^2} + \frac{1}{\pi}\int_{-1}^{-1/\sqrt{3}} \frac{\mathrm{d}x}{1+x^2} = \frac{2}{\pi}\int_{1/\sqrt{3}}^{1} \frac{\mathrm{d}x}{1+x^2} = \frac{1}{6}. \end{aligned}$$

The distribution function is

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = \frac{1}{\pi}\left[\arctan x - \left(-\frac{\pi}{2}\right)\right] = \frac{1}{2} + \frac{1}{\pi}\ \arctan x.$$

Of course $F_X(-\infty) = 0$ and $F_X(\infty) = 1$, as one expects of a distribution function, as well as $P\left(\frac{1}{3} \le X^2 \le 1\right) = F_X(1) - F_X(1/\sqrt{3}) + F_X(-1/\sqrt{3}) - F_X(-1)$. ◁

2.7 Transformation of Random Variables

In this section we learn how to determine the distribution of a random variable calculated from another random variable with a known distribution. Let the random variable X be distributed according to the density f_X. We are interested in the distribution of the variable Y which is some given function of X,

$$Y = h(X),$$

where $h : D \to R \subset \mathbb{R}$ is differentiable and monotonous on D (increasing or decreasing everywhere): this means that $h(x_1) = h(x_2)$ implies $x_1 = x_2$, i.e. its inverse is unique (bijective mapping from one interval to another). Suppose that h increases on D (the contrary case is derived analogously). Then

$$F_Y(y) = P(Y \le y) = P\big(h(X) \le y\big) = P\big(X \le h^{-1}(y)\big)$$
$$= \int_{-\infty}^{h^{-1}(y)} f_X(x)\,\mathrm{d}x = \int_{-\infty}^{y} f_X\left(h^{-1}(z)\right)\left|\frac{\mathrm{d}}{\mathrm{d}z}\,h^{-1}(z)\right|\,\mathrm{d}z,$$

where we have transformed the independent variable, $x = h^{-1}(z)$, and the upper integration boundary, $z = h(x) = h(h^{-1}(y)) = y$. If f_X is continuous at $h^{-1}(y)$, then F_Y is differentiable at y, hence the desired result is

$$\frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = f_Y(y) = f_X\left(h^{-1}(y)\right)\left|\frac{\mathrm{d}}{\mathrm{d}y}\,h^{-1}(y)\right|. \tag{2.15}$$

Example Let us commence with an example from a nuclear physicist's daily lab routine: a planar problem of a point radioactive source and a linear detector (e.g. an electrode in a particle-tracking wire chamber) at a distance d from the source at its nearest point (Fig. 2.6).

The source radiates isotropically (from $\phi = -\pi$ to $\phi = \pi$ in the whole plane or from $\phi = -\pi/2$ to $\phi = \pi/2$ in the lower half-plane), and the detector has a constant sensitivity. The random variable Φ (this used to be X in (2.15)), which measures the emission angle ϕ of the radiated particle, is therefore uniformly distributed,

$$f_\Phi(\phi) = \frac{\mathrm{d}F_\Phi}{\mathrm{d}\phi} = \frac{1}{\pi}, \qquad -\frac{\pi}{2} \le \phi \le \frac{\pi}{2}.$$

But what is the distribution of radiation along the wire? We must convert the distribution over ϕ to a distribution over x. From the figure we infer

$$y = y_0 + d\tan\phi, \tag{2.16}$$

therefore $\phi = h^{-1}(y) = \arctan((y - y_0)/d)$. According to (2.15) we obtain

$$f_Y(y) = f_\Phi\big(h^{-1}(y)\big)\left|\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y)\right| = \frac{1}{\pi}\left|\frac{\mathrm{d}}{\mathrm{d}y}\,\arctan\frac{y-y_0}{d}\right| = \frac{1}{\pi}\,\frac{d}{d^2 + (y-y_0)^2},$$

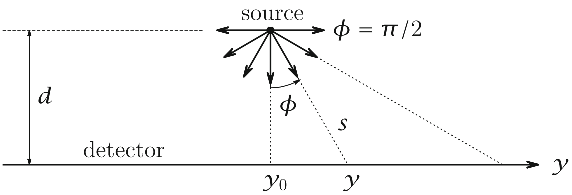


Fig. 2.6 A planar problem with a point radioactive source and infinitely long thin detector. The isotropic radiation from the source (uniform distribution over angles ϕ) is distributed according to the Cauchy distribution along the detector (y coordinate)

which is a correctly normalized Cauchy distribution, since $\int_{-\infty}^{\infty} f_Y(y)\,\mathrm{d}y = 1$. If we read the above equation in reverse, we learn something else: the values of y, randomly distributed according to the Cauchy distribution, can be obtained by randomly picking numbers ϕ, uniformly distributed between $-\pi/2$ and $\pi/2$, and calculating y by using (2.16).

What, then, is the distribution of *flight path lengths* of particles flying from the source to the detector? (The question is relevant because different flight paths imply different energy losses, meaning that the particles will be detected with different energies along the wire.) The flight path length is $s = h(\phi) = d/\cos\phi$: now the functional form of h is different, see Figs. 2.6 and 2.7 (left). Thus $\phi = h^{-1}(s) = \arccos(d/s)$, and the same rule as above yields

$$f_S(s) = f_\Phi\left(h^{-1}(s)\right)\left|\frac{\mathrm{d}}{\mathrm{d}s}h^{-1}(s)\right| = \frac{1}{\pi}\left|\frac{\mathrm{d}}{\mathrm{d}s}\arccos\left(\frac{d}{s}\right)\right| = \frac{1}{\pi}\frac{d}{s\sqrt{s^2-d^2}}. \qquad (2.17)$$

The variable s is defined on $d \le s < \infty$. When we wish to check whether the distribution with the density f_S is normalized, a surprise is in store:

$$\int_d^\infty f_S(s)\,\mathrm{d}s = \frac{1}{2} \ne 1.$$

What went wrong? When the variable ϕ runs through its definition domain, the variable s runs through its respective domain *twice*. For a correct normalization we should therefore multiply the density (2.17) by 2. In other words: the inverse of the function $s(\phi)$ is not unique, since an arbitrary interval of s corresponds to *two* equally long intervals of ϕ, as shown in Fig. 2.7 (left). How this discrepancy is handled will be discussed in the following. ◁

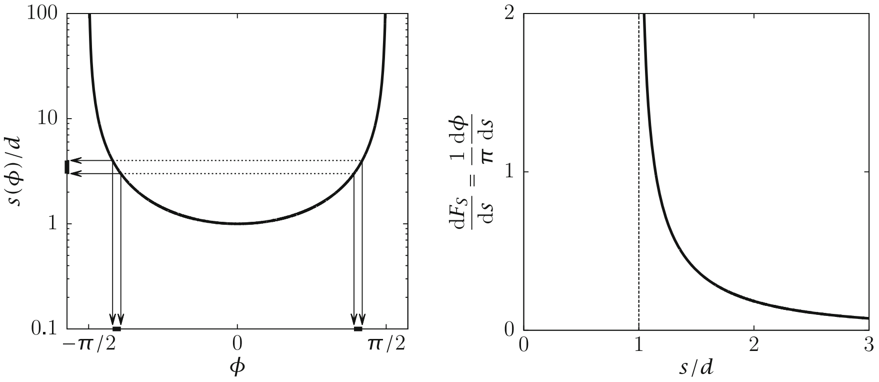


Fig. 2.7 [Left] Path lengths of particles flying from the source to the detector at angle ϕ in the setup of Fig. 2.6. One interval on the ordinate corresponds to two intervals on the abscissa: the inverse of $s(\phi)$ is not unique. [Right] Path-length distribution

2.7.1 What If the Inverse of $y = h(x)$ Is Not Unique?

When the function $h : D \to R$ is not bijective, an interval on the ordinate corresponds to two or more intervals on the abscissa (see Fig. 2.8). Suppose that for each $y \in h(D)$ there is a finite set $\Xi = \{x : h(x) = y\}$. Let $y = h(x)$ for some $x \in D$, and let h be differentiable, except in a countable number of points. By the inverse function theorem, there exists an open interval $I_D \subset D$ including x and an open interval $I_R \subset R$ including y, such that h (restricted to I_D) is bijective and its inverse $g = h^{-1} : I_R \to I_D$ exists and is differentiable. In other words, for each $x_i \in \Xi$ there exists a function g_i such that $(h \circ g_i)(\hat{y}) = \hat{y}$ for each $\hat{y}$ in the neighborhood of y and $(g_i \circ h)(\hat{x}) = \hat{x}$ for each $\hat{x}$ in the neighborhood of x_i. If needed, the interval I_R containing the values y for which all inverses g_i are defined, can be made small enough to render all $\{g_i(I_R)\}$ distinct. Assume $y_1 \le y \le y_2$, where $y_1, y_2 \in I_R$. Thus

$$\begin{aligned} P(y_1 \le Y \le y_2) &= P\Biggl(\bigcup_{x_i \in \Xi} \Bigl\{X \in g_i\bigl([y_1, y_2]\bigr)\Bigr\}\Biggr) = \sum_{x_i \in \Xi} P\Bigl(\Bigl\{X \in g_i\bigl([y_1, y_2]\bigr)\Bigr\}\Bigr) \\ &= \sum_{x_i \in \Xi} \int_{g_i([y_1, y_2])} f_X(x)\,\mathrm{d}x = \sum_{x_i \in \Xi} \int_{y_1}^{y_2} f_X\bigl(g_i(t)\bigr)\,\bigl|g_i'(t)\bigr|\,\mathrm{d}t \\ &= \int_{y_1}^{y_2} \sum_{x_i \in \Xi} f_X\bigl(g_i(t)\bigr)\,\bigl|g_i'(t)\bigr|\,\mathrm{d}t. \end{aligned}$$

Let $y_1 = y$, $y_2 = y + \Delta y$, differentiate both sides of the equation with respect to Δy and finally let $\Delta y \to 0$. It follows that

$$f_Y(y) = \sum_{x_i \in \Xi} f_X\bigl(g_i(y)\bigr)\,\bigl|g_i'(y)\bigr| = \sum_{x_i;\ h(x_i) = y} f_X(x_i)\,\frac{1}{|h'(x_i)|}. \tag{2.18}$$

Example A random variable X is distributed according to the density $f_X(x) = \mathcal{N}\mathrm{e}^{-x}$, where $\mathcal{N}$ is a normalization constant. We would like to calculate the distribution $f_Y(y)$ of the random variable $Y = h(X)$, where

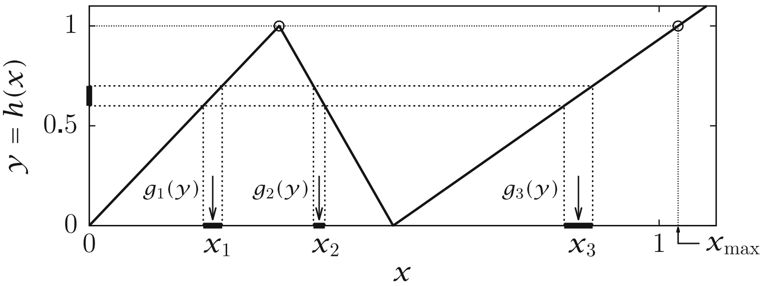


Fig. 2.8 An example of the mapping $y = h(x)$ whose inverse is not unique: an interval on the ordinate corresponds to three intervals on the abscissa; they must be accounted for separately when transforming the probability densities according to (2.18)

$$h(x) = \begin{cases} 3x & ; \ 0 < x \le \frac{1}{3}, \\ 1 - 5\left(x - \frac{1}{3}\right) & ; \ \frac{1}{3} < x \le \frac{8}{15}, \\ 2\left(x - \frac{8}{15}\right) & ; \ x > \frac{8}{15}. \end{cases} \tag{2.19}$$

This function is shown in Fig. 2.8. Let us restrict ourselves to $0 \le y \le 1$ which also dictates the normalization of f_X: the rightmost branch of h reaches the value $y = 1$ at $x = x_{\max}$, where $2(x_{\max} - 8/15) = 1$, so $x_{\max} = 31/30$ and

$$\frac{1}{\mathcal{N}} = \int_0^{x_{\max}} f_X(x)\,\mathrm{d}x \approx 0.644181.$$

Hence the correctly normalized density is

$$f_X(x) = \begin{cases} \mathcal{N}\mathrm{e}^{-x} & ; \ 0 \le x \le \frac{31}{30}, \\ 0 & ; \ \text{elsewhere.} \end{cases}$$

Any subinterval on the y-axis ($0 \le y \le 1$) corresponds to three distinct intervals on the x-axis, lying in the separate definition domains of (2.19). In the leftmost domain we have $y = h(x) = 3x$, so the inverse function there is $x = h^{-1}(x) = g_1(y) = y/3$. Similar results for the remaining two domains are readily obtained:

$$g_1(y) = \frac{y}{3}, \qquad g_2(y) = -\frac{y}{5} + \frac{8}{15}, \qquad g_3(y) = \frac{y}{2} + \frac{8}{15}.$$

We use (2.18) to calculate

$$\begin{aligned} f_Y(y) &= f_X\Big(g_1(y)\Big)\left|\frac{\mathrm{d}g_1(y)}{\mathrm{d}y}\right| + f_X\Big(g_2(y)\Big)\left|\frac{\mathrm{d}g_2(y)}{\mathrm{d}y}\right| + f_X\Big(g_3(y)\Big)\left|\frac{\mathrm{d}g_3(y)}{\mathrm{d}y}\right| \\ &= \mathcal{N}\left[\frac{1}{3}\,\mathrm{e}^{-y/3} + \frac{1}{5}\,\mathrm{e}^{y/5-8/15} + \frac{1}{2}\,\mathrm{e}^{-y/2-8/15}\right]. \end{aligned}$$

We also compute

$$\int_0^1 f_Y(y)\,\mathrm{d}y = 1,$$

hence the distribution with density f_Y is also correctly normalized. ◁

2.8 Two-Dimensional Discrete Distributions

It is not hard to generalize our discussion to two-dimensional probability distributions: first we are dealing with two discrete random variables X and Y, for which we define a *joint probability [mass] function*,

$$P(X = x, Y = y) = f_{X,Y}(x, y),$$

with the properties $f_{X,Y}(x,y) \geq 0$ and $\sum_{x,y} f_{X,Y}(x,y) = 1$. Suppose that X assumes the values $\{x_1, x_2, \ldots, x_m\}$ and Y assumes the values $\{y_1, y_2, \ldots, y_n\}$. By analogy to (2.13) the probability that $X = x_i$ and $Y = y_j$ equals

$$P(X = x_i, Y = y_j) = f_{X,Y}(x_i, y_j).$$

An example of a two-dimensional discrete distribution with the probability function $f_{X,Y}(x_i, y_j) = \mathcal{N}\big(\sin \pi x_i + \sin \pi y_j\big)$, where $x_i \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, $y_j \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, and $\mathcal{N}$ is a normalization factor, is shown in Fig. 2.9 (left). The probability that $X = x_i$ (regardless of Y) is obtained by summing the contributions of all y_j, while the probability for $Y = y_j$ (regardless of X) is calculated by summing the contributions of all x_i:

$$P(X = x_i) = f_X(x_i) = \sum_{j=1}^{n} f_{X,Y}(x_i, y_j),$$

$$P(Y = y_j) = f_Y(x_j) = \sum_{i=1}^{m} f_{X,Y}(x_i, y_j).$$

As usual, the symbols f_X and f_Y denote the projections of the two-dimensional distribution $f_{X,Y}$ to the corresponding distributions pertaining to the variables X and Y alone. Such one-dimensional projections are called *marginal distributions*. One must ensure the overall normalization

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{X,Y}(x_i, y_j) = 1,$$

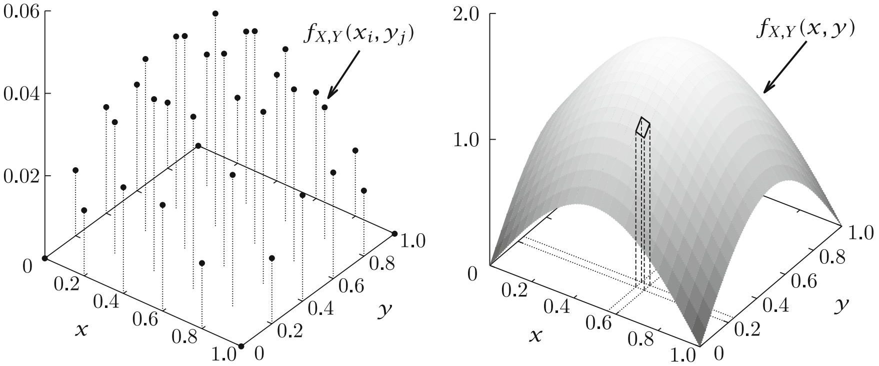


Fig. 2.9 [Left] Discrete two-dimensional probability distribution with a joint probability function $f_{X,Y}(x_i, y_j) = \mathcal{N}\big(\sin \pi x_i + \sin \pi y_j\big)$. [Right] Continuous two-dimensional distribution with a joint density $f_{X,Y}(x, y) = \mathcal{N}\big(\sin \pi x + \sin \pi y\big)$

hence also

$$\sum_{i=1}^{m} f_X(x_i) = \sum_{j=1}^{n} f_Y(y_j) = 1.$$

(Compute the normalization factor $\mathcal{N}$ introduced above as an exercise!) We define the two-dimensional (joint) cumulative distribution function as the sum of all contributions to probability for which $X \le x$ and $Y \le y$, i.e.

$$P(X \le x, Y \le y) = F_{X,Y}(x, y) = \sum_{u \le x} \sum_{v \le y} f_{X,Y}(u, v).$$

If the events $X = x$ and $Y = y$ are independent for all x and y, it holds that

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Such random variables X and Y are called *independent*. In that case we also have

$$F_{X,Y}(x, y) = F_X(x)F_Y(y). \tag{2.20}$$

2.9 Two-Dimensional Continuous Distributions

By now a generalization of continuous probability distributions to two dimensions should not prove a tough nut to crack. One introduces a *joint probability density [function]*, which is non-negative throughout the definition domain,

$$f_{X,Y}(x, y) \ge 0,$$

and normalized,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y = 1.$$

An example of a probability distribution with such probability density on the domain $(x, y) \in [0, 1] \times [0, 1]$ is shown in Fig. 2.9 (right)—calculate the appropriate normalization factor $\mathcal{N}$! The probability that a continuous random variable X takes a value between a and b *and* a continuous random variable Y takes a value between c and d, is equal to the integral

$$P(a \le X \le b,\ c \le Y \le d) = \int_{x=a}^{b} \int_{y=c}^{d} f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y,$$

indicated in the figure by the rectangular cut-out $0.56 \le x \le 0.6$ and $0.20 \le y \le 0.24$. The probability in this example is the volume of the column under the cut-out of the $f_{X,Y}(x, y)$ graph. The corresponding joint distribution function is

$$P(X \le x, Y \le y) = F_{X,Y}(x, y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f_{X,Y}(u, v)\,\mathrm{d}u\,\mathrm{d}v.$$

Analogously to the discrete case we obtain the probabilities for $X \le x$ by integrating the joint density over the whole domain of Y, and vice-versa:

$$P(X \le x) = F_X(x) = \int_{-\infty}^{x} \mathrm{d}u \int_{-\infty}^{\infty} f_{X,Y}(u, v)\,\mathrm{d}v, \tag{2.21}$$

$$P(Y \le y) = F_Y(y) = \int_{-\infty}^{y} \mathrm{d}v \int_{-\infty}^{\infty} f_{X,Y}(u, v)\,\mathrm{d}u. \tag{2.22}$$

Hence, the marginal probability densities are

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, v)\,\mathrm{d}v, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(u, y)\,\mathrm{d}u. \tag{2.23}$$

In the continuous case we call the variables X and Y *independent* if the events $X \le x$ and $Y \le y$ are independent for all x and y, i.e.

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y),$$

which is equivalent to

$$F_{X,Y}(x, y) = F_X(x)F_Y(y). \tag{2.24}$$

It is important that precisely in this case we also have

$$f_{X,Y}(x, y) = f_X(x)f_Y(y). \tag{2.25}$$

Let us complete the story on joint densities by adding the concept of conditional probability. We inquire about the probability that event A occurred (e.g. $Y = y$), with the additional information ("condition") that B also occurred (e.g. $X = x$), glancing at (1.10). With continuous distributions, however, it is meaningless to speak of the probability that a random variable assumed some *precise* value; the statement must therefore be understood in the density sense:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{2.26}$$

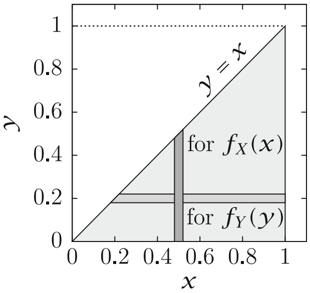
and

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}. \tag{2.27}$$

For example, the probability that $c \le Y \le d$, given $x \le X \le x + \mathrm{d}x$, is

$$P(c \le Y \le d \,|\, x \le X \le x + \mathrm{d}x) = \int_c^d f_{Y|X}(y|x)\,\mathrm{d}y.$$

(This can also be interpreted as the definition of the conditional density $f_{Y|X}$.)



Example Let the continuous random variables X and Y possess the joint probability density

$$f_{X,Y}(x, y) = \begin{cases} 8xy \; ; & 0 \le x \le 1, 0 \le y \le x, \\ 0 \; ; & \text{elsewhere}, \end{cases}$$

shown in the figure. Note that the domain is only the shaded part of [0, 1] × [0, 1]! What are the marginal probability densities $f_X(x)$ and $f_Y(y)$, and the conditional densities $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$? We first check the normalization of $f_{X,Y}$:

$$\int_{x=0}^1 \mathrm{d}x \int_{y=0}^x \mathrm{d}y f_{X,Y}(x, y) = \int_0^1 \mathrm{d}x\, 8x\frac{x^2}{2} = 4\frac{x^4}{4}\bigg|_0^1 = 1.$$

The marginal density $f_X(x)$ is obtained by integrating $f_{X,Y}$ over all possible values of Y, which is from $y = 0$ to $y = x$ (vertical dark-shaded band),

$$f_X(x) = \int_0^x f_{X,Y}(x, y)\,\mathrm{d}y = \int_0^x 8xy\,\mathrm{d}y = \begin{cases} 4x^3 \; ; & 0 \le x \le 1, \\ 0 \; ; & \text{elsewhere}, \end{cases}$$

while the marginal density $f_Y(y)$ is calculated by integrating $f_{X,Y}$ over all values of X, i.e. from $x = y$ to $x = 1$ (horizontal band):

$$f_Y(y) = \int_y^1 f_{X,Y}(x, y)\,\mathrm{d}x = \int_y^1 8xy\,\mathrm{d}y = \begin{cases} 4y(1 - y^2) \; ; & 0 \le y \le 1, \\ 0 \; ; & \text{elsewhere}. \end{cases}$$

The conditional probability densities are

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} 2x/(1-y^2)\ ; & 0 \le y \le x \le 1, \\ 0\ ; & \text{elsewhere,} \end{cases} \tag{2.28}$$

where x is a variable and y a parameter, and

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \begin{cases} 2y/x^2\ ; & 0 \le y \le x \le 1, \\ 0\ ; & \text{elsewhere;} \end{cases} \tag{2.29}$$

here y is a variable and x is a parameter. If our calculation was right, all densities should be correctly normalized, as we have only been tailoring the integration to the desired density. By elementary integration we indeed find out

$$\int_0^1 f_X(x)\,\mathrm{d}x = \int_0^1 f_Y(y)\,\mathrm{d}y = \int_y^1 f_{X|Y}(x|y)\,\mathrm{d}x = \int_0^x f_{Y|X}(y|x)\,\mathrm{d}y = 1.$$

Final question: are X and Y independent? The form of the function $f_{X,Y}(x,y) = 8xy$ might mislead us into believing that a factorization like, for example, $f_{X,Y}(x,y) = 4x \cdot 2y$, already implies that X and Y are independent. But for independence we have required (2.25), which certainly does not apply here, since

$$f_{X,Y}(x,y) = 8xy \ne f_X(x)f_Y(y) = 4x^3 \cdot 4y(1-y^2).$$

The culprit, of course, is the narrowing of the domain $[0, 1] \times [0, 1]$ to the triangle: the $y \le x$ restriction prevents the variables X and Y from grazing freely. ◁

2.10 Transformation of Variables in Two and More Dimensions

In Sect. 2.7 we learned how a probability distribution of a single random variable can be transformed into a distribution of another variable which is a function of the former. We would like to generalize the result of (2.18)—disregarding the issue of uniqueness, already (2.15)—to several dimensions. Instead of the scalar function of a scalar variable, $y = h(x)$, we are now dealing with vector quantities: $\boldsymbol{X} \in \mathbb{R}^n$ is a vector of n independent variables, distributed according to the probability density $f_{\boldsymbol{X}}$, and $\boldsymbol{h}$ is a vector-valued function, which uniquely maps $\boldsymbol{X}$ to a corresponding vector $\boldsymbol{Y} \in \mathbb{R}^n$, so that for the values $\boldsymbol{x}$ and $\boldsymbol{y}$ we have

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \boldsymbol{h}(\boldsymbol{x}) = \begin{pmatrix} h_1(x_1, x_2, \ldots, x_n) \\ h_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ h_n(x_1, x_2, \ldots, x_n) \end{pmatrix}.$$

The n-dimensional generalization of the derivative of h is the Jacobi *total derivative matrix:*

$$\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial h_1}{\partial x_1}(\boldsymbol{x}) & \frac{\partial h_1}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial h_1}{\partial x_n}(\boldsymbol{x}) \\ \frac{\partial h_2}{\partial x_1}(\boldsymbol{x}) & \frac{\partial h_2}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial h_2}{\partial x_n}(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial x_1}(\boldsymbol{x}) & \frac{\partial h_n}{\partial x_2}(\boldsymbol{x}) & \cdots & \frac{\partial h_n}{\partial x_n}(\boldsymbol{x}) \end{pmatrix}. \tag{2.30}$$

Comparing this to (2.15) it is easy to see that the probability density $f_{\boldsymbol{Y}}$ of the variable $\boldsymbol{Y}$ is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}\big(\boldsymbol{h}^{-1}(\boldsymbol{y})\big)\big|J_{\boldsymbol{h}^{-1}}(\boldsymbol{y})\big|, \tag{2.31}$$

where

$$J_{\boldsymbol{h}}(\boldsymbol{x}) = \det\left(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}}(\boldsymbol{x})\right).$$

It is also useful to note that

$$\big|J_{\boldsymbol{h}^{-1}}(\boldsymbol{y})\big| = \big|J_{\boldsymbol{h}}\big(\boldsymbol{h}^{-1}(\boldsymbol{y})\big)\big|^{-1}. \tag{2.32}$$

Example Let X and Y be independent random variables with the joint density

$$f_{X,Y}(x, y) = \frac{1}{2\pi}\mathrm{e}^{-(x^2+y^2)/2}. \tag{2.33}$$

The function $\boldsymbol{h}$ maps a pair of variables (X, Y) into a pair (D, Φ), such that for their values, arranged as vectors $\boldsymbol{x} = (x, y)^{\mathrm{T}}$ and $\boldsymbol{r} = (d, \phi)^{\mathrm{T}}$, it holds that

$$\boldsymbol{r} = \big(d, \phi\big)^{\mathrm{T}} = \boldsymbol{h}(\boldsymbol{x}) = \big(x^2 + y^2, \arctan(y/x)\big)^{\mathrm{T}}, \tag{2.34}$$

where the arctan function is sensitive to the quadrant of the pair (x, y). The inverse of $\boldsymbol{h}$ is

$$\boldsymbol{h}^{-1}(\boldsymbol{r}) = \left(\sqrt{d}\cos\phi, \sqrt{d}\sin\phi\right)^{\mathrm{T}},$$

and the corresponding 2×2 Jacobi matrix (2.30) is

$$\frac{\partial \boldsymbol{h}^{-1}}{\partial \boldsymbol{r}}(\boldsymbol{r}) = \begin{pmatrix} \dfrac{1}{2\sqrt{d}} \cos\phi & -\sqrt{d}\sin\phi \\ \dfrac{1}{2\sqrt{d}} \sin\phi & \sqrt{d}\cos\phi \end{pmatrix}.$$

Its determinant is

$$J_{\boldsymbol{h}^{-1}}(\boldsymbol{r}) = \det\left(\frac{\partial \boldsymbol{h}^{-1}}{\partial \boldsymbol{r}}(\boldsymbol{r})\right) = \frac{1}{2}\cos\phi^2 + \frac{1}{2}\sin\phi^2 = \frac{1}{2}. \tag{2.35}$$

From (2.31) it then follows that

$$f_{D,\Phi}(d,\phi) = f_{X,Y}(\underbrace{x}_{\sqrt{d}\cos\phi}, \underbrace{y}_{\sqrt{d}\sin\phi})\, J_{\boldsymbol{h}^{-1}}(\boldsymbol{r}) = \frac{1}{2\pi}\mathrm{e}^{-(d\cos^2\phi + d\sin^2\phi)/2}\frac{1}{2} = \frac{1}{2}\mathrm{e}^{-d/2}\frac{1}{2\pi}.$$

This means that the variable D is exponentially distributed (as we learn later, "with parameter $1/2$"), while Φ is uniformly distributed, with values on the interval $[0, 2\pi)$. Besides, D are Φ independent. (Explain why!)

The example can also be read in reverse. Start with an exponentially distributed variable D (with parameter $1/2$) and a uniformly distributed, independent variable $\Phi \sim U[0, 2\pi)$, and combine them as

$$\boldsymbol{g}(\boldsymbol{r}) = \left(\sqrt{d}\cos\phi, \sqrt{d}\sin\phi\right)^{\mathrm{T}}.$$

Then $\boldsymbol{g}^{-1}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})$, where the function $\boldsymbol{h}$ is already known from (2.34). The Jacobi matrix corresponding to the inverse function $\boldsymbol{g}^{-1}$ is

$$\frac{\partial \boldsymbol{g}^{-1}}{\partial \boldsymbol{x}}(\boldsymbol{x}) = \begin{pmatrix} 2x & 2y \\ -\dfrac{y}{x^2+y^2} & \dfrac{x}{x^2+y^2} \end{pmatrix}$$

and has determinant 2, as one can see from (2.35) and (2.32) without even computing the matrix, since $J_{\boldsymbol{g}^{-1}}(\boldsymbol{x}) = J_{\boldsymbol{h}}(\boldsymbol{x}) = [J_{\boldsymbol{h}^{-1}}(\boldsymbol{h}(\boldsymbol{x}))]^{-1} = (1/2)^{-1} = 2$. Hence

$$f_{X,Y}(x,y) = f_{D,\Phi}(\underbrace{d}_{x^2+y^2}, \phi) \cdot 2 = \frac{1}{2\pi}\mathrm{e}^{-(x^2+y^2)/2},$$

which is precisely (2.33). We have learned in passing how one can form a pair of independent, normally distributed (Gaussian) variables: pick a value of D from the exponential distribution with parameter $1/2$ and a value of Φ from the interval $[0, 2\pi)$, then calculate $\boldsymbol{g}(D, \Phi)$. See also Sect. C.2.5. ◁

Long example (Retold after [2].) Continuous random variables X and Y are distributed according to the joint probability density

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{96}\,xy\ ; & 0 < x < 4,\ 1 < y < 5, \\ \quad 0\ ; & \text{elsewhere}, \end{cases}$$

which is already normalized, since $\int_0^4 \int_1^5 (xy/96)\,\mathrm{d}x\,\mathrm{d}y = 1$. Find the probability density corresponding to the linear combination $U = X + 2Y$! If we wish to reap the fruits of our previous lessons, we must assign to the pair of variables X and Y another pair $U = h_1(X, Y)$ and $V = h_2(X, Y)$ such that the mapping will be unique. Let

$$u = h_1(x, y) = x + 2y, \qquad v = h_2(x, y) = x.$$

The first choice is motivated by the problem statement itself while the second choice will soon become clear: in short, it allows us to keep the integration bounds as simple as possible and ensure a non-zero Jacobi determinant. We solve both equations for x and y—in other words, we find the inverse functions $X = g_1(U, V)$ and $Y = g_2(U, V)$:

$$x = g_1(u, v) = v, \qquad y = g_2(u, v) = (u - x)/2 = (u - v)/2.$$

From here we can infer that with respect to the original domains of X and Y,

$$0 < x < 4, \qquad 1 < y < 5,$$

the variables U and V span the ranges

$$0 < v < 4, \qquad 2 < u - v < 10,$$

denoted by the shaded area in Fig. 2.10 (left).

Just as in the previous example the new density $f_{U,V}$ is calculated by evaluating the old density $f_{X,Y}$ with transformed arguments and multiplying the result by the absolute value of the Jacobi determinant:

$$f_{U,V}(u, v) = f_{X,Y}\big(\underbrace{g_1(u, v)}_{v},\ \underbrace{g_2(u, v)}_{(u-v)/2}\ \big) \left\| \begin{matrix} \frac{\partial g_1}{\partial u} & \frac{\partial g_1}{\partial v} \\ \frac{\partial g_2}{\partial u} & \frac{\partial g_2}{\partial v} \end{matrix} \right\| = \frac{v(u - v)}{192} \left\| \begin{matrix} 0 & 1 \\ \frac{1}{2} & -\frac{1}{2} \end{matrix} \right\|,$$

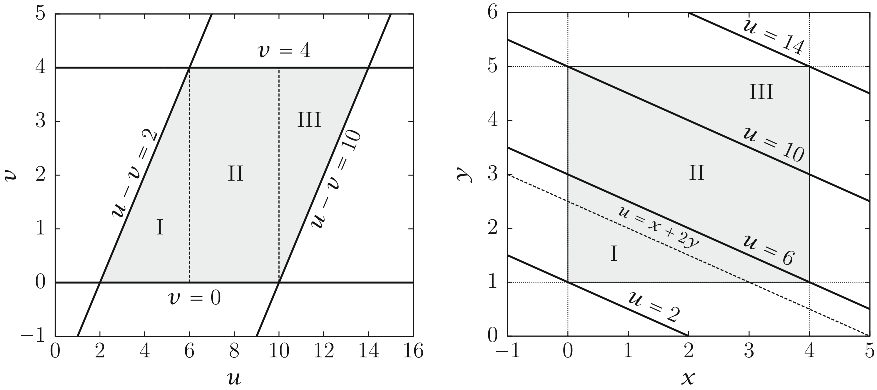


Fig. 2.10 Finding the probability density of the variable $U = X + 2Y$. [Left] Calculation in the domain of the transformed variables U and V. [Right] Calculation in the domain of the original variables X and Y

therefore

$$f_{U,V}(u,v) = \begin{cases} \frac{1}{384}\,v(u-v)\ ; & 0 < v < 4,\, 2 < u-v < 10, \\ 0\ ; & \text{elsewhere.} \end{cases}$$

If this two-dimensional probability density (corresponding to random variables U and V) is integrated over v, we obtain the one-dimensional density corresponding to the variable $U = X + 2Y$. In doing so we must pay attention to correct integration boundaries of the areas denoted by I, II and III in Fig. 2.10:

$$f_U(u) = \begin{cases} \displaystyle\int_0^{u-2} f_{U,V}(u,v)\,\mathrm{d}v = \frac{(u-2)^2(u+4)}{2304} & ;\ 2 < u < 6, \\ \displaystyle\int_0^{4} f_{U,V}(u,v)\,\mathrm{d}v = \frac{3u-8}{144} & ;\ 6 < u < 10, \\ \displaystyle\int_{u-10}^{4} f_{U,V}(u,v)\,\mathrm{d}v = \frac{348u - u^3 - 2128}{2304} & ;\ 10 < u < 14. \end{cases}$$

Alternative Method

There is another path leading to the same goal. We first calculate the distribution function of the variable U, i.e. the probability for "event" $U = X + 2Y \le u$,

$$F_U(u) = P(X + 2Y \le u) = \iint_{x+2y\le u} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y,$$

where $0 < x < 4$ and $1 < y < 5$, and differentiate the resulting function with respect to u. The integration boundaries must again be tailored carefully: we are now integrating over a rectangular domain of the original variables x and y, but the condition $x + 2y \le u$ slices it into three distinct sub-domains shown in Fig. 2.10 (right). In domain I we obtain

$$F_U(u) = \int_0^{u-2} \mathrm{d}x \int_1^{(u-x)/2} \frac{xy}{96}\,\mathrm{d}y = \int_0^{u-2} \frac{x\bigl((u-x)^2-4\bigr)}{768}\,\mathrm{d}x = \frac{(u-2)^3(u+6)}{9216},$$

where we have first integrated over y and then over x. Reversing the order of integration would yield the same result, but the upper integration boundaries must be adjusted accordingly:

$$\int_{x=0}^{u-2}\int_{y=1}^{(u-x)/2} \cdots = \int_{y=1}^{u/2}\int_{x=0}^{u-2y} \cdots .$$

Finally

$$f_U(u) = \frac{\mathrm{d}F_U(u)}{\mathrm{d}u} = \frac{(u-2)^2(u+4)}{2304}, \qquad 2 < u < 6, \tag{2.36}$$

which is exactly the same expression as before. We exploit the same machinery to handle the contributions from regions II and III. In the end, we should also check the normalization: we find $\int_2^{14} f_U(u)\,\mathrm{d}u = 1$, as expected.

Swiss Army Knife Approach

We can perhaps shed a different light on the direct integration of the joint density by resorting to the Dirac delta tool. All cuts through the definition domain in Fig. 2.10 (right) are straight lines of the form $u = x + 2y$. By inserting the Dirac delta with the argument $u - x - 2y = 0$ in the integrand this straight-line constraint is enforced, while the integral over x becomes trivial. We get

$$f_U(u) = \iint f_{X,Y}(x,y)\delta(u-x-2y)\,\mathrm{d}x\,\mathrm{d}y = \int_{I(u)} f_{X,Y}(u-2y,y)\,\mathrm{d}y,$$

we just need to pay attention to the intervals for the integration over y. These depend on the values of u, but in such a way that x never leaves the interval $[0, 4]$ and y never leaves $[1, 5]$. As before, this results in three domains,

$$I(u) = \begin{cases} [1, u/2] & ;\ 2 < u < 6, \\ [(u-4)/2, u/2] & ;\ 6 < u < 10, \\ [(u-4)/2, 5] & ;\ 10 < u < 14, \end{cases}$$

on which the final integral over y should be evaluated. In the first domain, for example, we obtain

$$f_U(u) = \frac{1}{96}\int_1^{u/2}(u-2y)y\,\mathrm{d}y = \frac{1}{96}\left(\frac{uy^2}{2}-\frac{2y^3}{3}\right)\Big|_1^{u/2} = \frac{u^3-12u+16}{2304},$$

which is identical to (2.36). ◁

2.11 Problems

2.11.1 Black-Body Radiation

The distribution of energy spectral density of black-body radiation with respect to wavelengths λ at temperature T is given by the Planck formula

$$u(\lambda; T) = \frac{4\pi}{c}\frac{\mathrm{d}j}{\mathrm{d}\lambda} = \frac{8\pi hc}{\lambda^5}\frac{1}{\exp(hc/\lambda k_\mathrm{B}T)-1},$$

where h is the Planck and k_B is the Boltzmann constant (Fig. 2.11 (left)). Calculate the distribution over the frequencies ν (Fig. 2.11 (right)) and show that the maxima of the two distributions are not at the same location, i.e. $\lambda_\mathrm{max} \neq c/\nu_\mathrm{max}$!

✎ Only one independent variable is involved, so the frequency distribution is obtained by the chain rule for derivatives

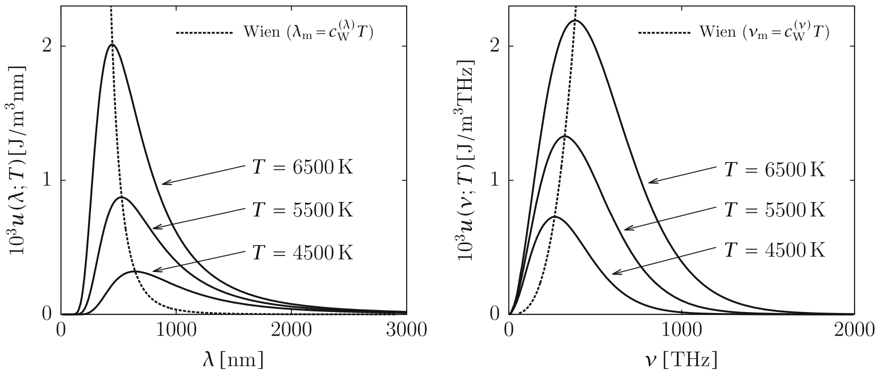


Fig. 2.11 Planck law of black-body radiation. [Left] Temperature dependence of the energy spectral density in terms of wavelengths. [Right] Temperature dependence of the density in terms of frequencies. Wien curves connecting the maxima of both distributions are also shown [3]

$$u(\nu; T) = \frac{4\pi}{c}\frac{\mathrm{d}j}{\mathrm{d}\nu} = \frac{4\pi}{c}\frac{\mathrm{d}j}{\mathrm{d}\lambda}\left|\frac{\mathrm{d}\lambda}{\mathrm{d}\nu}\right| = \frac{8\pi h\nu^3}{c^3}\frac{1}{\exp(h\nu/k_\mathrm{B}T) - 1}.$$

The value on the abscissa which corresponds to the maximum of the distributions can be obtained by solving the equation for the local extremum. In the case of the wavelength spectrum one needs to solve the equation $\mathrm{d}^2 j/\mathrm{d}\lambda^2 = 0$, whence

$$(x_\lambda - 5)\,\mathrm{e}^{x_\lambda} + 5 = 0, \qquad x_\lambda = hc/\lambda k_\mathrm{B}T,$$

while in the case of the frequency distribution we need to solve $\mathrm{d}^2 j/\mathrm{d}\nu^2 = 0$, which becomes

$$(x_\nu - 3)\,\mathrm{e}^{x_\nu} + 3 = 0, \qquad x_\nu = h\nu/k_\mathrm{B}T.$$

These dimensionless equations have analytic solutions (see [4], p. 94), but they can be harnessed numerically by iteration. The solution of the first equation is $x_\lambda \approx 4.965$. At the temperature on the surface of the Sun ($T \approx 6000\,\mathrm{K}$) this means $\lambda_\mathrm{max} \approx 490\,\mathrm{nm}$, i.e. blue light (visible part of the spectrum). The solution of the second equation is $x_\nu \approx 2.821$. At the same temperature T, $\lambda(\nu_\mathrm{max}) = c/\nu_\mathrm{max} \approx 860\,\mathrm{nm}$, which is in the infra-red.

2.11.2 Energy Losses of Particles in a Planar Detector

(Adapted from [5].) Consider a planar detector consisting of two parallel infinite plates spaced apart by h (Fig. 2.12 (left)). A radioactive source attached to the bottom plate radiates α particles with energy E_0. The space between the plates is filled with gas in which particles lose energy. A particle flying unhindered loses all its energy

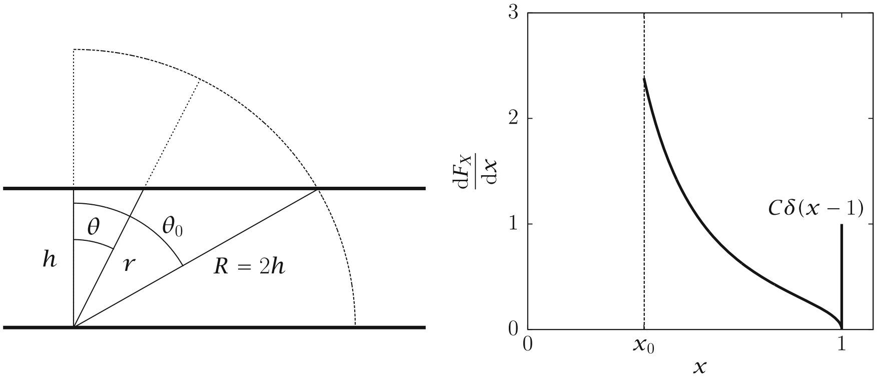


Fig. 2.12 [Left] Planar detector with plates at a distance h. The radioactive source is at the origin. The ratio $C = h/R = 1/2$ is given. [Right] The expected distribution of particle energy losses

along a distance called the *range* (R), which scales roughly as $R = kE_0^{3/2}$, where k is a known constant. The electric signals picked up on the plates are proportional to the energy loss E of particles in the gas. We are interested in the distribution of the pulse heights.

✎ In radioactive decay of nuclei at rest no direction of the sky is privileged; the α particles are emitted isotropically into the solid angle Ω, hence $\mathrm{d}F/\mathrm{d}\Omega = 1/\Omega$. This implies $\mathrm{d}F_\Phi/\mathrm{d}\phi = \text{const.} = 1/2\pi$ and $\mathrm{d}F_\Theta/\mathrm{d}(\cos\theta) = \text{const.} = 1$ (upper hemisphere). The remaining energy of a particle emitted under the angle θ and hitting the top plate after flying over a distance of $r = h/\cos\theta$ is $E_0 - E$. If the plate were not there, the particle could have flown an additional distance $R - r = k(E_0 - E)^{3/2}$. Let us introduce dimensionless quantities $x = E/E_0$ and $\rho = r/R$, so that the equation for the range becomes $1 - \rho = (1 - x)^{3/2}$. We read off $\rho = C/\cos\theta$ from the figure, therefore $\cos\theta = C/[1 - (1 - x)^{3/2}]$. Our task is to express the original distribution over $\cos\theta$ by the distribution over x, which we accomplish by the derivative chain rule:

$$f_X(x) = \frac{\mathrm{d}F_X}{\mathrm{d}x} = \frac{\mathrm{d}F_\Theta}{\mathrm{d}(\cos\theta)}\left|\frac{\mathrm{d}(\cos\theta)}{\mathrm{d}x}\right| = \frac{3C}{2}\frac{\sqrt{1-x}}{\left[1 - (1-x)^{3/2}\right]^2}. \tag{2.37}$$

In Fig. 2.12 (right) this probability density is shown by the curve on the interval from $x = x_0$ to $x = 1$. The lower edge of the interval corresponds to the smallest possible energy loss of the particle in the gas: it occurs if the particle flies vertically upwards from the source, so that $r = h$ and therefore $x_0 = 1 - (1 - C)^{2/3} \approx 0.370$.

What about the particles that are emitted under large enough angles to lose all their energy (meaning $x = 1$) and never reach the top plate? From the geometry we deduce that the fraction of such particles is $\cos\theta_0 = h/R = C$, and they contribute to the energy loss distribution with an additional term

$$C\delta(x - 1). \tag{2.38}$$

Only then the sum of (2.37) and (2.38) is correctly normalized, so that $\int_0^1 f_X(x)\,\mathrm{d}x = 1$.

2.11.3 Computing Marginal Probability Densities from a Joint Density

Continuous random variables X and Y are distributed according to the joint probability density

$$f_{X,Y}(x,y) = \begin{cases} C(2x + y)\ ; & 2 < x < 6, 0 < y < 5, \\ 0\ ; & \text{elsewhere}, \end{cases}$$

where C is the normalization constant. ① Determine C and calculate the probabilities $P(X > 3, Y > 2)$, $P(X > 3)$ and $P(X + Y < 4)$. ② Compute the distribution functions $F_X(x)$ and $F_Y(y)$, then differentiate them with respect to x and y to obtain the marginal probability densities $f_X(x)$ and $f_Y(y)$.

✎ ① The normalization constant is determined by integrating the density $f_{X,Y}$ over the whole definition domain:

$$C \int_2^6 \mathrm{d}x \int_0^5 (2x + y)\,\mathrm{d}y = C \int_2^6 \left(2xy + \frac{y^2}{2}\right)\Bigg|_{y=0}^{y=5} \mathrm{d}x$$
$$= C \int_2^6 \left(10x + \frac{25}{2}\right)\mathrm{d}x = 210C = 1,$$

hence $C = 1/210$. The required probabilities are

$$P(X > 3, Y > 2) = \int_3^6 \mathrm{d}x \int_2^5 f_{X,Y}(x,y)\,\mathrm{d}y = \frac{15}{28},$$
$$P(X > 3) = \int_3^6 \mathrm{d}x \int_0^5 f_{X,Y}(x,y)\,\mathrm{d}y = \frac{23}{28},$$
$$P(X + Y < 4) = \int_2^4 \mathrm{d}x \int_0^{4-x} f_{X,Y}(x,y)\,\mathrm{d}y = \frac{2}{35}.$$

② The cumulative distributions can be computed by resorting to (2.21) and (2.22):

$$F_X(x) = P(X \le x) = \begin{cases} 0 & ;\ x < 2, \\ \displaystyle\int_2^x \mathrm{d}u \int_0^5 f_{X,Y}(u,v)\,\mathrm{d}v = \frac{2x^2 + 5x - 18}{84} & ;\ 2 \le x < 6, \\ 1 & ;\ x \ge 6, \end{cases}$$
$$F_Y(y) = P(Y \le y) = \begin{cases} 0 & ;\ y < 0, \\ \displaystyle\int_0^y \mathrm{d}v \int_2^6 f_{X,Y}(u,v)\,\mathrm{d}u = \frac{y^2 + 16y}{105} & ;\ 0 \le y < 5, \\ 1 & ;\ y \ge 5, \end{cases}$$

and thence

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} = \frac{4x + 5}{84}, \qquad 2 \le x < 6,$$
$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{2y + 16}{105}, \qquad 0 \le y < 5,$$

while one has $f_X(x) = f_Y(y) = 0$ outside of the specified regions. Of course we can reach the same conclusion more easily, through (2.23).

2.11.4 Independence of Random Variables in Two Dimensions

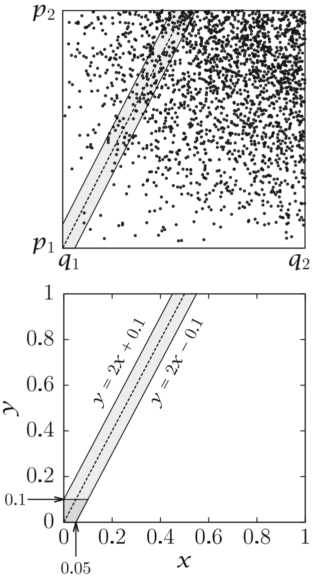
Suppose that the coordinate q and the velocity p of some body are random, that they always lie on the intervals $[q_1, q_2]$ and $[p_1, p_2]$, and that by measuring them a distribution of points in phase space (q, p) is observed (top figure). What is the fraction of points in the phase space corresponding to a 10% deviation from the linear relation

$$p = p_1 + 2\frac{\Delta p}{\Delta q}(q - q_1),$$

where $\Delta q = q_2 - q_1$ and $\Delta p = p_2 - p_1$? The relation is denoted by the dashed line, while the condition

$$\left| \frac{p - p_1}{\Delta p} - 2\,\frac{q - q_1}{\Delta q} \right| \le 0.1$$

is indicated by the shaded area. In dimensionless variables $X = (Q - q_1)/\Delta q$ and $Y = (P - p_1)/\Delta p$ the condition becomes $|Y - 2X| \le 0.1$ (see bottom panel).



Of course the events in the indicated region can be simply counted, but let us try to envision a simple model. The gradual increase of the density of points from q_1 towards q_2 and from p_1 to p_2 suggests that perhaps the mechanism behind the observed pattern can be described by a distribution of continuous variables X and Y with a joint density

$$f_{X,Y}(x,y) = \begin{cases} Cxy \,;\ 0 \le x \le 1,\ 0 \le y \le 1, \\ \quad 0 \,;\ \text{elsewhere}, \end{cases}$$

where C is a constant. Normalize the distribution, then calculate ① $P(|Y - 2X| \le 0.1)$, i.e. the probability that the values of X and Y are restricted to the shaded band. ② Calculate the one-dimensional probability densities $f_X(x)$ and $f_Y(y)$. Are the variables X and Y independent? (Do not forget that this is only a model!)

✎ First we normalize the joint density:

$$1 = \int_0^1 \mathrm{d}x \int_0^1 \mathrm{d}y f_{X,Y}(x,y) = C \int_0^1 x \,\mathrm{d}x \int_0^1 y \,\mathrm{d}y = C \frac{x^2}{2}\bigg|_0^1 \frac{y^2}{2}\bigg|_0^1 = \frac{C}{4} \quad \Longrightarrow \quad C = 4.$$

① The required probability is obtained by integrating the density over two regions: the dark-shaded region defined by $0 \le x \le (y + 0.1)/2$ and $0 \le y \le 0.1$, and the light-shaded region defined by $(y - 0.1)/2 \le x \le (y + 0.1)/2$ and $0.1 \le y \le 1$:

$$\begin{aligned} P(|Y - 2X| \le 0.1) &= \int_0^{0.1} \mathrm{d}y \int_0^{(y+0.1)/2} f_{X,Y}(x,y)\,\mathrm{d}x + \int_{0.1}^1 \mathrm{d}y \int_{(y-0.1)/2}^{(y+0.1)/2} f_{X,Y}(x,y)\,\mathrm{d}x \\ &= \frac{1}{2}\left[\int_0^{0.1} y(y+0.1)^2\,\mathrm{d}y + \int_{0.1}^1 0.4y^2\,\mathrm{d}y\right] \approx 0.0667. \end{aligned}$$

② The variables X and Y are independent:

$$\begin{aligned} f_X(x) &= \int_0^1 f_{X,Y}(x,y)\,\mathrm{d}y = 4x \int_0^1 y\,\mathrm{d}y = 2x, \qquad 0 \le x \le 1, \\ f_Y(y) &= \int_0^1 f_{X,Y}(x,y)\,\mathrm{d}x = 4y \int_0^1 x\,\mathrm{d}x = 2y, \qquad 0 \le y \le 1, \end{aligned}$$

therefore we indeed observe $f_{X,Y}(x,y) = 4xy = f_X(x) f_Y(y) = 2x \cdot 2y$.

2.11.5 Transformation of Variables in Two Dimensions

Let two independent continuous variables X and Y be described by the joint probability density

$$f_{X,Y}(x,y) = \frac{1}{2\pi}\,\mathrm{e}^{-(x^2+y^2)/2}, \qquad -\infty < x, y < \infty.$$

Calculate the joint probability density of random variables U and V, where ① $U = \sqrt{X^2 + Y^2}$, $V = \arctan(X/Y)$ and ② $U = \sqrt{X^2 + Y^2}$, $V = X/Y$!

✎ In the case ① the system of equations relating x and y to u and v has a unique solution

$$x = g_1(u, v) = u \sin v, \qquad y = g_2(u, v) = u \cos v.$$

The Jacobi determinant of this system is

$$J = \begin{vmatrix} \dfrac{\partial g_1}{\partial u} & \dfrac{\partial g_1}{\partial v} \\ \dfrac{\partial g_2}{\partial u} & \dfrac{\partial g_2}{\partial v} \end{vmatrix} = \begin{vmatrix} \sin v & u \cos v \\ \cos v & -u \sin v \end{vmatrix} = -u \sin^2 v - u \cos^2 v = -u,$$

and its absolute value is $|J| = u$. From (2.31) we obtain

$$f_{U,V}(u, v) = f_{X,Y}(u \sin v, u \cos v) \cdot u = \frac{1}{2\pi} \mathrm{e}^{-(u^2 \sin^2 v + u^2 \cos^2 v)/2} \cdot u = \frac{u}{2\pi} \mathrm{e}^{-u^2/2}.$$

Let us check the normalization of the density $f_{U,V}$! Suitable definition domains of the transformed variables must be considered, $0 \le u < \infty$ and $0 \le v < 2\pi$. Then indeed

$$\int_0^\infty u\, \mathrm{e}^{-u^2/2} \mathrm{d}u \int_0^{2\pi} \frac{\mathrm{d}v}{2\pi} = 1.$$

In the case ② the system has two solutions:

$$x = g_1(u, v) = s \frac{uv}{\sqrt{1 + v^2}}, \qquad y = g_2(u, v) = s \frac{u}{\sqrt{1 + v^2}}, \tag{2.39}$$

where $s = \pm 1$. The Jacobi determinant for the first solution is

$$J = \begin{vmatrix} \dfrac{\partial g_1}{\partial u} & \dfrac{\partial g_1}{\partial v} \\ \dfrac{\partial g_2}{\partial u} & \dfrac{\partial g_2}{\partial v} \end{vmatrix} = \begin{vmatrix} \dfrac{v}{\sqrt{1 + v^2}} & u \left(\dfrac{1}{\sqrt{1 + v^2}} - \dfrac{v^2}{(1 + v^2)^{3/2}} \right) \\ \dfrac{1}{\sqrt{1 + v^2}} & -\dfrac{uv}{(1 + v^2)^{3/2}} \end{vmatrix} = -\frac{u}{1 + v^2},$$

and its absolute value is $|J| = u/(1 + v^2)$. Equation (2.31) then yields

$$f_{U,V}(u, v) = f_{X,Y} \left(\frac{uv}{\sqrt{1 + v^2}}, \frac{u}{\sqrt{1 + v^2}} \right) \cdot \frac{u}{1 + v^2} = \frac{1}{2\pi} \frac{u}{1 + v^2} \mathrm{e}^{-u^2/2}.$$

Let us again check the normalization:

$$\frac{1}{2\pi} \int_0^\infty u\, \mathrm{e}^{-u^2/2} \mathrm{d}u \int_{-\infty}^\infty \frac{\mathrm{d}v}{1 + v^2} = \frac{1}{2}.$$

What have we missed? In the case ② the mapping from (u, v) to (x, y) is not unique, so the problem must be split into such domain segments that on each of them the inverses $x = g_1(u, v)$ and $y = g_2(u, v)$ are unique—precisely in the same spirit as in the one-dimensional problem of Sect. 2.7.1. We must evaluate (2.31) on each of these segments and sum the contributions. Since $|J|$ is the same for both solutions of (2.39), all that is missing for the correct probability density is a factor of 2, thus

$$f_{U,V}(u, v) = \frac{1}{\pi} \frac{u}{1 + v^2} \, \mathrm{e}^{-u^2/2}.$$

2.11.6 Distribution of Maximal and Minimal Values

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed continuous random variables. ① What is the distribution of their maximal value

$$U = \max\{X_1, X_2, \ldots, X_n\}?$$

(One can inquire about the distribution of U because U itself is a random variable.) Derive the general expression and apply it in the case that all X_i are described by the exponential probability density of the form $f_X(x) = \lambda \, \mathrm{e}^{-\lambda x}$, where $x \geq 0$, $\lambda > 0$. ② What is the distribution of the *minimum* of such variables,

$$V = \min\{X_1, X_2, \ldots, X_n\}?$$

✎ Problem ① can be solved by using distribution functions. If the maximal value of all X_i should be smaller than some x, *all* X_i simultaneously should be smaller than x, hence

$$F_U(x) = P(U \leq x) = P(X_1 < x, \; X_2 < x, \ldots, \; X_n < x) = \bigl[P(X_1 \leq x)\bigr]^n = \bigl[F_X(x)\bigr]^n.$$

For an individual exponentially distributed X it holds that

$$P(X \leq x) = \int_0^x f_X(t) \, \mathrm{d}t = F_X(x) = \int_0^x \lambda \, \mathrm{e}^{-\lambda t} \mathrm{d}t = 1 - \mathrm{e}^{-\lambda x},$$

therefore

$$F_U(x) = \bigl[1 - \mathrm{e}^{-\lambda x}\bigr]^n.$$

To obtain the probability density corresponding to U, we only need to compute the derivative of F_U,

$$f_U(x) = F_U'(x) = n\lambda \, \mathrm{e}^{-\lambda x} \bigl[1 - \mathrm{e}^{-\lambda x}\bigr]^{n-1}.$$

Problem ② can be solved analogously. If all X_i are simultaneously larger than some x, the minimal X_i is certainly also larger than x:

$$\begin{aligned} P(V > x) &= 1 - P(V \le x) = 1 - F_V(x) \\ &= P(X_1 > x,\, X_2 > x, \ldots, X_n > x) = \bigl[P(X_1 > x)\bigr]^n. \end{aligned}$$

For exponentially distributed X we have

$$P(X > x) = \int_x^\infty \lambda\, \mathrm{e}^{-\lambda t} \mathrm{d}t = \mathrm{e}^{-\lambda x},$$

whence

$$F_V(x) = 1 - \bigl[\mathrm{e}^{-\lambda x}\bigr]^n = 1 - \mathrm{e}^{-n\lambda x}.$$

The probability density corresponding to variable V is then

$$f_V(x) = F_V'(x) = n\lambda\, \mathrm{e}^{-n\lambda x}.$$

In other words, if each of the n independent variables $\{X_i\}_{i=1}^n$ is exponentially distributed (with parameter λ), their minimal value is also exponentially distributed, but with parameter $n\lambda$.

References

1. J.C. Ferreira, Introduction to the theory of distributions, *Pitman Monographs and Surveys in Pure and Applied Mathematics* (Addison Wesley Longman Ltd., Harlow, 1997)
2. M.R. Spiegel, J. Schiller, R.A. Srinivasan, *Theory and Problems of Probability and Statistics*, 4th edn. (McGraw-Hill, New York, 2012)
3. J.J. Brehm, W.J. Mullin, *Introduction to the Structure of Matter* (Wiley, New York, 1989)
4. S. Širca, M. Horvat, *Computational Methods for Physicists* (Springer, Berlin, 2012)
5. I. Kuščer, A. Kodre, *Mathematik in Physik und Technik* (Springer, Berlin, 1993)

Chapter 3
Special Continuous Probability Distributions

Abstract Particular continuous distributions encountered on a daily basis are discussed: the simplest uniform distribution, the exponential distribution characterizing the decay of unstable atoms and nuclei, the ubiquitous normal (Gauss) distribution in both its general and standardized form, the Maxwell velocity distribution in its vector and scalar form, the Pareto (power-law) distribution, and the Cauchy (Lorentz, Breit–Wigner) distribution suitable for describing spectral line shapes and resonances. Three further distributions are introduced (χ^2-, Student's t- and F-distributions), predominantly used in problems of statistical inference based on samples. Generalizations of the exponential law to hypo- and hyper-exponential distributions are presented.

In this chapter we become acquainted with the most frequently used continuous probability distributions that physicists typically deal with on a daily basis.

3.1 Uniform Distribution

Its name says it all: the *uniform distribution* describes outcomes of random experiments—a set of measured values of a random variable—where all values between the lowest (a) and the highest possible (b) are equally probable. A bus that runs on a 15-min schedule, will turn up at our stop anywhere between $a = 0$ min and $b = 15$ min from now: our waiting time X is a continuous random variable distributed uniformly between a and b, which one denotes as

$$X \sim U(a, b).$$

The probability density corresponding to the uniform distribution $U(a, b)$ is

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} \; ; & a \le x \le b, \\[2ex] 0 \; ; & \text{elsewhere}, \end{cases} \tag{3.1}$$

(Fig. 3.1 (left)) and its distribution function is

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_3

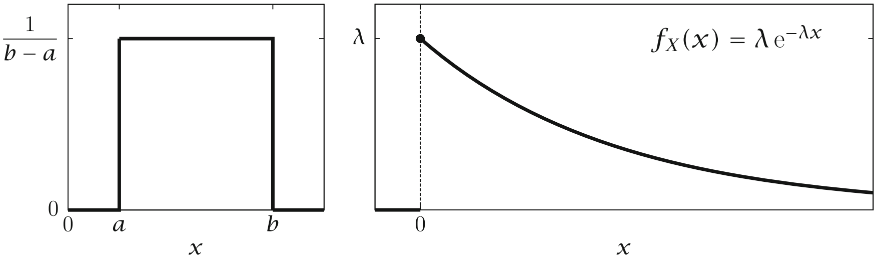


Fig. 3.1 [Left] The probability density of the uniform distribution $U(a, b)$. [Right] The probability density of the exponential distribution with parameter λ

$$P(X \le x) = F_X(x) = \begin{cases} 0 \;;\; x < a, \\ \dfrac{x-a}{b-a} \;;\; a \le x \le b, \\ 1 \;;\; x > b. \end{cases}$$

If we show up at the bus stop at a random instant, the probability that our waiting time will not exceed 10 min, is

$$P(X \le 10) = F_X(10) = \frac{10-0}{15-0} = \frac{2}{3}.$$

Example On a hot day, a house-fly mostly sits still, but occasionally takes off to stretch its legs. Suppose that the time T of its buzzing around is uniformly distributed between 0 and 30 s, i.e. $T \sim U(0, 30)$. What is the probability that it will fly for more than 20 s (event A) *given that* it flies for more than 10 s (condition B)? Due to the additional information B the probability density is no longer $f_T(t) = 1/30\,\mathrm{s}$ but $\widetilde{f}_T(t) = 1/((30\text{–}10)\mathrm{s}) = 1/20\,\mathrm{s}$, hence

$$P\big(T > 20\,\mathrm{s} \,|\, T > 10\,\mathrm{s}\big) = \int_{20\,\mathrm{s}}^{30\,\mathrm{s}} \widetilde{f}_T(t)\,\mathrm{d}t = \frac{30\,\mathrm{s} - 20\,\mathrm{s}}{20\,\mathrm{s}} = \frac{1}{2}.$$

The same result can be obtained by using the original density $f_T(t)$ and direct application of the conditional probability formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \int_{20\,\mathrm{s}}^{30\,\mathrm{s}} f_T(t)\,\mathrm{d}t \bigg/ \int_{10\,\mathrm{s}}^{30\,\mathrm{s}} f_T(t)\,\mathrm{d}t = \frac{1/3}{2/3} = \frac{1}{2}.$$

No matter how trivial the example is, do not forget that computing a conditional probability imposes a restriction on the sample space! ◁

3.2 Exponential Distribution

The exponential distribution is used to describe processes in which the probability of a certain event per unit time is constant: the classical example is the time-dependence of the radioactive decay of nuclei, but it is also used in modeling the distribution of waiting times in queues or durations of fault-free operation (lifetimes) of devices like light bulbs or computer disks.

The decay of an unstable atomic nucleus is a random process *par excellence* (see also Sects. C.3 and 3.2.1). For a single nucleus, it is impossible to predict the precise moment of its decay; the probability for it to decay in some time interval depends only on the length of this interval, Δt, not on the age of the nucleus. We say that the nuclei "do not age" and that radioactive decay is a "memory-less" process: suppose that we have been waiting in vain for time t for the nucleus to decay; the probability that the decay finally occurs after $t + \Delta t$, is independent of t,

$$P(T > t + \Delta t \,|\, T > t) = P(T > \Delta t). \tag{3.2}$$

If the interval Δt is short enough, we can assume that the decay probability is proportional to Δt, and then the only choice becomes

$$P(\text{decay}) = \lambda \Delta t \qquad \text{or} \qquad P(\text{no decay}) = 1 - \lambda \Delta t,$$

where $\lambda = 1/\tau$ is the *decay probability per unit time* [s^{-1}], also called the *decay constant*, while τ is the *characteristic* or *decay time*. The probability that a nucleus has *not decayed yet* after $n\Delta t$ is $(1 - \lambda\Delta t)^n$. The probability that it has not decayed after a longer time $t = n\Delta t$, meaning that it will decay at some time $T > t = n\Delta t$, is therefore

$$P(T > t) = \lim_{n\to\infty} (1 - \lambda\Delta t)^n = \lim_{n\to\infty} \left(1 - \frac{\lambda t}{n}\right)^n = \mathrm{e}^{-\lambda t}. \tag{3.3}$$

Since $P(T > t) = 1 - P(T \le t) = 1 - F_T(t)$, we can immediately calculate the corresponding probability density,

$$f_T(t) = \frac{\mathrm{d}F_T(t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(1 - \mathrm{e}^{-\lambda t}\right) = \lambda\,\mathrm{e}^{-\lambda t}, \qquad t \ge 0, \tag{3.4}$$

shown in Fig. 3.1 (right). (As an exercise, check the validity of (3.2)!) Let us think in a complementary way: the probability that the nucleus has not decayed until time t must equal the probability that it will decay at some instant from t until ∞, i.e. the corresponding integral of the density we have just derived. Indeed

$$\int_t^\infty f_T(t')\,\mathrm{d}t' = \int_t^\infty \lambda\,\mathrm{e}^{-\lambda t'}\,\mathrm{d}t' = \mathrm{e}^{-\lambda t}.$$

It is incredible how many wrong interpretations of these considerations can be heard, so let us reiterate: Equation (3.3) gives the probability that until time t the nucleus has not decayed. At time zero this probability equals 1 and exponentially drops to zero henceforth: every unstable nucleus will decay at some time. The rate of change of the number of nuclei—nuclei *still available for decay*—is given by the differential equation $\mathrm{d}N(t)/\mathrm{d}t = -\lambda N(t)$ with the initial condition $N(t=0) = N_0$, and its solution is

$$N(t) = N_0\,\mathrm{e}^{-\lambda t}. \tag{3.5}$$

The decay constant λ is determined experimentally by counting the number of decays $R(t)$ until time t. Since $N_0 = N(t) + R(t)$, it follows from above that $\mathrm{e}^{-\lambda t} = 1 - R(t)/N_0$, therefore

$$\lambda t = -\log\left(1 - \frac{R(t)}{N_0}\right).$$

By fitting this functional dependence to the measured data we extract $\lambda = 1/\tau$.

Mini-example Two counters in a bank are busy serving a single customer each: the first person has just arrived, while the other has been there for 10 min. Which counter should we choose in order to be served as quickly as possible? If the waiting times are exponentially distributed, it does not matter. ◁

Example You do not believe the Mini-example? Let the variable T measure the time between consecutive particle hits in a Geiger–Müller counter, where T is exponentially distributed, with a characteristic time of $\tau = 84$ s [1]. The probability that we detect a particle $\Delta t = 30$ s after the counter has been switched on, is

$$P\big(T \le \Delta t\big) = F_T(\Delta t) = 1 - \mathrm{e}^{-\Delta t/\tau} \approx 0.30. \tag{3.6}$$

Now imagine that we switch on the detector and three minutes ($t = 180$ s) elapse without a single particle being detected. What is the probability to detect a particle within the next $\Delta t = 30$ s? Intuitively we expect that after three minutes the next particle is "long over-due". But we need the *conditional* probability

$$P(T \le t + \Delta t \,|\, T > t) = \frac{P(t < T \le t + \Delta t)}{P(T > t)}.$$

Here

$$P(t < T \le t{+}\Delta t) = F_T(t + \Delta t){-}F_T(t) = \big[1 - \mathrm{e}^{-(t+\Delta t)/\tau}\big]{-}\big[1 - \mathrm{e}^{-t/\tau}\big] \approx 0.035$$

and $P(T > t) = 1 - F_T(t) = \mathrm{e}^{-t/\tau} \approx 0.117$, thus $P(T \le t + \Delta t \,|\, T > t) = 0.035/0.117 \approx 0.30$, which is the same as (3.6). The fact that we have waited 3 minutes without detecting a particle, has no influence whatsoever on the probability of detection within the next 30 s. $\triangleleft$

Example Customers A and B arrive simultaneously at two bank counters. Their service time is an exponentially distributed random variable with parameters λ_A and λ_B, respectively. What is the probability that B leaves before A?

Let T_A and T_B be random variables measuring the actual service time. The probability that A has not been served until t_A is $\mathrm{e}^{-\lambda_\mathrm{A} t_\mathrm{A}}$. The corresponding probability for customer B is $\mathrm{e}^{-\lambda_\mathrm{B} t_\mathrm{B}}$. Since the waiting processes are independent, their joint probability density is the product of individual probability densities:

$$f_{T_\mathrm{A},T_\mathrm{B}}\big(t_\mathrm{A}, t_\mathrm{B}\big) = \lambda_\mathrm{A}\, \mathrm{e}^{-\lambda_\mathrm{A} t_\mathrm{A}} \cdot \lambda_\mathrm{B}\, \mathrm{e}^{-\lambda_\mathrm{B} t_\mathrm{B}}.$$

Therefore the required probability is

$$P\big(T_\mathrm{B} < T_\mathrm{A}\big) = \int_0^\infty \mathrm{d}t_\mathrm{A} \int_0^{t_\mathrm{A}} f_{T_\mathrm{A},T_\mathrm{B}}\big(t_\mathrm{A}, t_\mathrm{B}\big)\, \mathrm{d}t_\mathrm{B} = \int_0^\infty \mathrm{d}t_\mathrm{A} \lambda_\mathrm{A}\, \mathrm{e}^{-\lambda_\mathrm{A} t_\mathrm{A}} \Big(1 - \mathrm{e}^{-\lambda_\mathrm{B} t_\mathrm{A}}\Big) = \frac{\lambda_\mathrm{B}}{\lambda_\mathrm{A} + \lambda_\mathrm{B}}.$$

The limits are also sensible: if the clerk serving B is very slow ($\lambda_\mathrm{B} \to 0$), then $P\big(T_\mathrm{B} < T_\mathrm{A}\big) \to 0$, while in the opposite case $P\big(T_\mathrm{B} < T_\mathrm{A}\big) \to 1$. $\triangleleft$

The conviction that exponential distributions are encountered only in random processes involving time in some manner, is quite false. Imagine a box containing many balls with diameter d. The fraction of black and white balls is p and $1 - p$, respectively [2]. We draw the balls from the box and arrange them in a line, one touching the other. Suppose we have just drawn a black ball. What is the probability that the distance x between its center and the center of the next black ball is exactly iD, ($i = 1, 2, \ldots$)? We are observing the sequences of drawn balls or "events" of the form

$$\bullet|\ \bullet, \qquad \bullet|\ \circ\ \bullet, \qquad \bullet|\ \circ\ \circ\ \bullet, \qquad \bullet|\ \underbrace{\circ\ \circ \cdots \circ\ \circ}_{(i-1)D}\ \bullet,$$

so the required probability is obviously

$$P(x = iD) = (1 - p)^{i-1} p.$$

Since these events are exclusive, the corresponding probability function is a sum of all probabilities for individual sequences:

$$F_X(x) = P(x \le iD) = p + (1 - p)p + \cdots + (1 - p)^{i-1} p = 1 - (1 - p)^i.$$

Abbreviating $D = 1/n$ and $np = \lambda$ this can be written as

$$F_X(x) = 1 - \left(1 - \frac{\lambda}{n}\right)^{nx},$$

since $i = x/D = nx$. Suppose we take the limits $n \to \infty$ and $p \to 0$ (i.e. there are very few black balls in the box and they have very small diameters), such that λ and x remain unchanged: then $F_X(x) \to 1 - \mathrm{e}^{-\lambda x}$, and the corresponding density is $f_X(x) = \mathrm{d}F_X/\mathrm{d}x = \lambda\,\mathrm{e}^{-\lambda x}$, which is indeed the same as (3.4).

3.2.1 Is the Decay of Unstable States Truly Exponential?

The exponential distribution offers an excellent phenomenological description of the time dependence of the decay of nuclei and other unstable quantum-mechanical states, but its theoretical justification implies many approximations and assumptions, some of which might be questionable in the extremes $t/\tau \ll 1$ and $t/\tau \gg 1$. Further reading can be found in [3] and the classic textbooks [4–6].

3.3 Normal (Gauss) Distribution

It is impossible to resist the temptation of beginning this Section by quoting the famous passage from Poincaré's *Probability calculus* published in 1912 [7]:

> [The law of the distribution of errors] does not follow from strict deduction; many seemingly correct derivations are poorly argued, among them the one resting on the assumption that the probability of deviation is proportional to the deviation. Everyone trusts this law, as I have recently been told by Mr. Lippmann, since the experimentalists believe it is a mathematical theorem, while the theorists think it is an experimental fact.[1]

The normal (Gauss) distribution describes—at least approximately—countless quantities from any sphere of human existence and Nature, for example, diameters of screws being produced in their thousands on a lathe, body masses of people, exam grades and velocities of molecules. A partial explanation and justification for this ubiquity of the Gaussian awaits us in Sect. 6.3 and in particular in Chap. 11. For now let us simply become acquainted with the bell-shaped curve of its two-parameter probability density

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty, \tag{3.7}$$

[1] In the original: "Elle ne s'obtient pas par des déductions rigoureuses; plus d'une démonstration qu'on a voulu en donner est grossière, entre autres celle qui s'appuie sur l'affirmation que la probabilité des écarts est proportionelle aux écarts. Tout le monde y croit cependant, me disait un jour M. Lippmann, car les expérimentateurs s'imaginent que c'est un théorème de mathématiques, et les mathématiciens que c'est un fait expérimental."

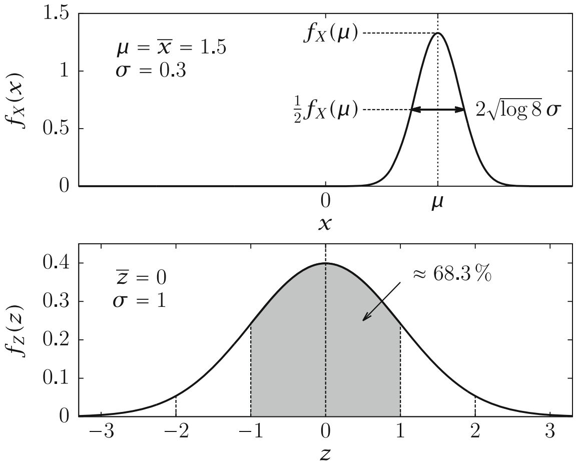


Fig. 3.2 [Top] Normal distribution $N(\mu = 1.5, \sigma^2 = 0.09)$ with average (mean) μ and positive parameter σ determining the peak width. Regardless of σ the area under the curve equals one. [Bottom] Standardized normal distribution $N(0, 1)$

shown in Fig. 3.2 (top).

The definition domain itself makes it clear why the normal distribution is just an approximation in many cases: body masses can not be negative and exam grades can not be infinite. The distribution is symmetric around the value of μ, while the width of its peak is driven by the *standard deviation* σ; at $x = \mu \pm \sigma$ the function f_X has an inflection. The commonly accepted "abbreviation" for the normal distribution is $N(\mu, \sigma^2)$. In Chap. 4 we will see that μ is its *average* or *mean* and σ^2 is its *variance*.

The cumulative distribution function corresponding to density (3.7) is

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right],$$

where

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z \mathrm{e}^{-t^2}\mathrm{d}t, \qquad \mathrm{erf}(-z) = -\mathrm{erf}(z), \tag{3.8}$$

is the so-called *error function* which is tabulated (see Tables D.1 and D.2 and the text below). The probability that a continuous random variable, distributed according to the density (3.7), takes a value between a and b, is

$$P(a \le X \le b) = F_X(b) - F_X(a) = \frac{1}{2}\left[\mathrm{erf}\left(\frac{b-\mu}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{a-\mu}{\sqrt{2}\sigma}\right)\right]. \tag{3.9}$$

3.3.1 Standardized Normal Distribution

When handling normally distributed data it makes sense to eliminate the dependence on the origin and the width by subtracting μ from the variable X and divide out σ, thereby forming a new, *standardized* random variable

$$Z = \frac{X - \mu}{\sigma}.$$

The distribution of Z is then called *standardized normal* and is denoted by $N(0, 1)$ (zero mean, unit variance). It corresponds to the probability density

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \, e^{-z^2/2}, \tag{3.10}$$

while the distribution function is

$$\Phi(z) = P(Z \le z) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^z e^{-t^2/2} \, dt = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}} \right) \right]. \tag{3.11}$$

The values of definite integrals of the standardized normal distribution

$$\frac{1}{\sqrt{2\pi}} \int_0^z e^{-t^2/2} dt = \frac{1}{2} \operatorname{erf}\left(\frac{z}{\sqrt{2}} \right) \tag{3.12}$$

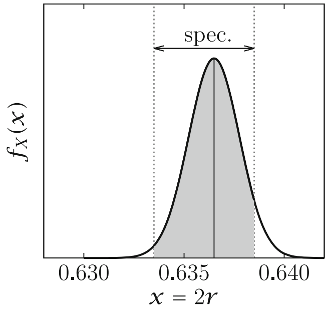
for z between 0 and 5 in steps of 0.01, which is sufficient for everyday use, are listed in Table D.1. The abscissas $x = \mu \pm n\sigma$ or $z = \pm n$ $(n = 1, 2, \ldots)$ are particularly important. The areas under the curve $f_Z(z)$ on these intervals,

$$P_{n\sigma} = P(\mu - n\sigma \le X \le \mu + n\sigma) = P(-n \le Z \le n) = \operatorname{erf}\left(\frac{n}{\sqrt{2}} \right),$$

are equal to

$$P_{1\sigma} \approx 0.683, \quad P_{2\sigma} \approx 0.955, \quad P_{3\sigma} \approx 0.997, \quad P_{4\sigma} \approx 0.9999367\ldots \tag{3.13}$$

(see Fig. 3.2 (bottom)) and tell us what fraction of the data (diameters, masses, exam grades, velocities) is within these—completely arbitrary—intervals and what fraction is outside. For example, if we establish a normal mass distribution of a large sample of massless particles (smeared around zero due to measurement errors), while a few counts lie above 3σ, one may say: "The probability that the particle actually has a non-zero mass, is 0.3%." But if the distribution of measurement error is indeed Gaussian, then even the extreme 0.3% events in the distribution tail may be genuine! However, by increasing the upper bound to 4σ, 5σ,... we can be more and more confident that the deviation is not just a statistical fluctuation. In modern nuclear and particle physics the discovery of a new particle, state or process the mass difference or the signal-to-noise ratio must typically be larger than 5σ.



Example (Adapted from [1].) The diameter of the computer disk axes is described by a normally distributed random variable $X = 2R$ with average $\mu = 0.63650\,\mathrm{cm}$ and standard deviation $\sigma = 0.00127\,\mathrm{cm}$, as shown in the figure. The required specification (shaded area) is $(0.6360 \pm 0.0025)\,\mathrm{cm}$. Let us calculate the fraction of the axes that fulfill this criterion: it is equal to the probability $P(0.6335\,\mathrm{cm} \le X \le 0.6385\,\mathrm{cm})$, which can be computed by converting to the standardized variables $z_1 = (0.6335\,\mathrm{cm} - \mu)/\sigma = -2.36$, corresponding to the lower specified bound, and $z_2 = (0.6385\,\mathrm{cm} - \mu)/\sigma = 1.57$, which corresponds to the upper one. Hence the probability is $P(-2.36 \le Z \le 1.57)$ and can be computed by using the values from Table D.1 (see also Fig. D.1):

$$\begin{aligned} P(-2.36 \le Z \le 1.57) &= P(Z \le 1.57) - P(Z \le -2.36) \\ &= P(Z \le 1.57) - \big[1 - P(Z \le 2.36)\big] \\ &= \tfrac{1}{2} + 0.4418 - \big[1 - \big(\tfrac{1}{2} + 0.4909\big)\big] = 0.9327. \end{aligned}$$

If the machining tool is modified so as to produce the axes with the required diameter of 0.6360 cm, but with the same uncertainty as before, σ, the standardized variables become $z_2 = -z_1 = (0.6335\text{–}0.6360\,\mathrm{cm})/\sigma = 1.97$, thus

$$P(z_1 \le Z \le z_2) = P(-z_2 \le Z \le z_2) = 2\,P(0 \le Z \le z_2) = 2 \cdot 0.4756 = 0.9512.$$

The fraction of useful axes is thereby increased by about 2%.

3.3.2 *Measure of Peak Separation*

A practical quantity referring to the normal distribution is its *full width at half-maximum* (FWHM), see double-headed arrow in Fig. 3.2 (top). It can be obtained by simple calculation: $f_X(x)/f_X(0) = 1/2$ or $\exp[-x^2/(2\sigma^2)] = 1/2$, hence $x = \sigma\sqrt{2\log 2}$. The FWHM is just twice this number,

$$\mathrm{FWHM} = 2\sqrt{2\log 2}\,\sigma \approx 2.35\,\sigma.$$

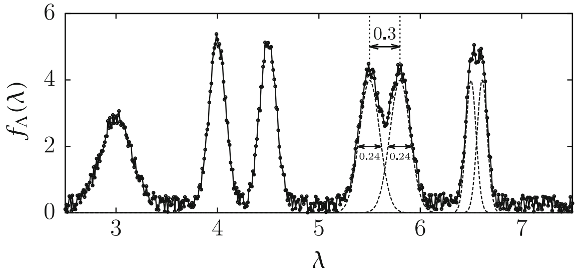


Fig. 3.3 Illustration of the measure of peak separation. The centers of the fourth and fifth peak from the left are 0.3 apart, which is just slightly above the value of FWHM = 0.24 for individual peaks, so they can still be separated. The three leftmost peaks can also be separated. The structure at the right consists of two peaks which are too close to each other to be cleanly separated. In practice, similar decisions are almost always complicated by the presence of noise

FWHM offers a measure of how well two Gaussian peaks in a physical spectrum can be separated. By convention we can distinguish neighboring peaks with equal amplitudes and equal σ if their centers are at least FWHM apart (Fig. 3.3).

3.4 Maxwell Distribution

The Maxwell distribution describes the velocities of molecules in thermal motion in thermodynamic equilibrium. In such motion the velocity components of each molecule, $\boldsymbol{v} = (v_x, v_y, v_z)$, are stochastically independent, and the average velocity (as a vector) is zero. The directions x, y and z correspond to kinetic energies $mv_x^2/2$, $mv_y^2/2$ and $mv_z^2/2$, and the probability density in velocity space at given temperature T decreases exponentially with energy. The probability density for $\boldsymbol{v}$ is the product of three one-dimensional Gaussian densities:

$$f_V(\boldsymbol{v}) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^3 \exp\left(-\frac{v_x^2 + v_y^2 + v_z^2}{2\sigma^2}\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{3/2} \exp\left(-\frac{v^2}{2\sigma^2}\right), \tag{3.14}$$

where $v^2 = v_x^2 + v_y^2 + v_z^2$ and $\sigma^2 = k_\mathrm{B}T/m$. The distribution over $\boldsymbol{v}$ is spherically symmetric, so the appropriate distribution in magnitudes $v = |\boldsymbol{v}|$ is obtained by evaluating $f_V(\boldsymbol{v})$ in a thin spherical shell with volume $4\pi v^2 \mathrm{d}v$, thus

$$f_V(v) = \frac{\mathrm{d}F_V}{\mathrm{d}v} = \left(\frac{m}{2\pi k_\mathrm{B}T}\right)^{3/2} 4\pi v^2 \exp\left(-\frac{mv^2}{2k_\mathrm{B}T}\right). \tag{3.15}$$

An example of such distribution for nitrogen molecules at temperatures 193 and 393 K is shown in Fig. 3.4 (left).

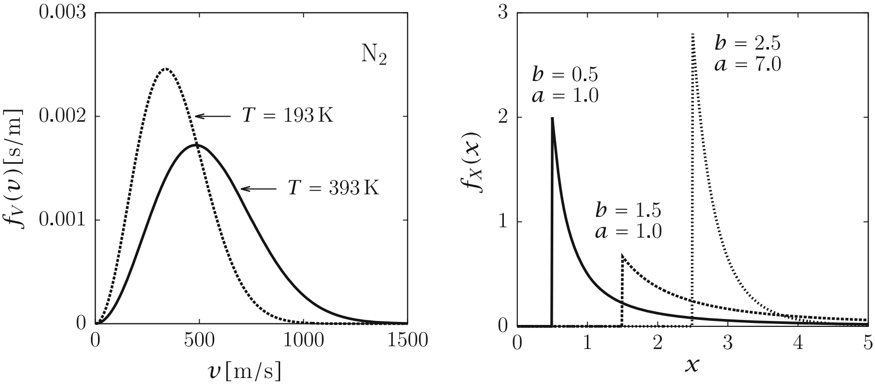


Fig. 3.4 [Left] Maxwell distribution of velocities of nitrogen molecules at $T = 193\,\mathrm{K}$ and $T = 393\,\mathrm{K}$. See also Fig. 4.1 (*right*) and Problem 3.10.4. [Right] Pareto distribution with parameters $b \equiv x_{\min}$ (minimum value on the abscissa) and a (shape parameter)

3.5 Pareto Distribution

Probability distributions of many quantities that can be interpreted as random variables have relatively narrow ranges of values. The height of an average adult, for example, is 180 cm, but nobody is 50 or 500 cm tall. The data acquired by the WHO [8] show that the body mass index (ratio of the mass in kilograms to the square of the height in meters) is restricted to a range between ≈15 and ≈50.

But one also frequently encounters quantities that span many orders of magnitude, for example, the number of inhabitants of human settlements (ranging from a few tens in a village to tens of millions in modern city conglomerates). Similar “processes” with a large probability for small values and small probability for large values are: frequency of specific given names, size of computer files, number of citations of scientific papers, number of web-page accesses and the quantities of sold merchandise (see Example on p. 97), but also quantities measured in natural phenomena, like step lengths in random walks (anomalous diffusion), magnitudes of earthquakes, diameters of lunar craters or the intensities of solar X-ray bursts [9–11]. A useful approximation for the description of such quantities is the Pareto (power law) distribution with the probability density

$$f_X(x) = \frac{ab^a}{x^{a+1}} = \frac{a}{b}\left(\frac{b}{x}\right)^{a+1}, \qquad 0 < b \le x, \tag{3.16}$$

where b is the minimal allowed x (Fig. 3.4 (right)), and a is a parameter which determines the relation between the prominence of the peak near the origin and the strength of the tail at large x. It is this flexibility in parameters that renders the Pareto distribution so useful in modeling the processes and phenomena enumerated above. As an example, Fig. 3.5 (left) shows the distribution of the lunar craters in terms

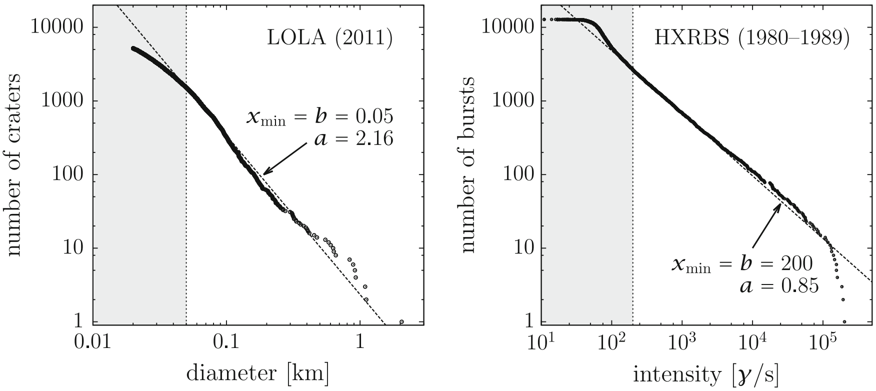


Fig. 3.5 [Left] Distribution of lunar craters with respect to their diameter, as determined by researchers of the Lunar Orbiter Laser Altimeter (LOLA) project [12, 13] up to 2011. [Right] The distribution of hard X-rays in terms of their intensity, measured by the Hard X-Ray Burst Spectrometer (HXRBS) between 1980 and 1989 [14]. The straight lines represent the approximate power-law dependencies, also drawn in the shaded areas, although the Pareto distributions commence only at their right edges ($x \geq x_{\min}$)

of their diameter, and Fig. 3.5 (right) shows the distribution of solar X-ray bursts in terms of their intensity.

The Pareto distribution is normalized on the interval $[b, \infty)$ and frequently one does not use its distribution function $F_X(x) = P(X \leq x)$ but rather its complement,

$$1 - F_X(x) = P(X > x) = \int_x^\infty f_X(t)\,\mathrm{d}t = ab^a \int_x^\infty \frac{\mathrm{d}t}{t^{a+1}} = \left(\frac{b}{x}\right)^a, \qquad x \geq b, \tag{3.17}$$

as it is easier to normalize and compare it to the data: the ordinate simply specifies the number of data points (measurements, events) that were larger than the chosen value on the abscissa. By plotting the data in this way, one avoids histogramming in bins, which is not unique. The values $x_{\min} = b$ should not be set to the left edge of the interval on which measurements are available (e.g. 20 m in LOLA measurements), but to the value above which the description in terms of a power-law appears reasonable ($\approx$50 m). The parameter a can be determined by fitting the power function to the data, but in favor of better stability [9] we recommend the formula

$$a = n\left[\sum_{i=1}^n \log\frac{x_i}{b}\right]^{-1},$$

which we derive later ((8.11)).

Hint If we wish to plot the cumulative distribution for the data $\{x_i, y_i\}_{i=1}^n$, we can use the popular graphing tool GNUPLOT. We first sort the data, so that x_i are arranged in increasing order (two-column file `data`). The cumulative distribution can then be plotted by the command

```
gnuplot > plot "data" using 1 : ($0/n) with lines
```

3.5.1 Estimating the Maximum x in the Sample

Having at our disposal a sample of n measurements presumably originating from a power-law distribution with known parameters a and b, a simple consideration allows us to estimate the value of the largest expected observation [9]. Since we are dealing with a continuous distribution, we should refer to the probability that its value falls in the interval $[x, x + \mathrm{d}x]$. The probability that a data point is larger than x, is given by (3.17), while the probability for the opposite event is $1 - P(X > x)$. The probability that a particular measurement will be in $[x, x + \mathrm{d}x]$ and that *all others will be smaller* is therefore $[1 - P(X > x)]^{n-1} f_X(x)\,\mathrm{d}x$. Because the largest measurement can be chosen in n ways, the total probability is

$$n\,[1 - P(X > x)]^{n-1}\, f_X(x)\,\mathrm{d}x.$$

The expected value of the largest measurement—such quantities will be discussed in the next chapter—is obtained by integrating x, weighted by the total probability, over the whole definition domain:

$$\begin{aligned}\overline{x}_{\max} &= n\int_b^\infty x f_X(x)\,[1 - P(X > x)]^{n-1}\,\mathrm{d}x = na\int_b^\infty \left(\frac{b}{x}\right)^a \left[1 - \left(\frac{b}{x}\right)^a\right]^{n-1} \mathrm{d}x \\ &= nb\int_0^1 t^{n-1}(1-t)^{-1/a}\,\mathrm{d}t = nb\,B\left(n, \frac{a-1}{a}\right),\end{aligned}$$

where $B(p, q)$ is the beta function. We have substituted $t = 1 - (b/x)^a$ in the intermediate step. For the sample in Fig. 3.5 (left), which contains $n = 1513$ data points, $a = 2.16$ and $b = 0.05$ km, we obtain $x_{\max} \approx 2.5$ km. If the sample were ten times as large, we would anticipate $x_{\max} \approx 7.1$ km.

3.6 Cauchy Distribution

The Cauchy distribution with probability density

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \qquad -\infty < x < \infty, \tag{3.18}$$

is already familiar to us from the Example on p. 41. In fact, we should have discussed it along with the exponential, as the Fourier transform of the exponential function in the time scale is the Cauchy function in the energy scale:

$$g(t) = \mathrm{e}^{-|t|/\tau} \quad \Longrightarrow \quad \frac{1}{2\pi}\int_{-\infty}^{\infty} g(t)\,\mathrm{e}^{-\mathrm{i}\,2\pi\nu t}\,\mathrm{d}t = \frac{1}{\pi}\frac{1/\tau}{(1/\tau)^2 + 4\pi^2\nu^2}. \tag{3.19}$$

In other words, the energy distribution of the states decaying exponentially in time is given by the Cauchy distribution. It is therefore suitable for the description of spectral line shapes in electromagnetic transitions of atoms and molecules (Fig. 3.6 (left)) or for modeling the energy dependence of cross-sections for the formation of resonances in hadronic physics (Fig. 3.6 (right)). With this in mind, it makes sense to furnish it with the option of being shifted by x_0 and with a parameter s specifying its width:

$$f_X(x; x_0, s) = \frac{1}{\pi}\frac{s}{s^2 + (x - x_0)^2}. \tag{3.20}$$

In spectroscopy the Cauchy distribution is also known as the Lorentz curve, while in the studies of narrow, isolated resonant states in nuclear and particle physics it is called the Breit–Wigner distribution: in this case it is written as

$$f(W; W_0, \Gamma) = \frac{1}{\pi}\frac{\Gamma/2}{(W - W_0)^2 + \Gamma^2/4},$$

where W_0 is the resonance energy and Γ is the resonance width.

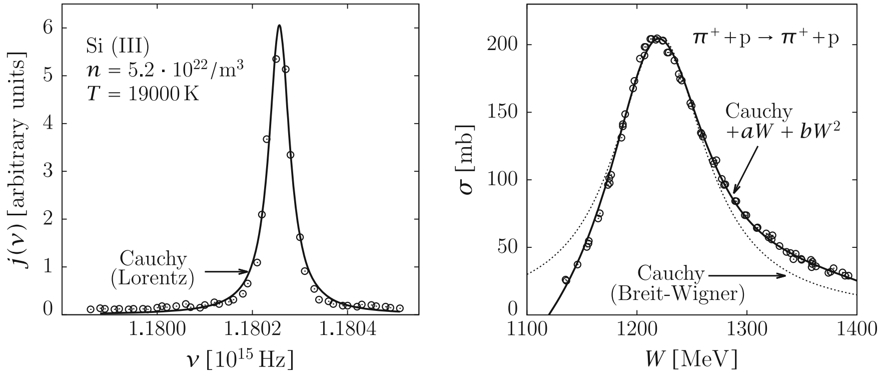


Fig. 3.6 [Left] A spectral line in the emission spectrum of silicon (centered at $\lambda = 254.182\,\mathrm{nm}$) at a temperature of 19,000 K and particle density $5.2 \times 10^{22}/\mathrm{m}^3$ [15], along with the Cauchy (Lorentz) approximation. Why the agreement with the measured values is imperfect and how it can be improved will be revealed in Problem 6.9.2. [Right] Energy dependence of the cross-section for scattering of charged pions on protons. In this process a resonance state is formed whose energy distribution in the vicinity of the maximum can also be described by the Cauchy (Breit–Wigner) distribution

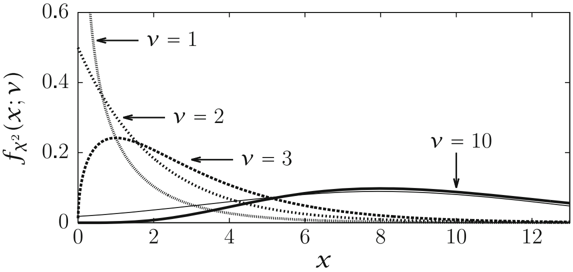


Fig. 3.7 The density of the χ^2 distribution for four different parameters (degrees of freedom) ν. The maximum of the function $f_{\chi^2}(x;\nu)$ for $\nu > 2$ is located at $x = \nu - 2$. For large ν the χ^2 density converges to the density of the normal distribution with average $\nu - 2$ and variance 2ν. The thin curve just next to $f_{\chi^2}(x; 10)$ denotes the density of the $N(8, 10)$ distribution

3.7 The χ^2 distribution

The χ^2 distribution, a one-parameter probability distribution with the density

$$f_{\chi^2}(x;\nu) = \frac{1}{2^{\nu/2}} \frac{1}{\Gamma\big(\nu/2\big)} x^{\nu/2-1} \mathrm{e}^{-x/2}, \qquad x > 0, \tag{3.21}$$

will play its role in the our discussion on statistics (Chaps. 7–10). The parameter ν is called the *number of degrees of freedom*. The probability density of the χ^2 distribution for four values of ν is shown in Fig. 3.7. The corresponding distribution function is

$$F_{\chi^2}(x;\nu) = P\,(X \le x) = \frac{1}{2^{\nu/2}} \frac{1}{\Gamma\big(\nu/2\big)} \int_0^x t^{\nu/2-1} \mathrm{e}^{-t/2}\,\mathrm{d}t.$$

In practical work one usually does not need this definite integral but rather the answer to the opposite question, the cut-off value x at given P. These values are tabulated: see Fig. D.1 (top right) and Table D.3.

3.8 Student's Distribution

The Student's distribution (or the t distribution)[2] is also a one-parameter probability distribution that we shall encounter in subsequent chapters devoted to statistics. Its density is

[2] The Student's distribution acquired its first peculiar name from a paper [16] that an English statistician W.S. Gosset published under the pseudonym Student, and the second one from a specific random variable (see formula (7.18)).

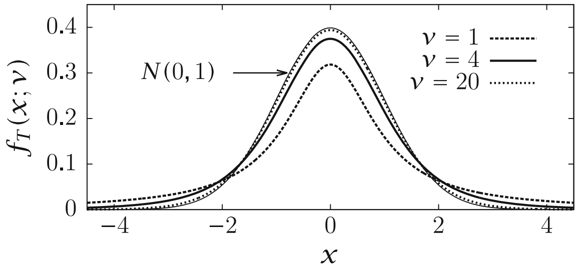


Fig. 3.8 The density of the Student's (t) distribution with $\nu = 1$, $\nu = 4$ and $\nu = 20$ degrees of freedom. The distribution is symmetric about the origin and approaches the standardized normal distribution $N(0, 1)$ with increasing ν (*thin curve*), from which it is hardly discernible beyond $\nu \approx 30$

$$f_T(x;\nu) = \frac{1}{\sqrt{\nu}\, B\left(\frac{\nu}{2}, \frac{1}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \qquad -\infty < x < \infty, \tag{3.22}$$

where ν is the number of degrees of freedom and B is the beta function. The graphs of its density for $\nu = 1$, $\nu = 4$ and $\nu = 20$ are shown in Fig. 3.8. In the limit $\nu \to \infty$ the Student's distribution tends to the standardized normal distribution.

3.9 *F* distribution

The F distribution is a two-parameter distribution with the probability density

$$f_F(x;\nu_1,\nu_2) = \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{\Gamma\big((\nu_1+\nu_2)/2\big)}{\Gamma\big(\nu_1/2\big)\Gamma\big(\nu_2/2\big)}\, x^{\nu_1/2-1} \left(1 + \frac{\nu_1}{\nu_2}\, x\right)^{-(\nu_1+\nu_2)/2}, \tag{3.23}$$

where ν_1 is the number of degrees of freedom "in the numerator" and ν_2 is the number of degrees of freedom "in the denominator". Why this distinction is necessary will become clear in Sect. 7.2.3: there we shall compare *ratios* of particular random variables, distributed according to (3.23). The probability densities of the F distribution are shown in Fig. 3.9 for several typical (ν_1, ν_2) pairs.

3.10 Problems

3.10.1 In-Flight Decay of Neutral Pions

A complicated transformation of a uniform distribution may still turn out to be a uniform distribution, as we learn by solving the classical problem in relativistic

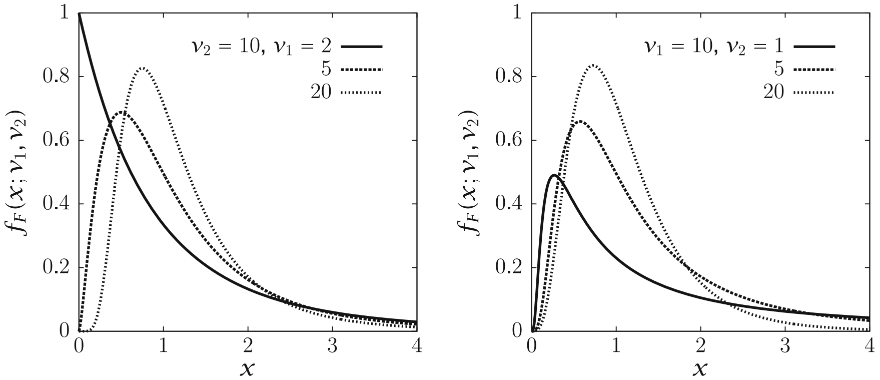


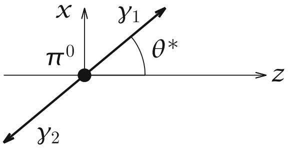
Fig. 3.9 [Left] The probability density of the F distribution for $\nu_1 = 10$ degrees of freedom (numerator) and three different degrees of freedom ν_2 (denominator). [Right] The density of the F distribution for $\nu_2 = 10$ and three different values of ν_1

kinematics, the in-flight neutral pion decay to two photons, $\pi^0 \to \gamma + \gamma$. Calculate the energy distribution of the decay photons, $\mathrm{d}N/\mathrm{d}E_\gamma$!

✎ Let the π^0 meson fly in the laboratory frame along the z-axis with velocity v_π. The decay in the π^0 rest frame is isotropic. Due to azimuthal symmetry (ϕ) this implies a uniform distribution over the *cosine* of the angle θ^* (see Sect. C.2.2):

$$f(\cos\theta^*) = \frac{\mathrm{d}N}{\mathrm{d}(\cos\theta^*)} = \frac{1}{2}, \qquad -1 \le \cos\theta^* \le 1,$$

where θ^* is the emission angle of the first photon in the rest frame, as shown in the figure:



The energy distribution of the photons is obtained by the derivative chain-rule:

$$\frac{\mathrm{d}N}{\mathrm{d}E_\gamma} = \frac{\mathrm{d}N}{\mathrm{d}(\cos\theta^*)}\frac{\mathrm{d}(\cos\theta^*)}{\mathrm{d}E_\gamma} = \frac{1}{2}\frac{\mathrm{d}(\cos\theta^*)}{\mathrm{d}E_\gamma}. \tag{3.24}$$

We therefore need to establish a relation between θ^* and E_γ, and it is offered by the Lorentz transformation from the π^0 rest frame to the laboratory frame. Of course, the energies of the photons in the rest frame are equal, $E^*_{\gamma,1} = E^*_{\gamma,2} = E^*_\gamma = p^*_\gamma c = m_\pi c^2/2$, and their four-vectors are

$$\left(E^*_{\gamma,i},\ p^*_{\gamma x,i}c,\ p^*_{\gamma y,i}c,\ p^*_{\gamma z,i}c\right) = \tfrac{1}{2}m_\pi c^2\left(1, \pm\sin\theta^*, 0, \pm\cos\theta^*\right), \qquad i = 1, 2.$$

The Lorentz transformation that gives us their energies in the laboratory frame is

$$E_{\gamma,i} = \gamma E^*_{\gamma,i} + \gamma\beta p^*_{\gamma z,i} c = \tfrac{1}{2} m_\pi c^2 \gamma \bigl(1 \pm \beta \cos\theta^*\bigr),$$

where $\beta = v_\pi/c = p_\pi c/E_\pi$ and $\gamma = 1/\sqrt{1-\beta^2} = E_\pi/(m_\pi c^2)$. It follows that

$$\frac{\mathrm{d}E_{\gamma,i}}{\mathrm{d}(\cos\theta^*)} = \tfrac{1}{2} m_\pi c^2 \gamma\beta = \tfrac{1}{2} p_\pi c$$

i.e.

$$\frac{\mathrm{d}(\cos\theta^*)}{\mathrm{d}E_\gamma} = \frac{2}{p_\pi c}.$$

When this is inserted in (3.24), we obtain the required energy distribution, which is indeed uniform:

$$\frac{\mathrm{d}N}{\mathrm{d}E_\gamma} = \frac{1}{p_\pi c},$$

namely on the interval between the minimal and maximal values

$$E_\gamma^{\min} = \tfrac{1}{2}(E_\pi - p_\pi c) = \tfrac{1}{2}E_\pi(1-\beta), \qquad E_\gamma^{\max} = \tfrac{1}{2}(E_\pi + p_\pi c) = \tfrac{1}{2}E_\pi(1+\beta).$$

Let us check our findings by a simple simulation, observing the decay of pions with a velocity of $0.7c$ ($\beta = 0.7$). We use a computer to generate 100000 uniformly distributed values $-1 \le \cos\theta^* \le 1$ (Fig. 3.10 (left)), and then use each of these values to calculate the photon energies in the laboratory frame, $E_{\gamma,1}$ and $E_{\gamma,2}$. A uniform distribution over E_γ on the interval between $E_\gamma^{\min}$ and $E_\gamma^{\max}$ should appear. It can be seen in Fig. 3.10 (right) that we were not mistaken.

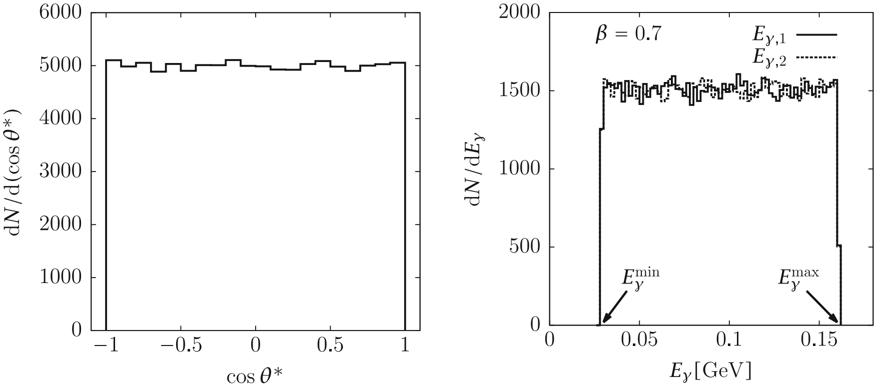
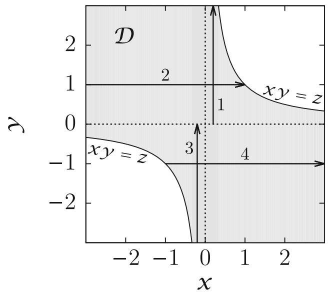


Fig. 3.10 The $\pi^0 \to \gamma + \gamma$ decay. [Left] Uniform distribution of events over $\cos\theta^*$ in the π^0 rest frame. [Right] Uniform energy distribution of the decay pions in the laboratory frame

3.10.2 Product of Uniformly Distributed Variables

(Adapted from [17].) Let two continuous random variables X and Y be described by a known probability density $f_{X,Y}(x, y)$. ① Calculate the probability density $f_Z(z)$ of the product random variable $Z = XY$ in the most general case and in the case that X and Y are independent. ② Discuss the special case of independent variables X and Y, both of which are uniformly distributed on the interval $(0, 1)$.



✎ Define the domain $\mathcal{D} = \{(x, y) : xy < z\}$ (shown for positive z as the shaded region in the figure) which determines the distribution function of the variable Z:

$$P\big((X, Y) \in \mathcal{D}\big) = F_Z(z) = \int_0^{\infty} \mathrm{d}y \int_{-\infty}^{z/y} f_{X,Y}(x, y)\,\mathrm{d}x + \int_{-\infty}^{0} \mathrm{d}y \int_{z/y}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x.$$

To facilitate the determination of integration boundaries, the intervals of four integrations in this equation—read from left to right—are denoted by numbers 1 to 4 in the figure. (The derivation for negative z proceeds analogously.) ① The corresponding probability density is then obtained by differentiation:

$$f_Z(z) = \frac{\mathrm{d}F_Z(z)}{\mathrm{d}z} = \int_0^{\infty} \frac{1}{y}\, f_{X,Y}\left(\frac{z}{y}, y\right) \mathrm{d}y - \int_{-\infty}^{0} \frac{1}{y}\, f_{X,Y}\left(\frac{z}{y}, y\right) \mathrm{d}y.$$

If X and Y are independent, possessing probability densities $f_X(x)$ and $f_Y(y)$, one has $f_{X,Y}(x, y) = f_X(x) f_Y(y)$, thus

$$f_Z(z) = \int_0^{\infty} \frac{1}{y}\, f_X\left(\frac{z}{y}\right) f_Y(y)\,\mathrm{d}y - \int_{-\infty}^{0} \frac{1}{y}\, f_X\left(\frac{z}{y}\right) f_Y(y)\,\mathrm{d}y. \qquad (3.25)$$

② The product of uniformly distributed variables X and Y is always positive and less than 1, hence the probability density $f_Z(z)$ of the variable $Z = XY$ may be non-zero only on the interval $(0, 1)$. On this interval it can be determined by using (3.25), in which only the first term survives due to this very requirement, and even here the integrand is positive only if $0 < z/y < 1$ and $0 < y < 1$, i.e. when $z < y < 1$. It follows that

$$f_Z(z) = \int_z^1 \frac{\mathrm{d}y}{y} = -\log z, \qquad 0 < z < 1,$$

while $f_Z(z) = 0$ elsewhere.

3.10.3 Joint Distribution of Exponential Variables

Let X and Y be independent random variables distributed exponentially with parameters $\lambda_1 = 1$ and $\lambda_2 = 3$,

$$f_X(x) = \lambda_1\, \mathrm{e}^{-\lambda_1 x}, \qquad f_Y(y) = \lambda_2\, \mathrm{e}^{-\lambda_2 y}, \qquad x, y \geq 0.$$

Imagine a square region $S = [0, a] \times [0, a]$. ① Calculate the value of a, for which the probability that a randomly drawn (x, y) pair falls into S, equals $1/2$. ② Calculate the conditional joint probability density of the variables X and Y, given that $X \geq a$ and $Y \geq a$.

✎ The variables X and Y are independent, hence their joint probability density is

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) = \lambda_1 \lambda_2\, \mathrm{e}^{-\lambda_1 x} \mathrm{e}^{-\lambda_2 y}, \qquad x, y \geq 0.$$

The probability that a random pair of values (x, y) finds itself in S, equals

$$P_{aa} \equiv P\big(0 \leq X \leq a, 0 \leq Y \leq a\big) = \int_0^a \int_0^a f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y = \Big(1 - \mathrm{e}^{-\lambda_1 a}\Big)\Big(1 - \mathrm{e}^{-\lambda_2 a}\Big).$$

① We are looking for a such that $P_{aa} = 1/2$. This equation is best solved by Newton's method, in spite of its known pitfalls: with the function $f(x) = (1 - \mathrm{e}^{-\lambda_1 x})(1 - \mathrm{e}^{-\lambda_2 x}) - 1/2$ (plot it!) and its derivative $f'(x) = \lambda_1 \mathrm{e}^{-\lambda_1 x} + \lambda_2 \mathrm{e}^{-\lambda_2 x} - (\lambda_1 + \lambda_2)\mathrm{e}^{-(\lambda_1+\lambda_2)x}$ we start the iteration $x_{n+1} = x_n - f(x_n)/f'(x_n)$, $n = 0, 1, 2, \ldots$ With the initial approximation $x_0 = 0.5$ just a few iteration steps lead to $a = x_\infty \approx 0.7987$.

② We first form the conditional distribution function

$$\begin{aligned} F_{X,Y}\big(x, y | X \geq a, Y \geq a\big) &= P\big(X \leq x, Y \leq y | X \geq a, Y \geq a\big) \\ &= \frac{P\big(a \leq X \leq x \cap a \leq Y \leq y\big)}{P\big(X \geq a \cap Y \geq a\big)} = \frac{P(a \leq X \leq x) P(a \leq Y \leq y)}{P(X \geq a) P(Y \geq a)} \\ &= \frac{\int_a^y \mathrm{d}v \int_a^x f_{X,Y}(u, v)\, \mathrm{d}u}{\int_a^\infty \mathrm{d}v \int_a^\infty f_{X,Y}(u, v)\, \mathrm{d}u} = \frac{\big(\mathrm{e}^{-\lambda_1 a} - \mathrm{e}^{-\lambda_1 x}\big)\big(\mathrm{e}^{-\lambda_2 a} - \mathrm{e}^{-\lambda_2 y}\big)}{\mathrm{e}^{-\lambda_1 a} \mathrm{e}^{-\lambda_2 a}}, \end{aligned}$$

where we have taken into account that X and Y are independent. The probability density can then be calculated by differentiating $F_{X,Y}$ with respect to x and y:

$$f_{X,Y}\big(x, y | X \geq a, Y \geq a\big) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}\big(x, y | X \geq a, Y \geq a\big) = \lambda_1 \lambda_2 \mathrm{e}^{-\lambda_1 (x-a)} \mathrm{e}^{-\lambda_2 (y-a)}.$$

We should also check the normalization which must be fulfilled—as for any probability density—also for the calculated conditional density. Indeed we find

$$\int_a^\infty \int_a^\infty f_{X,Y}\bigl(x, y|X \geq a, Y \geq a\bigr)\,\mathrm{d}x\,\mathrm{d}y = 1,$$

where $[a, \infty]^2$ is the definition domain of the conditional joint probability density.

3.10.4 Integral of Maxwell Distribution over Finite Range

What fraction of nitrogen ($\mathrm{N_2}$) molecules at temperature $T = 393\,\mathrm{K}$ have velocities between $v_1 = 500$ and $v_2 = 1000\,\mathrm{m/s}$, if the velocity distribution is of the Maxwell type (see Fig. 3.4 (left))?

✎ Let us rewrite (3.15) in a slightly more compact form

$$f_V(v) = \sqrt{\frac{16\alpha^3}{\pi}}\, v^2\, \mathrm{e}^{-\alpha v^2}, \qquad \alpha = \sqrt{\frac{m}{2k_\mathrm{B}T}}\ .$$

The required fraction of molecules is equal to the definite integral of the probability density from v_1 to v_2,

$$P(v_1 \leq V \leq v_2) = \int_{v_1}^{v_2} f_V(v)\,\mathrm{d}v = \sqrt{\frac{16\alpha^3}{\pi}} \int_{v_1}^{v_2} v^2\,\mathrm{e}^{-\alpha v^2}\,\mathrm{d}v.$$

Such integrals are typically handled by resorting to integration by parts, in which the power of the variable x in the integrand is gradually reduced:

$$\begin{aligned}\int x^n\,\mathrm{e}^{-\alpha x^2}\mathrm{d}x &= -\frac{1}{2\alpha}\int x^{n-1}(-2\alpha x)\,\mathrm{e}^{-\alpha x^2}\,\mathrm{d}x \\ &= -\frac{1}{2\alpha}\left[x^{n-1}\,\mathrm{e}^{-\alpha x^2} - \int (n-1)x^{n-2}\,\mathrm{e}^{-\alpha x^2}\,\mathrm{d}x\right].\end{aligned}$$

In our case we only need the integral with $n = 2$, therefore

$$I(v) = \int\limits_0^v x^2\,\mathrm{e}^{-\alpha x^2}\,\mathrm{d}x = \frac{1}{2\alpha}\int\limits_0^v \mathrm{e}^{-\alpha x^2}\,\mathrm{d}x - \frac{1}{2\alpha}\,v\,\mathrm{e}^{-\alpha v^2} = \frac{\sqrt{\pi}}{4\alpha^{3/2}}\,\mathrm{erf}\,\bigl(\sqrt{\alpha}\,v\bigr) - \frac{v}{2\alpha}\,\mathrm{e}^{-\alpha v^2}.$$

From Table D.2 we read off $\mathrm{erf}(\sqrt{\alpha}v_1) \approx 0.4288$ and $\mathrm{erf}(\sqrt{\alpha}v_2) \approx 0.4983$, and all that is needed is to merge the expressions to

$$P(v_1 \le V \le v_2) = \sqrt{\frac{16\alpha^3}{\pi}}\big[I(v_2) - I(v_1)\big] \approx 0.5065.$$

(The result by computing the erf functions accurately is 0.5066.)

3.10.5 Decay of Unstable States and the Hyper-exponential Distribution

Organic scintillator is a material in which charged particles promote electrons to excited states, which get rid of the excess energy by photon emission. The time dependence of the intensity of emitted light can be approximated by a sum of two independent excitation mechanisms (occurring almost instantaneously) and de-excitations proceeding with two different decay times, as shown in Fig. 3.11. ① Write down the corresponding probability density and the functional form of the decay curve. ② Generalize the expressions to multiple time components. Does the same physical picture apply to a mixture of radioactive isotopes, if each of them has only a single decay mode?

✎ The mechanisms of light generation in scintillators are poorly understood, but the predominant opinion seems to be that the type of relaxation (fast or slow) is determined already during excitation. ① We are thus dealing with exclusive (incompatible) events, hence the probability density is

$$f_T(t) = P\lambda_1\,\mathrm{e}^{-\lambda_1 t} + (1-P)\lambda_2\,\mathrm{e}^{-\lambda_2 t}.$$

The time dependence of the light curve is then given by the distribution function:

$$N(t)/N_0 = 1 - F_T(t) = 1 - \int_0^t f_T(t')\,\mathrm{d}t' = P\,\mathrm{e}^{-\lambda_1 t} + (1-P)\,\mathrm{e}^{-\lambda_2 t}.$$

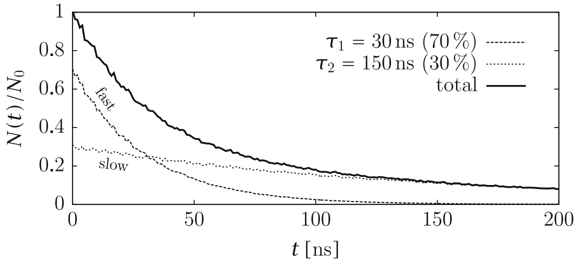


Fig. 3.11 Typical time dependence of a light pulse emanating from an organic scintillator: in this case, the total intensity consists of a fast relaxation component with decay time $\tau_1 = 30$ ns (frequency 70%) and a slow one with decay time $\tau_2 = 150$ ns (30%)

② Obviously one can generalize this to multiple (k) time components by writing

$$f_T(t) = \sum_{i=1}^{k} P_i \lambda_i \, \mathrm{e}^{-\lambda_i t}, \qquad \sum_{i=1}^{k} P_i = 1. \tag{3.26}$$

The distribution with such probability density is known as the *k-phase hyper-exponential distribution*. It can be used to model the superposition of k independent events, e.g. the response time of a system of k parallel computer servers, in which the ith server is assigned with probability P_i to handle our request, and the distribution of its service time is exponential with parameter λ_i (Fig. 3.12 (left)). Such a distribution also describes the lifetime of a product manufactured on several parallel assembly lines or in factories with different levels of manufacturing quality.

At first sight, radioactive decay in a sample containing various isotopes (for example, a mixture of ^{137}Cs, ^{235}U and ^{241}Am) resembles such a k-phase process. But the key difference is that the decays of individual isotopes are not mutually exclusive: in a chosen time interval Δt we might detect the decay of a single isotope, two, or all three. In this case the hyper-exponential distribution is not justified.

Similar conclusions can be drawn for the decay of unstable particles with multiple decay modes, each occurring with a distinct probability. Suppose that particle X decays into the final state A consisting of two or more lighter particles. The usual decay law (3.5) applies:

$$\dot{N}_{\mathrm{X}\to\mathrm{A}}(t) = -\lambda_{\mathrm{A}} N(t).$$

If multiple final states A, B, C, ... are allowed, we must sum over all contributions: the time derivative of the number of particles still available for decay at time t is

$$\dot{N}(t) = \dot{N}_{\mathrm{X}\to\mathrm{anything}}(t) = \dot{N}_{\mathrm{X}\to\mathrm{A}}(t) + \dot{N}_{\mathrm{X}\to\mathrm{B}}(t) + \cdots = -\big(\lambda_{\mathrm{A}} + \lambda_{\mathrm{B}} + \cdots\big) N(t) \equiv -\lambda N(t).$$

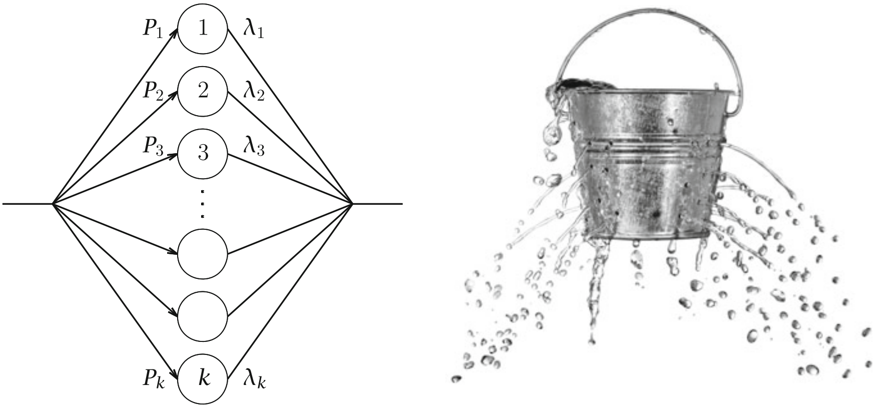


Fig. 3.12 [Left] A set of k parallel independent processes ("phases") with a single output, described by the hyper-exponential distribution. [Right] An illustration of the decay modes of a sample of unstable particles

The extinction of N is therefore driven by a single time constant, $\lambda = \lambda_A + \lambda_B + \cdots$! Just prior to the decay, Nature does not think about the type of the final state, but rather just chooses the time of the decay by exponential law with parameter λ,

$$N(t) = N_0\,\mathrm{e}^{-\lambda t} = N_0\,\mathrm{e}^{-t/\tau},$$

where τ is the average decay time. Instead of τ we sometimes prefer to specify the conjugated variable in the Heisenberg sense (time and energy, position and linear momentum, angle and angular momentum), known as the *total decay width:*

$$\Gamma = \frac{\hbar}{\tau} = \hbar\lambda = \hbar\big(\lambda_A + \lambda_B + \cdots\big) = \Gamma_A + \Gamma_B + \cdots .$$

The total width Γ is a sum of the *partial widths* Γ_A, Γ_B, ... It is only at the very moment of decay that the particle randomly "picks" a certain final state. The probabilities for the transitions to specific final states can be expressed by *branching ratios* or *branching fractions:* for individual decay modes we have

$$\mathrm{Br}_A = \frac{\dot{N}_{X\to A}}{\dot{N}_{X\to \text{anything}}} = \frac{\Gamma_A}{\Gamma}, \qquad \mathrm{Br}_B = \frac{\Gamma_B}{\Gamma}, \qquad \mathrm{Br}_C = \frac{\Gamma_C}{\Gamma}, \qquad \ldots \tag{3.27}$$

Conservation of probability (a particle *must* decay into some final state after all) of course ensures

$$\mathrm{Br}_A + \mathrm{Br}_B + \mathrm{Br}_C + \cdots = 1.$$

As an example, Table 3.1 shows the partial widths and branching fractions in the decay of the Z^0 bosons produced in collisions of electrons and positrons at invariant energies around 91 GeV; see Fig. 3.12 (right). From the total decay width we compute the average decay time $\tau = \hbar/\Gamma \approx 2.6 \times 10^{-25}$s. The energy dependence of the Z^0 resonance is described by the Breit-Wigner distribution (Fig. 3.6 (right)) with the center at approximately 91.2 GeV and a width of about 2.5 GeV.

Table 3.1 The dominant decay modes of the Z^0 boson, the corresponding partial decay widths and the branching fractions

Decay mode	Width (MeV)	Branching fraction (%)
$Z^0 \to e^+e^-$	83.9 ± 0.1	3.363 ± 0.004
$Z^0 \to \mu^+\mu^-$	84.0 ± 0.1	3.366 ± 0.007
$Z^0 \to \tau^+\tau^-$	84.0 ± 0.1	3.367 ± 0.008
$Z^0 \to \nu_l\bar{\nu}_l$ ($l = e, \mu, \tau$)	499.0 ± 1.5	20.00 ± 0.66
$Z^0 \to q\bar{q}$ (hadrons)	1744.4 ± 2.0	69.91 ± 0.06
$Z^0 \to$ anything	2495.2 ± 2.3	100

3.10.6 Nuclear Decay Chains and the Hypo-exponential Distribution

In nuclear decay chains an unstable nucleus decays with characteristic time τ_1 to a lighter nucleus, which in turn decays with characteristic time τ_2 to an even lighter nucleus, and so on. Such decay chains with consecutive emissions (mostly α particles or electrons) are typical of heavy nuclei. Figure 3.13 (left) shows a segment of the uranium decay chain where each subsequent isotope has a single decay mode, but with a different characteristic time. Find the probability distribution to describe such processes!

✎ Suppose that the decay chain is initiated by type 1 isotopes with no daughter nuclei present at time zero, and that no other isotope decays into this type. The time evolution of the decay chain is then governed by the set of differential equations

$$\begin{aligned}
\dot{N}_1 &= -\lambda_1 N_1, \\
\dot{N}_2 &= -\lambda_2 N_2 + \lambda_1 N_1, \\
\dot{N}_3 &= -\lambda_3 N_3 + \lambda_2 N_2, \\
\cdots &= \quad \cdots ,
\end{aligned}$$

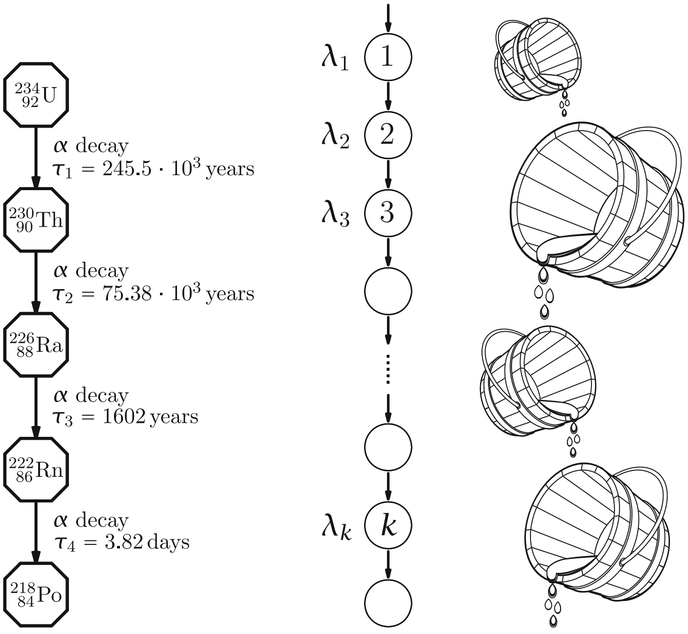


Fig. 3.13 [Left] A segment of the uranium decay chain where only one type of decay is allowed at each stage. [Center] Depiction of k serial processes with a single output, described by the hypo-exponential distribution. [Right] Illustration of a nuclear decay chain; compare it to Fig. 3.12 (*right*)

with initial conditions

$$N_1(0) = N_0, \qquad N_2(0) = N_3(0) = \cdots = 0.$$

We already know the solution of the first line:

$$N_1(t) = N_0 \, \mathrm{e}^{-\lambda_1 t}.$$

The next component of the chain is obtained by first multiplying the second line of the system by $\mathrm{e}^{\lambda_2 t}$ and exploiting the previously calculated solution for $N_1(t)$,

$$\mathrm{e}^{\lambda_2 t} \dot{N}_2(t) = -\lambda_2 \, \mathrm{e}^{\lambda_2 t} N_2 + \lambda_1 N_0 \, \mathrm{e}^{(\lambda_2 - \lambda_1)t}.$$

We move the first term on the right to the left,

$$\mathrm{e}^{\lambda_2 t} \dot{N}_2(t) + \lambda_2 \, \mathrm{e}^{\lambda_2 t} N_2 = \left(\mathrm{e}^{\lambda_2 t} N_2(t)\right)^{\cdot} = \lambda_1 N_0 \, \mathrm{e}^{(\lambda_2 - \lambda_1)t},$$

and integrate to get

$$\mathrm{e}^{\lambda_2 t} N_2(t) = \frac{\lambda_1}{\lambda_2 - \lambda_1} N_0 \, \mathrm{e}^{(\lambda_2 - \lambda_1)t} + C.$$

The constant C is dictated by the condition $N_2(0) = 0$, whence $C = -\lambda_1 N_0 / (\lambda_2 - \lambda_1)$ and

$$N_2(t) = \frac{\lambda_1}{\lambda_1 - \lambda_2} N_0 \left[\mathrm{e}^{-\lambda_2 t} - \mathrm{e}^{-\lambda_1 t}\right].$$

The same trick can be used to obtain the remaining elements of the chain: in the ith line of the system we always multiply $\dot{N}_i$ by $\mathrm{e}^{\lambda_i t}$, carry over $-\lambda_i \, \mathrm{e}^{\lambda_i t} N_i$ to the left where it can be joined with its neighbor into a derivative of a product, grab the result from the previous step, and integrate. For the third element of the chain, for example, we obtain

$$N_3(t) = \lambda_1 \lambda_2 N_0 \left[\frac{\mathrm{e}^{-\lambda_1 t}}{(\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1)} + \frac{\mathrm{e}^{-\lambda_2 t}}{(\lambda_1 - \lambda_2)(\lambda_3 - \lambda_2)} + \frac{\mathrm{e}^{-\lambda_3 t}}{(\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3)}\right].$$

It is obvious that this can be generalized to

$$N_k(t) = \sum_{i=1}^{k} \left(\prod_{\substack{j=1 \\ j \neq i}}^{k} \frac{\lambda_j}{\lambda_j - \lambda_i} \right) \lambda_i \, \mathrm{e}^{-\lambda_i t}, \tag{3.28}$$

except that we must replace $\lambda_k \to N_0$ in the numerator of all fractions. Such a distribution, which in general describes a sum of independent, exponentially distributed variables, each with its own parameter λ_i, is called *hypo-exponential*.

References

1. D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, 5th edn (John Wiley & Sons, New York, 2010)
2. G.H. Jowett, The exponential distribution and its applications. Inco. Stat. **8**, 89 (1958)
3. P.J. Aston, Is radioactive decay really exponential?, Europhys. Lett. 97 (2012) 52001. See also the reply C. A. Nicolaides, Comment on Is radioactive decay really exponential?. Europhys. Lett. **101**, 42001 (2013)
4. R.G. Newton, *Scattering Theory of Waves and Particles*, 2nd edn (Springer-Verlag, Berlin, 1982)
5. E. Merzbacher, *Quantum Mechanics*, 3rd edn (Wiley & Sons Inc, New York, 1998)
6. M.L. Goldberger, K.M. Watson, *Collision Theory*, (John Wiley & Sons, New York 1964) (Chapter 8)
7. H. Poincaré, *Calcul des Probabilités* (Gauthier-Villars, Paris, 1912)
8. WHO Global InfoBase team, *The SuRF Report 2*. Country-level data and comparable estimates. (World Health Organization, Geneva, Surveillance of chronic disease Risk Factors, 2005)
9. M.E.J. Newman, Power laws, Pareto distributions and Zipf's law. Contemp. Phys. **46**, 323 (2005)
10. A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data. SIAM Rev. **51**, 661 (2009)
11. M. Schroeder, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* (W.H. Freeman, New York, 1991)
12. J.W. Head, C.I. Fassett, S.J. Kadish, D.E. Smith, M.T. Zuber, G.A. Neumann, E. Mazarico, Global distribution of large lunar craters: implications for resurfacing and impactor populations. Science **329**, 1504 (2010)
13. S.J. Kadish, C.I. Fassett, J.W. Head, D.E. Smith, M.T. Zuber, G.A. Neumann, E. Mazarico, A Global Catalog of Large Lunar Crater ($\geq$ 20 km) from the Lunar Orbiter Laser Altimeter. Lunar Plan. Sci. Conf., XLII, abstract 1006 (2011)
14. B.R. Dennis, L.E. Orwig, G.S. Kennard, G.J. Labow, R.A. Schwartz, A.R. Shaver, A.K. Tolbert, *The Complete Hard X-ray Burst Spectrometer Event List, 1980–1989*, NASA Technical Memorandum 4332 (NASA, 1991)
15. S. Bukvić, S. Djeniže, A. Srećković, Line broadening in the Si I, Si II, Si III, and Si IV spectra in the helium plasma, Astron. Astrophys. **508**, 491 (2009)
16. Student [W.S. Gosset], The probable error of a mean. Biometrika **6**, 1 (1908)
17. R. Jamnik, *Verjetnostni račun* (Mladinska knjiga, Ljubljana, 1971)

Chapter 4
Expected Values

Abstract Finding expected values of distributions is one of the main tasks of any probabilistic analysis. The expected value in the narrower sense of the average (mean), which is a measure of distribution location, is introduced first, followed by the related concepts of the median and distribution quantiles. Expected values of functions of random variables are presented, as well as the variance as the primary measure of the distribution scale. The discussion is extended to moments of distributions (skewness, kurtosis), as well as to two- and d-dimensional generalizations. Finally, propagation of errors is analyzed.

In this chapter we discuss quantities that one may anticipate for individual random variables or their functions—with respect to the probability distributions of these variables—after multiple repetitions of random experiments: they are known as *expected values* or *expectations* of random variables. The most important such quantity is the *average value*, which is the expected value in the basic, narrowest sense of the word; further below we also discuss other expected values in the broader sense.

4.1 Expected (Average, Mean) Value

The expected value of a *discrete* random variable X, which can assume the values x_i $(i = 1, 2, \ldots)$, is computed by weighting (multiplying) each of these values by the probability $P(X = x_i) = f_X(x_i)$ that in a large number of trials this particular value turns up (see (2.13)), then sum all such products:

$$\overline{X} = E[X] = \sum_{i=1}^{n} x_i \, P(X = x_i). \tag{4.1}$$

The average is denoted by E or by a line across the random variable (or its function) being averaged. Both $E[X]$ and $\overline{X}$, as well as the frequently used symbol μ_X imply the "averaging operation" performed on the variable X. (We emphasize this because

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_4

we occasionally also use the slightly misleading expression “expected value of a distribution”: what usually changes in random processes is the value of a variable, not its distribution!) In Chaps. 4–6 the symbols

$$E[X], \quad \overline{X}, \quad \mu_X, \tag{4.2}$$

signify the one and the same thing, while in Chaps. 7–10 the symbols $\overline{X}$ and $\overline{x}$ will denote the average value of a *sample* and $E[\bullet]$ will be used strictly as expected value. The only symbol that really would not make any sense, is $E[x]$.

It can not hurt to recall the formula to compute the center of mass of a one-dimensional system of point-like masses with a total mass $M = \sum_{i=1}^{n} m_i$:

$$x_{\mathrm{cm}} = \frac{\sum_{i=1}^{n} x_i m_i}{\sum_{i=1}^{n} m_i} = \sum_{i=1}^{n} x_i \frac{m_i}{M}.$$

If all probabilities in (4.1) are equal, we get a simple expression for the usual arithmetic average

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

The expected value of a *continuous* random variable X is obtained by replacing the sum by the integral and integrating the product of the variable value x and the corresponding probability density over the whole definition domain,

$$\overline{X} = E[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x. \tag{4.3}$$

(Beware: this expected value may not exist for certain types of densities f_X.) The analogy from mechanics is again the center of mass of a three-dimensional inhomogeneous body, which is calculated by integrating the product of the position vector with the position-dependent density over the whole volume:

$$\boldsymbol{r}_{\mathrm{cm}} = \overline{\boldsymbol{r}} = \frac{1}{m} \int_V \boldsymbol{r}\,\mathrm{d}m = \frac{1}{m} \int_V \boldsymbol{r} \rho(\boldsymbol{r})\,\mathrm{d}^3\boldsymbol{r}.$$

Example In a casino we indulge in a game of dice with the following rules for each throw: 2 spots—win 10 €; 4 spots—win 30 €; 6 spots—lose 20 €; 1 spot, 3 spots or 5 spots—neither win nor lose. Any number of spots x_i is equally probable, $P(X = x_i) = \frac{1}{6}$, so the expected value of our earnings is

$$E[X] = \tfrac{1}{6}0\,€ + \tfrac{1}{6}10\,€ + \tfrac{1}{6}0\,€ + \tfrac{1}{6}30\,€ + \tfrac{1}{6}0\,€ + \tfrac{1}{6}(-20)\,€ \approx 3.67\,€.$$

If the casino wishes to profit from this game, the participation fee should be at least this much. ◁

4.2 Median

The median of a random variable X (discrete or continuous) is the value $x = \mathrm{med}[X]$, for which

$$P(X < x) \le \tfrac{1}{2} \quad \text{and} \quad P(X > x) \le \tfrac{1}{2}. \tag{4.4}$$

For a continuous variable X the inequalities become equalities,

$$P(X < x) = P(X > x) = \tfrac{1}{2} \quad \Longleftrightarrow \quad \mathrm{med}[X] = F_X^{-1}(1/2),$$

as it is always possible to find the value of x that splits the area under the probability density curve in two halves: the probabilities that X assumes a value above or below the median, respectively, are exactly 50%.

The median of a discrete variable X sometimes can not be determined uniquely, since the discrete nature of its distribution may cause the inequalities in (4.4) to be fulfilled simultaneously, but for many different x. For example, consider a discrete distribution with probability function $f_X(x) = 1/2^x$, where $x = 1, 2, \ldots$ We see that $P(X < x) = P(X > x) = \frac{1}{2}$ holds for *any* value $1 \le x \le 2$. In such cases the median is defined as the central point of the interval on which the assignment is ambiguous—in the present example we therefore set it to $\mathrm{med}[X] = 1.5$.

Example A continuous random variable has the probability density

$$f_X(x) = \begin{cases} \dfrac{4x(9-x^2)}{81} & ; \quad 0 \le x \le 3, \\ 0 & ; \quad \text{elsewhere}, \end{cases} \tag{4.5}$$

shown in Fig. 4.1 (left). Find the mode (location of maximum), median and the average (mean) of this distribution!

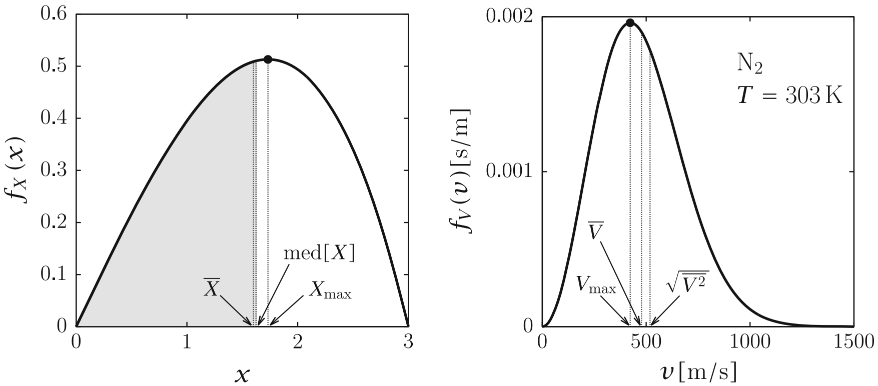


Fig. 4.1 [Left] Probability density f_X (see (4.5)) with its average, median and mode (maximum). [Right] Maxwell distribution with its mode ("most probable velocity"), average velocity and the root-mean-square velocity. See also Fig. 3.4 (left)

The mode is obtained by differentiating and setting the result to zero:

$$\left.\frac{\mathrm{d}f_X}{\mathrm{d}x}\right|_{X_{\max}} = \frac{36 - 12X_{\max}^2}{81} = 0 \quad \Longrightarrow \quad X_{\max} = \sqrt{3} \approx 1.73.$$

The median med$[X] \equiv a$ must split the area under the curve of f_X in two parts of $1/2$ each, thus

$$P\left(X < a\right) = P\left(X > a\right) = \frac{4}{81}\int_0^a x\left(9 - x^2\right)\mathrm{d}x = \frac{4}{81}\left(\frac{9a^2}{2} - \frac{a^4}{4}\right) \equiv \frac{1}{2}.$$

This results in the quadratic equation $2a^4 - 36a^2 + 81 = 0$ with two solutions, $a^2 = 9(1 \pm \sqrt{2}/2)$. Only the solution with the negative sign is acceptable as it is the only one that falls within the [0, 3] domain:

$$\mathrm{med}[X] = \sqrt{a^2} = \sqrt{9(1 - \sqrt{2}/2)} \approx 1.62.$$

The average is calculated by using the definition (4.3),

$$\overline{X} = \int_0^3 x f_X(x)\,\mathrm{d}x = \frac{4}{81}\int_0^3 x^2\left(9 - x^2\right)\mathrm{d}x = \frac{4}{81}\left.\left(3x^3 - \frac{x^5}{5}\right)\right|_0^3 \approx 1.60.$$

All three values are shown in Fig. 4.1 (left). ◁

4.3 Quantiles

The value of a random variable, below which a certain fraction of all events are found after numerous trials, is called the *quantile* of its distribution (lat. *quantum*, “how much”). For a continuous probability distribution this means that the integral of the probability density from $-\infty$ to x_α equals α (Fig. 4.2). For example, the 0.50th quantile of the standardized normal distribution is $x_{0.50} = 0$, while its 0.9985th quantile is $x_{0.9985} \approx 3$, see (3.13).

To express the αth quantile all values $0 \leq \alpha \leq 1$ are allowed, but several brethren terms are in wide use for specific values of α: integer values (in percent) express *percentiles*, the tenths of the whole range of α are delimited by *deciles* and the fourths by *quartiles:* $x_{0.20}$ defines the 20th percentile or the second decile of a distribution, $x_{0.25}$ and $x_{0.75}$ set the limits of its first and third quartile. Hence, $x_{0.50}$ carries no less than five names: it is the 0.50th quantile, the 50th percentile, the second quartile, the fifth decile and—the median. The difference $x_{0.75} - x_{0.25}$ is called the *inter-quartile range* (IQR). The interval $[x_{0.25}, x_{0.75}]$ contains half of all values; a quarter of them reside to its left and a quarter to its right.

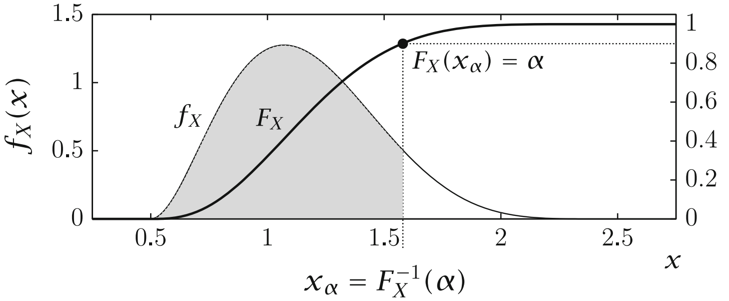


Fig. 4.2 Definition of the quantile of a continuous distribution. The integral of the density $f_X(x)$ from $-\infty$ (or the lowest edge of its domain) to $x = x_\alpha$ equals α. The figure shows the density $f_X(x) = \frac{21}{32}\left(x - \frac{1}{2}\right)^2\left(\frac{5}{2} - x\right)^5$, $0.5 \le x \le 2.5$, the corresponding distribution function, and the 90th percentile ($\alpha = 0.90$), which is $x_\alpha = 1.58$

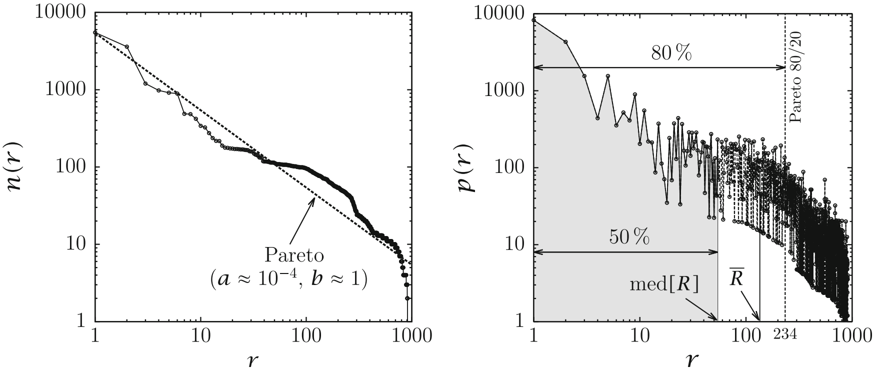


Fig. 4.3 [Left] Daily sales of fiction books as a function of sales rank. [Right] Daily earnings as a function of sales rank. In the book segment the online giant earns 50% by selling books with sales ranks above med[R] ≈ 53, while the average sales rank is $\overline{r} \approx 135$

Example Fig. 4.3 (left) shows the daily sales of fiction books from the 1000 bestseller list (sales rank r) of the AMAZON online bookstore in a certain time period. (Note the log-log scale: in linear scale the distribution has a sharp peak at $r = 1$ and a rapidly dropping tail, so it mostly occupies the region around the origin.)

To study the sales dynamics such discrete distributions are often approximated by continuous Pareto distributions (3.16). For many markets in the past, the "Pareto 80/20 principle" seemed to apply, stating that a relatively small fraction (≈20%) of products (in our case best-selling books) brings the most (≈80%) profit. Figure 4.3 (right) shows the daily earnings as a function of sales rank, as well as the median, average rank, and the sales rank up to which AMAZON earns 80% of the money: the latter is 234 (of 1000), neatly corresponding with the Pareto "principle". Still, it is obvious from the graph that the Pareto distribution under-estimates the actual sales at high ranks r. Analyses show [1, 2] that the distribution $n(r)$ has become flatter over the years, meaning that more and more profit is being squeezed from the ever increasing tail; see also [3]. ◁

4.4 Expected Values of Functions of Random Variables

The simplest functions of random variables are the sum $X + Y$ of two variables and the linear combination $aX + b$, where a and b are arbitrary real constants. Since the expected value of a continuous random variable, $E[X]$, is defined by an integral, the expected values of $E[X + Y]$ and $E[aX + b]$ inherit all properties of the integral, in particular linearity. (A similar conclusion follows in the discrete case where we are dealing with sums.) Therefore, for both continuous and discrete random variables it holds that

$$E[X + Y] = E[X] + E[Y], \tag{4.6}$$

as well as

$$E[X_1 + X_2 + \cdots + X_n] = \sum_{i=1}^{n} E[X_i]$$

and

$$E[aX + b] = aE[X] + b.$$

One needs to be slightly more careful in computing the expected values of more general functions of random variables. Suppose that X is a discrete random variable with probability distribution (probability function) f_X. Then $Y = g(X)$ is also a random variable and its probability function is

$$f_Y(y) = P(Y = y) = \sum_{\{x|g(x)=y\}} P(X = x) = \sum_{\{x|g(x)=y\}} f_X(x).$$

If X takes the values $x_1, x_2, \ldots, x_n$ and Y takes the values $y_1, y_2, \ldots, y_m$ $(m \le n)$, we have

$$\begin{aligned} E[Y] &= y_1 f_Y(y_1) + y_2 f_Y(y_2) + \cdots + y_m f_Y(y_m) \\ &= g(x_1)f_X(x_1) + g(x_2)f_X(x_2) + \cdots + g(x_n)f_X(x_n) = E\big[g(X)\big], \end{aligned}$$

hence

$$\overline{g(X)} = E\big[g(X)\big] = \sum_{i=1}^{n} g(x_i)f_X(x_i). \tag{4.7}$$

If X is a continuous random variable, we just need to replace the sum by the integral and the probability function by the probability density:

$$\overline{g(X)} = E\big[g(X)\big] = \int_{-\infty}^{\infty} g(x)f_X(x)\,\mathrm{d}x. \tag{4.8}$$

This is a good spot to comment on a very popular approximation that can be an ugly mistake or a good short-cut to a solution: it is the approximation

$$g\left(\overline{X}\right) \approx \overline{g(X)}. \tag{4.9}$$

The trick works well if the density f_X of X is a sharp, strongly peaked function, and not so well otherwise. Regardless of this, however, for any convex[1] function g, *Jensen's inequality* holds true:

$$g\left(\overline{X}\right) \le \overline{g(X)}, \tag{4.10}$$

that is,

$$g\left(\int x f_X(x)\,\mathrm{d}x\right) \le \int g(x) f_X(x)\,\mathrm{d}x.$$

4.4.1 Probability Densities in Quantum Mechanics

As physicists, we ceaselessly calculate expected values of the form (4.8) in any field related to statistical or quantum mechanics. We say: the expected value of an operator $\widehat{\mathcal{O}}$ in a certain state of a quantum-mechanical system (for example, ground state of the hydrogen atom) described by the wave-function ψ, is

$$\overline{\mathcal{O}} = \int_\Omega \psi^*(\boldsymbol{r})\widehat{\mathcal{O}}(\boldsymbol{r})\psi(\boldsymbol{r})\,\mathrm{d}V.$$

The operator $\widehat{\mathcal{O}}$ acts on the right part of the integrand, ψ, then the result is multiplied from the left by its complex conjugate ψ^*, and integrated over the whole domain. If $\widehat{\mathcal{O}}$ is multiplicative, for example $\widehat{\mathcal{O}}(\boldsymbol{r}) = z$—in this case we obtain the expectation value of the third Cartesian component of the electron's position vector in the hydrogen atom—we are computing just

$$\overline{\mathcal{O}} = \int_\Omega \widehat{\mathcal{O}}(\boldsymbol{r})\,\underbrace{|\psi(\boldsymbol{r})|^2}_{\rho(\boldsymbol{r})}\,\mathrm{d}V, \tag{4.11}$$

which is the integral of a product of two scalar functions, the second of which, $\rho(\boldsymbol{r})$, is nothing but the probability density of (4.8).

Example An electron moving in the electric field of a lead nucleus is described by the function

$$\psi(r) = \frac{1}{\sqrt{\pi}} r_\mathrm{B}^{-3/2} \mathrm{e}^{-r/r_\mathrm{B}},$$

where $r_\mathrm{B} \approx 6.46 \times 10^{-13}$ m. The nucleus may be imagined as a positively charged sphere with radius 7×10^{-15} m. How much time does the electron "spend" *in the*

[1] A function is defined to be convex if the line segment between any two points on the graph of the function lies above the graph.

nucleus, i. e. what is the probability that it resides within a sphere of radius R? All we are looking for is the expected value of the operator $\widehat{\mathcal{O}}(\boldsymbol{r}) = 1$ in (4.11); due to angular symmetry the volume element is simply $\mathrm{d}V = 4\pi r^2\,\mathrm{d}r$, thus

$$P = \int_0^R |\psi(r)|^2\, 4\pi r^2\,\mathrm{d}r \approx 1.67 \times 10^{-6}.$$

An almost identical result is obtained by assuming that ψ is practically constant on the interval $[0, R]$, which is reasonable, since $R \ll r_\mathrm{B}$. In this case we obtain $P = (1/\pi)r_\mathrm{B}^{-3}(4\pi R^3/3) = (4/3)(R/r_\mathrm{B})^3 \approx 1.69 \times 10^{-6}$. ◁

4.5 Variance and Effective Deviation

Computing the expected value of a random variable X tells us something about where within its domain its values will approximately land after many repetitions of the corresponding random experiment. Now we are also interested in the variation (scattering) of the values around their average $E[X] = \overline{X}$. A measure of this scattering is the *variance*, defined as

$$\mathrm{var}[X] = E\big[(X - E[X])^2\big] = \overline{(X - \overline{X})^2}.$$

A large variance means a large scatter around the average and vice-versa. The positive square root of the variance,

$$\sigma_X = \sqrt{\mathrm{var}[X]},$$

is known as *effective* or *standard deviation*—in particular with the normal distribution on our minds. In the following we shall also make use of the relation

$$\mathrm{var}[aX + b] = a^2\,\mathrm{var}[X]. \tag{4.12}$$

(Prove it as an exercise.) If X is a discrete random variable, which takes the values $x_1, x_2, \ldots, x_n$ and has the probability function f_X, its variance is

$$\sigma_X^2 = \sum_{i=1}^{n} \big(x_i - \overline{X}\big)^2 f_X(x_i). \tag{4.13}$$

In the case that all probabilities are equal, $f_X(x_i) = 1/n$, the variance is

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n} \big(x_i - \overline{X}\big)^2. \tag{4.14}$$

Note the factor $1/n$—not $1/(n-1)$, as one often encounters—as it will acquire an important role in random *samples* in Chap. 7.

If X is a continuous random variable with the probability density f_X, its variance is

$$\sigma_X^2 = \int_{-\infty}^{\infty} \left(x - \overline{X}\right)^2 f_X(x)\,\mathrm{d}x. \tag{4.15}$$

It can be shown that, *regardless of the distribution* obeyed by any (continuous or discrete) random variable X, it holds that

$$P\left(|X - \overline{X}| \geq a\right) \leq \frac{\sigma_X^2}{a^2}$$

for any constant $a > 0$, which is known as the Chebyshev inequality. It can also be formulated in terms of the slightly tighter Cantelli's constraints

$$P\left(X \geq \overline{X} + a\right) \leq \frac{\sigma_X^2}{\sigma_X^2 + a^2}, \qquad P\left(X \leq \overline{X} - a\right) \leq \frac{\sigma_X^2}{\sigma_X^2 + a^2}. \tag{4.16}$$

We may resort to this tool if we know only the expected value of the random variable, $\overline{X}$, and its variance, σ_X^2, but not the functional form of its distribution. In such cases we can still calculate the upper limits for probabilities of the form (4.16).

Example Suppose that the measured noise voltage at the output of a circuit has an average of $\overline{U} = 200\,\mathrm{mV}$ and variance $\sigma_U^2 = (80\,\mathrm{mV})^2$. The probability that the noise exceeds 300 mV (i. e. raises more than $\Delta U = 100\,\mathrm{mV}$ above the average), can be bounded from above as $P\left(U \geq \overline{U} + \Delta U\right) \leq \sigma_U^2 \big/ \left(\sigma_U^2 + (\Delta U)^2\right) \approx 0.39$. ◁

4.6 Complex Random Variables

A particular linear combination of real random variables X and Y is the *complex random variable*

$$Z = X + \mathrm{i}\,Y.$$

Its distribution function at $z = x + \mathrm{i}\,y$ is defined as

$$F_Z(z) = P(X \leq x, Y \leq y) = F_{X,Y}(x, y),$$

where $F_{X,Y}(x, y)$ is the distribution function of the pair—more precisely, the *random vector* (X, Y). The expected value of the variable Z is defined as

$$E[Z] = E[X] + \mathrm{i}\,E[Y].$$

Computing the expectation values of complex random variables is an additive and homogeneous operation: for arbitrary Z_1 and Z_2 it holds that

$$E[Z_1 + Z_2] = E[Z_1] + E[Z_2],$$

while for an arbitrary complex constant $c = a + \mathrm{i}\, b$ we have

$$E[cZ] = cE[Z].$$

The variance of a complex random variable is defined as

$$\mathrm{var}[Z] = E\Big[\big|Z - E[Z]\big|^2\Big].$$

A short calculation—do it!—shows that it is equal to the sum of the variances of its components,

$$\mathrm{var}[Z] = \mathrm{var}[X] + \mathrm{var}[Y].$$

The complex random variables $Z_1 = X_1 + \mathrm{i}\, Y_1$ and $Z_2 = X_2 + \mathrm{i}\, Y_2$ are mutually independent if random vectors[2] (X_1, X_2) and (Y_1, Y_2) are independent. (A generalization is at hand: complex random variables $Z_k = X_k + \mathrm{i}\, Y_k$ $(k = 1, 2, \ldots, n)$ are mutually independent if the same applies to random vectors (X_k, Y_k).) If Z_1 and Z_2 are independent and possess expected values, their product also possesses it, and it holds that

$$E[Z_1 Z_2] = E[Z_1]E[Z_2].$$

4.7 Moments

The average (mean) and the variance are two special cases of expected values in the broader sense called *moments:* the *pth raw* or *algebraic moment* M'_p of a random variable X is defined as the expected value of its pth power, that is, $M'_p = E[X^p]$:

$$\begin{aligned} M'_p &= \sum_{i=1}^{n} x_i^p f_X(x_i) \quad \text{(discrete case)}, \\ M'_p &= \int_{-\infty}^{\infty} x^p f_X(x)\, \mathrm{d}x \quad \text{(continuous case)}. \end{aligned} \tag{4.17}$$

[2] A random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ with a distribution function $F_{\boldsymbol{X}}(x_1, x_2, \ldots, x_m)$ and a random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ with a distribution function $F_{\boldsymbol{Y}}(y_1, y_2, \ldots, y_n)$ are mutually independent if $F_{\boldsymbol{X},\boldsymbol{Y}}(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n) = F_{\boldsymbol{X}}(x_1, x_2, \ldots, x_m) F_{\boldsymbol{Y}}(y_1, y_2, \ldots, y_n)$. This is an obvious generalization of (2.20) and (2.24).

Frequently we also require *central moments*, defined with respect to the corresponding average value of the variable, that is, $M_p = E\big[\big(X - \overline{X}\big)^p\big]$:

$$M_p = \sum_{i=1}^{n} \big(x_i - \overline{X}\big)^p f_X(x_i) \quad \text{(discrete case)},$$

$$M_p = \int_{-\infty}^{\infty} \big(x - \overline{X}\big)^p f_X(x)\,\mathrm{d}x \quad \text{(continuous case)}.$$

From here we read off $M'_0 = 1$ (normalization of probability distribution), $M'_1 = \overline{X}$ and $M_2 = \sigma_X^2$. The following relations (check them as an exercise) also hold:

$$M_2 = M'_2 - \overline{X}^2 = \overline{X^2} - \overline{X}^2,$$
$$M_3 = M'_3 - 3M'_2\overline{X} + 2\overline{X}^3,$$
$$M_4 = M'_4 - 4M'_3\overline{X} + 6M'_2\overline{X}^2 - 3\overline{X}^4.$$

In addition to the first (average) and second moment (variance) only the third and fourth central moment are in everyday use. The third central moment, divided by the third power of its effective deviation,

$$\rho = \frac{M_3}{\sigma^3}, \tag{4.18}$$

is called the *coefficient of skewness* or simply *skewness*. The coefficient ρ measures the asymmetry of the distribution around its average: $\rho < 0$ means that the distribution has a relatively longer tail to the left of the average value (Fig. 4.4 (left)), while $\rho > 0$ implies a more pronounced tail to its right (Fig. 4.4 (center)).

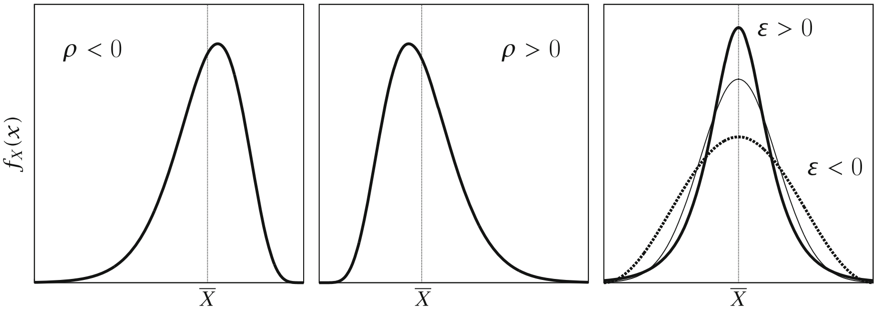


Fig. 4.4 [Left] A distribution with negative skewness: the tail protruding to the left of the average value is more pronounced than the one sticking to the right. [Center] A distribution with positive skewness. [Right] Examples of distributions with positive (*thick full curve*) and negative excess kurtosis (*thick dashed curve*) with respect to the normal distribution (*thin full curve*)

The fourth central moment, divided by the square of the variance,

$$\frac{M_4}{\sigma^4},$$

is known as *kurtosis* and tells us something about the "sharpness" or "bluntness" of the distribution. For the normal distribution we have $M_4/\sigma^4 = 3$, so we sometimes prefer to specify the quantity

$$\varepsilon = \frac{M_4}{\sigma^4} - 3, \tag{4.19}$$

called the *excess kurtosis:* $\varepsilon > 0$ indicates that the distribution is "sharper" than the normal (more prominent peak, faster falling tails), while $\varepsilon < 0$ implies a "blunter" distribution (less pronounced peak, stronger tails), see Fig. 4.4 (right).

The properties of the most important continuous distributions—average value, median, mode (location of maximum), variance, skewness (ρ) and kurtosis ($\varepsilon + 3$)—are listed in Table 4.1. See also Appendices B.2 and B.3, where we shall learn how to "automate" the calculation of moments by using generating and characteristic functions.

Table 4.1 Properties of select continuous distributions: average (mean) value, median, mode, variance, skewness ($M_3/\sigma^3 = \rho$) and kurtosis ($M_4/\sigma^4 = \varepsilon + 3$)

Distribution	Average	Median	Mode	Variance	ρ	$\varepsilon + 3$
$U(a, b)$ (3.1)	$\frac{a+b}{2}$	$\frac{a+b}{2}$	/	$\frac{(b-a)^2}{12}$	0	$\frac{9}{5}$
$\mathrm{Exp}(\lambda)$ (3.4)	$\frac{1}{\lambda}$	$\frac{\log 2}{\lambda}$	0	$\frac{1}{\lambda^2}$	2	9
$N(\mu, \sigma^2)$ (3.7)	μ	μ	μ	σ^2	0	3
Cauchy (3.20)	/	x_0	x_0	/	/	/
$\chi^2(\nu)$ (3.21)	ν	$\nu - \frac{2}{3}^{\dagger}$	$\nu - 2^{\ddagger}$	2ν	$\frac{2^{3/2}}{\sqrt{\nu}}$	$3 + \frac{12}{\nu}$
$t(\nu)$ (3.22)	$0^{\star}$	0	0	$\frac{\nu}{\nu-2}^{*}$	$0^{\P}$	$\frac{3(\nu-2)}{\nu-4}^{\S}$
$\mathrm{Pareto}(a, b)$ (3.16)	$\frac{ab}{a-1}^{\$}$	$b\sqrt[a]{2}$	b	$\frac{b^2/(a-2)}{(a-1)^2}^{\maltese}$	$\odot$	$\oplus$

Notes † approximate dependence for large ν | ‡ for $\nu > 2$ | * undefined for $\nu = 1$
* undefined for $\nu \le 2$ | ¶ undefined for $\nu \le 3$ | § undefined for $\nu \le 4$
$^{\$}$ defined for $a > 1$, otherwise ∞ | $^{\maltese}$ defined for $a > 2$, while $\alpha \in (1, 2]$ for ∞
$^{\odot}$ $[2(a+1)/(a-3)]\sqrt{(a-2)/a}$ for $a > 3$
$^{\oplus}$ $3 + 6(a^3 + a^2 - 6a - 2)/(a(a-3)(a-4))$ for $a > 4$

Example We are interested in the mode ("most probable velocity"), average velocity and the average velocity squared of N_2 gas molecules (molar mass $M = 28\,\mathrm{kg/kmol}$, mass of single molecule $m = M/N_\mathrm{A}$) at temperature $T = 303\,\mathrm{K}$. The velocity distribution of the molecules is given by the Maxwell distribution (3.15), whose maximum (mode) is determined by $\mathrm{d}f_V/\mathrm{d}v = 0$, hence

$$\left(2v - v^2 \frac{m}{2k_\mathrm{B}T}\, 2v\right)\bigg|_{V_\mathrm{max}} = 0 \quad\Longrightarrow\quad V_\mathrm{max} = \sqrt{\frac{2k_\mathrm{B}T}{m}} \approx 423\,\mathrm{m/s}.$$

The average value and the square root of the average velocity squared ("root-mean-square velocity") are computed from (4.3) and (4.17) with $p = 2$:

$$\overline{V} = \int_0^\infty v f_V(v)\,\mathrm{d}v = \sqrt{\frac{8k_\mathrm{B}T}{\pi m}} = \sqrt{\frac{4}{\pi}}\, V_\mathrm{max} \approx 478\,\mathrm{m/s},$$
$$\sqrt{\overline{V^2}} = \left(\int_0^\infty v^2 f_V(v)\,\mathrm{d}v\right)^{1/2} = \sqrt{\frac{3k_\mathrm{B}T}{m}} = \sqrt{\frac{3}{2}}\, V_\mathrm{max} \approx 518\,\mathrm{m/s},$$

where we have used $\int_0^\infty z^3 \exp(-z^2)\,\mathrm{d}z = 1/2$ and $\int_0^\infty z^4 \exp(-z^2)\,\mathrm{d}z = 3\sqrt{\pi}/8$. These three famous quantities are shown in Fig. 4.1 (right). ◁

4.7.1 Moments of the Cauchy Distribution

The Cauchy distribution $f_X(x) = (1/\pi)/(1+x^2)$ drops off so slowly at $x \to \pm\infty$ that its moments (average, variance, and so on) do not exist. For this reason its domain is frequently restricted to a narrower interval $[-x_\mathrm{max}, x_\mathrm{max}]$:

$$g_X(x) = \frac{f_X(x)}{\int_{-x_\mathrm{max}}^{x_\mathrm{max}} f_X(x')\,\mathrm{d}x'} = \frac{1}{2\,\arctan x_\mathrm{max}} \frac{1}{1+x^2}.$$

This is particularly popular in nuclear physics where the Breit–Wigner description of the shape of the resonance peak in its tails—see Fig. 3.6 (right)—is no longer adequate due to the presence of neighboring resonances or background. With the truncated density g_X both the average and the variance are well defined:

$$E[X] = \frac{1}{2\,\arctan x_\mathrm{max}} \int_{-x_\mathrm{max}}^{x_\mathrm{max}} \frac{x}{1+x^2}\,\mathrm{d}x = 0,$$
$$\mathrm{var}[X] = \frac{1}{2\,\arctan x_\mathrm{max}} \int_{-x_\mathrm{max}}^{x_\mathrm{max}} \frac{x^2}{1+x^2}\,\mathrm{d}x = \frac{x_\mathrm{max}}{\arctan x_\mathrm{max}} - 1.$$

Narrowing down the domain is a special case of a larger class of "distortions" of probability distributions used to describe, for example, non-ideal outcomes of a process or imperfect efficiencies for analyzing particles in a detector. If individual events are detected under different conditions, the ideal probability density, f_X, must be weighted by the *detection efficiency:*

$$g_X(x) = \frac{\int_{\Omega_y} f_X(x) P(y|x) \varepsilon(x, y) \,\mathrm{d}y}{\int_{\Omega_{x'}} \int_{\Omega_y} f_X(x') P(y|x') \varepsilon(x', y) \,\mathrm{d}x' \,\mathrm{d}y},$$

where y is an auxiliary variable over which the averaging is being performed, and $\varepsilon(x, y)$ is the probability density for the event being detected near $X = x$ and $Y = y$. An introduction to such weighted averaging procedures can be found in Sect. 8.5 of [4].

4.8 Two- and *d*-dimensional Generalizations

Let the continuous random variables X and Y be distributed according to the joint probability density $f_{X,Y}(x, y)$. In this case the expected values of the individual variable can be calculated by the obvious generalization of (4.3) to two dimensions. The density $f_{X,Y}$ is weighted by the variable whose expected value we are about to compute, while the other is left untouched:

$$\overline{X} = \mu_X = E[X] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y) \,\mathrm{d}x \,\mathrm{d}y,$$
$$\overline{Y} = \mu_Y = E[Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y) \,\mathrm{d}x \,\mathrm{d}y.$$

In the discrete case the extension to two variables requires a generalization of (4.1):

$$E[X] = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i f_{X,Y}(x_i, y_j), \qquad E[Y] = \sum_{i=1}^{n} \sum_{j=1}^{m} y_j f_{X,Y}(x_i, y_j).$$

By analogy to (4.15) and (4.13) we also compute the variances of variables in the continuous case,

$$\sigma_X^2 = E\big[(X - \mu_X)^2\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)^2 f_{X,Y}(x, y) \,\mathrm{d}x \,\mathrm{d}y,$$
$$\sigma_Y^2 = E\big[(Y - \mu_Y)^2\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \mu_Y)^2 f_{X,Y}(x, y) \,\mathrm{d}x \,\mathrm{d}y,$$

and the variances in the discrete case,

$$E\big[(X - \mu_X)^2\big] = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i - \mu_X)^2 f_{X,Y}(x_i, y_j),$$

$$E\big[(Y-\mu_Y)^2\big] = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_j-\mu_Y)^2 f_{X,Y}(x_i,y_j).$$

Henceforth we only give equations pertaining to continuous variables. The corresponding expressions for discrete variables are obtained, as usual, by replacing the probability densities $f_{X,Y}(x,y)$ by the probability [mass] functions $f_{X,Y}(x_i,y_j) = P(X=x_i, Y=y_j)$, and integrals by sums.

Since now *two* variables are at hand, we can define yet a third version of the double integral (or the double sum) in which the variables enter bilinearly—the so-called mixed moment known as the *covariance* of X and Y:

$$\sigma_{XY} = \mathrm{cov}[X,Y] = E\big[(X-\mu_X)(Y-\mu_Y)\big] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

One immediately sees that

$$\mathrm{cov}[aX,bY] = ab\,\mathrm{cov}[X,Y]$$

for arbitrary constants a and b, as well as

$$\begin{aligned}\sigma_{XY} &= E\big[(X-\mu_X)(Y-\mu_Y)\big] = E\big[XY-\mu_X Y-\mu_Y X+\mu_X\mu_Y\big]\\ &= E[XY]-\mu_X\underbrace{E[Y]}_{\mu_Y}-\mu_Y\underbrace{E[X]}_{\mu_X}+\mu_X\mu_Y = E[XY]-\mu_X\mu_Y.\end{aligned}$$

Therefore, if X and Y are mutually independent, then by definition (2.25) one also has $E[XY] = E[X]E[Y] = \mu_X\mu_Y$, and then

$$\sigma_{XY} = 0.$$

(The covariance of independent variables equals zero.) For a later discussion of measurement uncertainties the following relation between the variance and covariance of two variables is important:

$$\begin{aligned}\mathrm{var}[X\pm Y] &= \iint\big((x-\mu_X)\pm(y-\mu_Y)\big)^2 f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y\\ &= \iint(x-\mu_X)^2 f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y + \iint(y-\mu_Y)^2 f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y\\ &\quad\pm 2\iint(x-\mu_X)(y-\mu_Y)f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y\\ &= \mathrm{var}[X]+\mathrm{var}[Y]\pm 2\,\mathrm{cov}[X,Y].\end{aligned}\tag{4.20}$$

In other words,

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y \pm 2\sigma_{XY}.$$

By using the covariance and both effective deviations we define the *Pearson's coefficient of linear correlation* (also *linear correlation coefficient*)

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}, \quad -1 \le \rho_{XY} \le 1. \tag{4.21}$$

It is easy to confirm the allowed range of ρ_{XY} given above. Because of its power two the expression $E[(\lambda(X-\mu_X)-(Y-\mu_Y))^2]$ is non-negative for any $\lambda \in \mathbb{R}$. Let us expand it:

$$\lambda^2 \underbrace{E\big[(X-\mu_X)^2\big]}_{\sigma^2_X} - 2\lambda \underbrace{E\big[(X-\mu_X)(Y-\mu_Y)\big]}_{\sigma_{XY}} + \underbrace{E\big[(Y-\mu_Y)^2\big]}_{\sigma^2_Y} \ge 0.$$

The left side of this inequality is a real polynomial of second degree $a\lambda^2+b\lambda+c=0$ with coefficients $a=\sigma^2_X$, $b=-2\sigma_{XY}$, $c=\sigma^2_Y$, which is non-negative everywhere, so it can have at most one real zero. This implies that its discriminant can not be positive, so $b^2-4ac\le 0$. This tells us that $4\sigma^2_{XY}-4\sigma^2_X\sigma^2_Y \le 0$ or $|\sigma_{XY}/(\sigma_X\sigma_Y)| \le 1$, which is precisely (4.21).

The generalization of (4.20) to the sum of (not necessarily independent) random variables $X_1, X_2, \ldots, X_n$ is

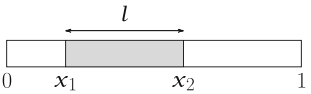
$$\mathrm{var}[X_1+X_2+\cdots+X_n] = \sum_{i=1}^n \mathrm{var}[X_i] + 2\sum_{i=1}^n \sum_{j=i+1}^n \mathrm{cov}[X_i, X_j].$$

If the variables $X_1, X_2, \ldots, X_n$ are mutually independent, this expression reduces to

$$\mathrm{var}[X_1+X_2+\cdots+X_n] = \sum_{i=1}^n \mathrm{var}[X_i]. \tag{4.22}$$

Example Many sticks with length 1 are broken at two random locations. What is the average length of the central pieces? At each hit, the stick breaks at $0 < x_1 < 1$ and $0 < x_2 < 1$, where the values x_1 and x_2 are uniformly distributed over the interval [0, 1], but one can have either $x_1 < x_2$ or $x_1 > x_2$. What we are seeking, then, is the expected value of the variable $L = |X_2 - X_1|$ (with values l) with respect to the probability density $f_{X,Y}(x_1, x_2) = 1$:

$$\overline{L} = \int_0^1\int_0^1 \big|x_2 - x_1\big|\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_0^1 \mathrm{d}x_2 \int_0^{x_2} (x_2-x_1)\,\mathrm{d}x_1 + \int_0^1 \mathrm{d}x_2 \int_{x_2}^1 (x_1-x_2)\,\mathrm{d}x_1 = \frac{1}{3}.$$



How would the result change if the probability that the stick breaks linearly increases from 0 at the origin to 1 at the opposite edge? ◁

Example Let the continuous random variables X and Y both be normally distributed, with averages μ_X and μ_Y and variances σ_X^2 and σ_Y^2. What is their joint probability density if X and Y are independent, and what are their joint and conditional densities in the dependent case, with correlation coefficient $\rho_{XY} = \rho$?

If X and Y are independent, their joint probability density—by (2.25)—is simply the product of the corresponding one-dimensional densities:

$$f_{X,Y}(x,y) = X(x)f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right] \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left[-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right].$$

The curves of constant values of $f_{X,Y}$ in the (x,y) plane are untilted ellipses in general $(\sigma_X \neq \sigma_Y)$, and circles in the special case $\sigma_X = \sigma_Y$. At any rate $\rho = 0$ for such a distribution. A two-dimensional normal distribution of dependent (and therefore correlated) variables is described by the probability density

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{1-\rho^2}\left[\frac{x'^2}{2\sigma_X^2} - 2\rho\frac{x'y'}{\sqrt{2}\sigma_X\sqrt{2}\sigma_Y} + \frac{y'^2}{2\sigma_Y^2}\right]\right\},$$

where we have denoted $x' = x - \mu_X$ and $y' = y - \mu_Y$. This distribution can not be factorized as $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, and its curves of constant values are tilted ellipses; for parameters $\mu_X = 10$, $\mu_Y = 0$, $\sigma_X = \sigma_Y = 1$ and $\rho = 0.8$ they are shown in Fig. 4.5 (left).

Conditional probability densities $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ can be computed by using (2.26) and (2.27). Let us treat the first case, the other one is obtained by simply replacing $x \leftrightarrow y$, $\mu_X \leftrightarrow \mu_Y$ and $\sigma_X \leftrightarrow \sigma_Y$ at appropriate locations:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_X^2}\left[x' - \rho\frac{\sigma_X}{\sigma_Y}y'\right]^2\right\}.$$

This conditional probability density is shown in Fig. 4.5 (right). By comparing it to definition (3.7) we infer that the random variable $X|Y$ is distributed as

$$X|Y \sim N\left(E[X] + \rho\frac{\sigma_X}{\sigma_Y}(Y - \mu_Y),\ (1-\rho^2)\sigma_X^2\right),$$

a feature also seen in the plot: the width of the band does not depend on y. ◁

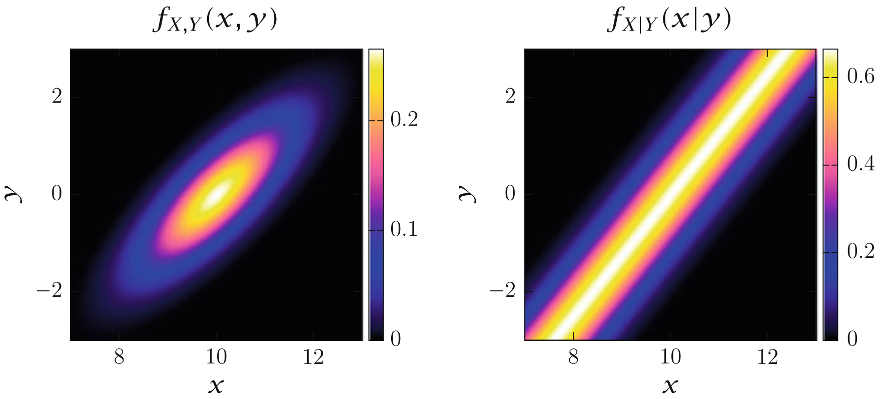


Fig. 4.5 [Left] Joint probability density of two dependent, normally distributed random variables X and Y with averages $\mu_X = 10$ and $\mu_Y = 0$, variances $\sigma_X^2 = \sigma_Y^2 = 1$ and linear correlation coefficient $\rho = 0.8$. [Right] Conditional probability density $f_{X|Y}(x|y)$

4.8.1 Multivariate Normal Distribution

This appears to be a good place to generalize the normal distribution of two variables (the so-called binormal or bivariate normal distribution) to d dimensions. We are dealing with a vector random variable

$$\boldsymbol{X} = \bigl(X_1, X_2, \ldots, X_d\bigr)^{\mathrm{T}} \in \mathbb{R}^d$$

and its average

$$E[\boldsymbol{X}] = \bigl(E[X_1], E[X_2], \ldots, E[X_d]\bigr)^{\mathrm{T}} = \bigl(\mu_1, \mu_2, \ldots, \mu_d\bigr)^{\mathrm{T}} = \boldsymbol{\mu}.$$

We construct the $d \times d$ *covariance matrix* Σ with the matrix elements

$$\Sigma_{ij} = \mathrm{cov}[X_i, X_j], \quad i, j = 1, 2, \ldots, d.$$

The covariance matrix is symmetric and at least positive semi-definite. It can even be strictly positive definite if none of the variables X_i is a linear combination of the others. The probability density of the *multivariate normal distribution* (compare it to its one-dimensional counterpart (3.10)) is then

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-d/2} \bigl(\det \Sigma\bigr)^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \, \Sigma^{-1} \, (\boldsymbol{x} - \boldsymbol{\mu}) \right\}. \qquad (4.23)$$

If $d = 2$ as in the previous Example, we have simply $\boldsymbol{X} = (X_1, X_2)^{\mathrm{T}} \to (X, Y)^{\mathrm{T}}$ and $\boldsymbol{\mu} = (\mu_1, \mu_2)^{\mathrm{T}} \to (\mu_X, \mu_Y)^{\mathrm{T}}$, while the covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}.$$

4.8.2 Correlation Does Not Imply Causality

A vanishing correlation coefficient of X and Y does not mean that these variables are stochastically independent: for each density $f_{X,Y}$ that is an even function of the deviations $x - \mu_X$ and $y - \mu_Y$, one has $\rho_{XY} = 0$. In other words, $\rho_{XY} = 0$ is just a necessary, but not sufficient condition for independence: see bottom part of Fig. 7.8 which illustrates the correlation in the case of finite samples.

Even though one observes a correlation in a pair of variables (sets of values, measurements, phenomena) this does not necessarily mean that there is a direct causal relation between them: *correlation does not imply causality*. When we observe an apparent dependence between two correlated quantities, often a third factor is involved, common to both X and Y. Example: the sales of ice-cream and the number of shark attacks at the beach are certainly correlated, but there is no causal relation between the two. (Does your purchase of three scoops of ice-cream instead of one triple your chances of being bitten by a shark?) The common factor of tempting scoops and aggressiveness of sharks is a hot summer day, when people wish to cool off in the water and sharks prefer to dwell near the shore.

Besides, one should be aware that correlation and causality are concepts originating in completely different worlds: the former is a statement on the basis of probability theory, while the latter signifies a strictly physical phenomenon, whose background is time and the causal connection between the present and past events.

4.9 Propagation of Errors

If we knew how to generalize (4.20) to an arbitrary function of an arbitrary number of variables, we would be able to answer the important question of *error propagation*. But what do we mean by “error of random variable”? In the introductory chapters we learned that each measurement of a quantity represents a single realization of a random variable whose value fluctuates statistically. Such a random deviation from its expected value is called the statistical uncertainty or “error”. By studying the propagation of errors we wish to find out how the uncertainties of a given set of variables translate into the uncertainty of a *function* of these variables. A typical example is the determination of the thermal power released on a resistor from the corresponding voltage drop: if the uncertainty of the voltage measurement is ΔU and the resistance R is known to an accuracy of no more than ΔR, what is the uncertainty of the calculated power $P = U^2/R$?

Let $X_1, X_2, \ldots, X_n$ be real random variables with expected values $\mu_1, \mu_2, \ldots, \mu_n$, which we arrange as vectors

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^{\mathrm{T}}$$

and

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^{\mathrm{T}},$$

just as in Sect. 4.8.1. Let $Y = Y(\boldsymbol{X})$ be an arbitrary function of these variables which, of course, is also a random variable. Assume that the covariances of all (X_i, X_j) pairs are known. We would like to estimate the variance of the variable Y. In the vicinity of $\boldsymbol{\mu}$ we expand Y in a Taylor series in $\boldsymbol{X}$ up to the linear term,

$$Y(\boldsymbol{X}) \approx Y(\boldsymbol{\mu}) + \sum_{i=1}^{n} (X_i - \mu_i) \frac{\partial Y}{\partial X_i}\bigg|_{\boldsymbol{X}=\boldsymbol{\mu}},$$

and resort to the approximation $E[Y(\boldsymbol{X})] \approx Y(\boldsymbol{\mu})$ (see (4.9) and (4.10)) to compute the variance. It follows that

$$\begin{aligned} \mathrm{var}[Y(\boldsymbol{X})] &= E\Big[\Big(Y(\boldsymbol{X}) - E\big[Y(\boldsymbol{X})\big]\Big)^2\Big] \approx E\Big[\Big(Y(\boldsymbol{X}) - Y(\boldsymbol{\mu})\Big)^2\Big] \\ &\approx \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\frac{\partial Y}{\partial X_i}\frac{\partial Y}{\partial X_j}\right)_{\boldsymbol{X}=\boldsymbol{\mu}} \Sigma_{ij}, \end{aligned} \tag{4.24}$$

where

$$\Sigma_{ij} = E\big[(X_i - \mu_i)(X_j - \mu_j)\big] = \mathrm{cov}\big[X_i, X_j\big]$$

is the covariance matrix of the variables X_i: its diagonal terms are the variances of the individual variables, $\mathrm{var}[X_i] = \sigma_{X_i}^2$, while the non-diagonal ones ($i \neq j$) are the covariances $\mathrm{cov}[X_i, X_j]$. Formula (4.24) is what we have been looking for: it tells us—within the specified approximations—how the "errors" in $\boldsymbol{X}$ map to the "errors" in Y. If X_i are mutually independent, we have $\mathrm{cov}[X_i, X_j] = 0$ for $i \neq j$ and the formula simplifies to

$$\mathrm{var}[Y(\boldsymbol{X})] \approx \sum_{i=1}^{n} \left(\frac{\partial Y}{\partial X_i}\right)^2_{\boldsymbol{X}=\boldsymbol{\mu}} \mathrm{var}[X_i]. \tag{4.25}$$

Example Let X_1 and X_2 be independent continuous random variables with the mean values μ_1 and μ_2 and variances σ_1^2 and σ_2^2. We are interested in the variance σ_Y^2 of their ratio $Y = X_1/X_2$. Since X_1 and X_2 are independent, we may apply formula (4.25). We need the derivatives

$$\left(\frac{\partial Y}{\partial X_1}\right)_{X=\mu} = \frac{1}{\mu_2}, \qquad \left(\frac{\partial Y}{\partial X_2}\right)_{X=\mu} = -\frac{\mu_1}{\mu_2^2}.$$

Therefore

$$\sigma_Y^2 \approx \left(\frac{1}{\mu_2}\right)^2 \sigma_1^2 + \left(\frac{\mu_1}{\mu_2^2}\right)^2 \sigma_2^2 = \frac{1}{\mu_2^4}\left[\mu_2^2\sigma_1^2 + \mu_1^2\sigma_2^2\right]$$

or

$$\frac{\sigma_Y^2}{\mu_Y^2} \approx \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2},$$

where $\mu_Y = E[Y] = \mu_1/\mu_2$. ◁

Example Let X and Y be independent random variables with the expected values μ_X and μ_Y and variances σ_X^2 and σ_Y^2 (with respective "uncertainties of measurements" σ_X and σ_Y). What is the variance σ_Z^2 of the product of their powers,

$$Z = X^m Y^n?$$

(This is a generalization of the function from the previous example to arbitrary powers m and n.) By formula (4.25) we again obtain

$$\sigma_Z^2 \approx \left(mX^{m-1}Y^n\right)^2_{\substack{X=\mu_X\\ Y=\mu_Y}} \sigma_X^2 + \left(nX^mY^{n-1}\right)^2_{\substack{X=\mu_X\\ Y=\mu_Y}} \sigma_Y^2.$$

Thus

$$\left(\frac{\sigma_Z}{\mu_Z}\right)^2 \approx \frac{m^2\mu_X^{2(m-1)}\mu_Y^{2n}}{\mu_X^{2m}\mu_Y^{2n}}\sigma_X^2 + \frac{n^2\mu_X^{2m}\mu_Y^{2(n-1)}}{\mu_X^{2m}\mu_Y^{2n}}\sigma_Y^2 = m^2\left(\frac{\sigma_X}{\mu_X}\right)^2 + n^2\left(\frac{\sigma_Y}{\mu_Y}\right)^2,$$

where we have denoted $\mu_Z = \mu_X^m\mu_Y^n$. ◁

4.9.1 Multiple Functions and Transformation of the Covariance Matrix

Let us now discuss the case of multiple scalar functions $Y_1, Y_2, \ldots, Y_m$, which all depend on variables $\boldsymbol{X}$,

$$Y_k = Y_k(X_1, X_2, \ldots, X_n) = Y_k(\boldsymbol{X}), \quad k = 1, 2, \ldots, m.$$

We arrange the function values in the vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)^{\mathrm{T}}$ and retrace the steps from the beginning of this section. We neglect all higher order terms in the Taylor expansion

$$Y_k(\boldsymbol{X}) = Y_k(\boldsymbol{\mu}) + \sum_{i=1}^{n}(X_i - \mu_i)\frac{\partial Y_k}{\partial X_i}\bigg|_{\boldsymbol{X}=\boldsymbol{\mu}} + \cdots , \quad k = 1, 2, \ldots, m,$$

and take into account that $E[Y_k(\boldsymbol{X})] \approx Y_k(\boldsymbol{\mu})$. Instead of (4.24) we now obtain a relation between the covariance matrix of variable $\boldsymbol{X}$ and the covariance matrix of the variables $\boldsymbol{Y}$,

$$\begin{aligned}\Sigma_{kl}(\boldsymbol{Y}) &\approx E\Big[\Big(Y_k(\boldsymbol{X}) - Y_k(\boldsymbol{\mu})\Big)\Big(Y_l(\boldsymbol{X}) - Y_l(\boldsymbol{\mu})\Big)\Big] \\ &\approx \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial Y_k}{\partial X_i}\frac{\partial Y_l}{\partial X_j}\right)_{\boldsymbol{X}=\boldsymbol{\mu}} \underbrace{E\big[(X_i - \mu_i)(X_j - \mu_j)\big]}_{\Sigma_{ij}(\boldsymbol{X})}.\end{aligned}$$

This relation becomes even more transparent if we write the Taylor expansion as

$$\boldsymbol{Y}(\boldsymbol{X}) = \boldsymbol{Y}(\boldsymbol{\mu}) + D\boldsymbol{X} + \cdots ,$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are n- and m-dimensional vectors, respectively, while D is an $m \times n$ matrix embodying the linear part of the expansion, namely

$$D_{ki} = \left(\frac{\partial Y_k}{\partial X_i}\right)_{\boldsymbol{X}=\boldsymbol{\mu}}, \tag{4.26}$$

Hence

$$\Sigma_{kl}(\boldsymbol{Y}) \approx \sum_{i=1}^{n}\sum_{j=1}^{n} D_{ki}\Sigma_{ij}(\boldsymbol{X})D_{jl}, \quad k, l = 1, 2, \ldots, m,$$

or, in brief,

$$\Sigma(\boldsymbol{Y}) \approx D\Sigma(\boldsymbol{X})D^{\mathrm{T}}. \tag{4.27}$$

The propagation of errors in higher dimensions can therefore be seen as a transformation of the covariance matrix. The variances $\sigma_{Y_k}^2$ of the variables Y_k are the diagonal matrix elements of $\Sigma(\boldsymbol{Y})$. In general they pick up terms from all elements of $\Sigma(\boldsymbol{X})$, even the non-diagonal ones, since

$$\Sigma_{kk}(\boldsymbol{Y}) \approx \sum_{i=1}^{n}\left(\frac{\partial Y_k}{\partial X_i}\frac{\partial Y_k}{\partial X_j}\right)_{\boldsymbol{X}=\boldsymbol{\mu}} \Sigma_{ij}(\boldsymbol{X}). \tag{4.28}$$

But if the variables X_i are mutually independent, only diagonal elements of $\Sigma(\boldsymbol{X})$ contribute to the right-hand side of the above equation, yielding

$$\sigma_{Y_k}^2 = \sum_{i=1}^{n}\left(\frac{\partial Y_k}{\partial X_i}\right)_{\boldsymbol{X}=\boldsymbol{\mu}}^2 \sigma_{X_i}^2. \tag{4.29}$$

Equations (4.28) and (4.29) are multi-dimensional equivalents of (4.24) and (4.25). Note that the non-diagonal elements of $\Sigma(\boldsymbol{Y})$ may be non-zero even though X_i are mutually independent! You can find an example of how to use these equations in the case of a measurement of the momentum of a particle in Problem 4.10.6.

4.10 Problems

4.10.1 Expected Device Failure Time

A computer disk is controlled by five circuits ($i = 1, 2, 3, 4, 5$). The time until an irreparable failure in each circuit is exponentially distributed, with individual time constants λ_i. The disk as a whole works if circuits 1, 2 and 3, circuits 3, 4 and 5, or, obviously, all five circuits work simultaneously. What is the expected time of disk failure?

✎ The probability that the ith element is not broken until time t (the probability that the failure time is larger than t) is exponentially decreasing and equals $\mathrm{e}^{-\lambda_i t}$. For the disk to fail, three key events are responsible:

$$\begin{array}{lll} \text{event } A : \text{circuits 1 and 2 fail after time } t & : P(A) = \mathrm{e}^{-(\lambda_1+\lambda_2)t}, \\ \text{event } B : \text{circuit 3 fails after time } t & : P(B) = \mathrm{e}^{-\lambda_3 t}, \\ \text{event } C : \text{circuits 4 and 5 fail after time } t & : P(C) = \mathrm{e}^{-(\lambda_4+\lambda_5)t}. \end{array}$$

The disk operates as long as $\big(A \cap B \cap \overline{C}\big) \cup \big(\overline{A} \cap B \cap C\big) \cup \big(A \cap B \cap C\big) \neq \{\}$. The probability that the disk still operates after time t, is therefore

$$\begin{aligned} P(t) &= P\big(A \cap B \cap \overline{C}\big) + P\big(\overline{A} \cap B \cap C\big) + P\big(A \cap B \cap C\big) \\ &= P(A)P(B)\big[1 - P(C)\big] + \big[1 - P(A)\big]P(B)P(C) + P(A)P(B)P(C) \\ &= P(B)\big[P(A) + P(C) - P(A)P(C)\big] \\ &= \mathrm{e}^{-(\lambda_1+\lambda_2+\lambda_3)t} + \mathrm{e}^{-(\lambda_3+\lambda_4+\lambda_5)t} - \mathrm{e}^{-(\lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5)t}. \end{aligned}$$

This is not yet our answer, since the expression still contains time! We are looking for the *expected value* of failure time, where we should recall that the appropriate probability density is $-P'(t)$ (see (3.4)), hence

$$\begin{aligned} \overline{T} = \int_0^\infty & t\,\big[(\lambda_1+\lambda_2+\lambda_3)\mathrm{e}^{-(\lambda_1+\lambda_2+\lambda_3)t} + (\lambda_3+\lambda_4+\lambda_5)\mathrm{e}^{-(\lambda_3+\lambda_4+\lambda_5)t} \\ & -(\lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5)\mathrm{e}^{-(\lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5)t}\big]\,\mathrm{d}t \\ = \frac{1}{\lambda_1+\lambda_2+\lambda_3} &+ \frac{1}{\lambda_3+\lambda_4+\lambda_5} - \frac{1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5}. \end{aligned}$$

4.10.2 Covariance of Continuous Random Variables

(Adopted from [5], Example 4.56.) Calculate the linear correlation coefficient of continuous random variables X and Y distributed according to the joint probability density

$$f_{X,Y}(x,y) = 2\,\mathrm{e}^{-x}\mathrm{e}^{-y}H(y)H(x-y), \quad -\infty < x, y < \infty,$$

where H is the Heaviside function (see (2.8)).

✎ The linear correlation coefficient ρ_{XY} of variables X and Y (see (4.21)) is equal to the ratio of covariance σ_{XY} to the product of their effective deviations σ_X and σ_Y. First we need to calculate the expected value of the product XY,

$$\begin{aligned} E[XY] = \overline{XY} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\,\mathrm{e}^{-x}\mathrm{e}^{-y}H(y)H(x-y)\,\mathrm{d}x\,\mathrm{d}y \\ &= 2\int_0^{\infty} x\,\mathrm{e}^{-x}\left[\int_0^x y\,\mathrm{e}^{-y}\,\mathrm{d}y\right]\mathrm{d}x \\ &= 2\int_0^{\infty} x\,\mathrm{e}^{-x}\Big[1-(1+x)\mathrm{e}^{-x}\Big]\mathrm{d}x = \ldots = 1, \end{aligned}$$

then the expected values of X, Y, X^2 and Y^2,

$$\begin{aligned} E[X] = \overline{X} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \frac{3}{2}, \\ E[Y] = \overline{Y} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2}, \\ E[X^2] = \overline{X^2} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \frac{7}{2}, \\ E[Y^2] = \overline{Y^2} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2 f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2}. \end{aligned}$$

It follows that

$$\sigma_X = \sqrt{\overline{X^2} - \overline{X}^2} \approx 1.118, \qquad \sigma_Y = \sqrt{\overline{Y^2} - \overline{Y}^2} = 0.5,$$

hence

$$\mathrm{cov}[X,Y] = \sigma_{XY} = \overline{XY} - \overline{X}\,\overline{Y} = 1 - \frac{3}{2}\frac{1}{2} = \frac{1}{4}$$

and

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \approx 0.447.$$

4.10.3 Conditional Expected Values of Two-Dimensional Distributions

Let us return to the Example on p. 49 involving two random variables, distributed according to the joint probability density

$$f_{X,Y}(x,y) = \begin{cases} 8xy\,; & 0 \le x \le 1, 0 \le y \le x, \\ 0\,; & \text{elsewhere.} \end{cases}$$

Find ① the conditional expected value of the variable Y, given $X = x$, and ② the conditional expected value of the variable X, given $Y = y$!

✎ We have already calculated the conditional densities $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ in (2.28) and (2.29), so the conditional expected value ① equals

$$E\big[Y|X=x\big] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\,\mathrm{d}y = \int_0^x y\,\frac{2y}{x^2}\,\mathrm{d}y = \frac{2x}{3},$$

and the conditional expected value ② is

$$E\big[X|Y=y\big] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x = \int_y^1 x\,\frac{2x}{1-y^2}\,\mathrm{d}x = \frac{2(1-y^3)}{3(1-y^2)} = \frac{2(1+y+y^2)}{3(1+y)}.$$

4.10.4 Expected Values of Hyper- and Hypo-exponential Variables

Calculate the expected value, the second moment and the variance of continuous random variables, distributed according to the ① hyper-exponential (see (3.26)) and ② hypo-exponential distribution (see (3.28)).

✎ ① The hyper-exponential distribution, which describes a mixture (superposition) of k independent phases of a parallel process, whose ith phase proceeds with probability P_i and time constant $\lambda_i = 1/\tau_i$, is defined by the probability density

$$f_X(x) = \sum_{i=1}^{k} P_i f_{X_i}(x) = \sum_{i=1}^{k} P_i \lambda_i\,\mathrm{e}^{-\lambda_i x}, \quad x \ge 0, \tag{4.30}$$

where $0 \le P_i \le 1$ and $\sum_{i=1}^{k} P_i = 1$. The expected value of a hyper-exponentially distributed variable X is

$$\overline{X} = E[X] = \int_0^{\infty} x f_X(x)\,\mathrm{d}x = \sum_{i=1}^{k} P_i \int_0^{\infty} \lambda_i x\,\mathrm{e}^{-\lambda_i x}\,\mathrm{d}x = \sum_{i=1}^{k} \frac{P_i}{\lambda_i}, \tag{4.31}$$

and its second moment is

$$\overline{X^2} = E[X^2] = \int_0^\infty x^2 f_X(x)\,\mathrm{d}x = \sum_{i=1}^{k} P_i \int_0^\infty \lambda_i x^2\,\mathrm{e}^{-\lambda_i x}\,\mathrm{d}x = 2\sum_{i=1}^{k} \frac{P_i}{\lambda_i^2}.$$

Its variance is therefore

$$\mathrm{var}[X] = \sigma_X^2 = E[X^2] - E[X]^2 = 2\sum_{i=1}^{k} \frac{P_i}{\lambda_i^2} - \left(\sum_{i=1}^{k} \frac{P_i}{\lambda_i}\right)^2. \qquad (4.32)$$

While $\sigma_X/\overline{X} = \lambda/\lambda = 1$ holds true for the usual single-exponential distribution, its hyper-exponential generalization always has $\sigma_X/\overline{X} > 1$, except when all λ_i are equal: this inequality is the origin of the root "hyper" in its name.

② The hypo-exponential distribution describes the distribution of the sum of k ($k \geq 2$) independent continuous random variables X_i, in which each term separately is distributed exponentially with parameter λ_i. The sum variable $X = \sum_{i=1}^{k} X_i$ has the probability density

$$f_X(x) = \sum_{i=1}^{k} \alpha_i \lambda_i\,\mathrm{e}^{-\lambda_i x}, \qquad (4.33)$$

where

$$\alpha_i = \prod_{\substack{j=1 \\ j \neq i}}^{k} \frac{\lambda_j}{\lambda_j - \lambda_i}, \quad i = 1, 2, \ldots, k.$$

By comparing (4.33) to (4.30) one might conclude that the coefficients α_i represent the probabilities P_i for the realization of the ith random variable, but we are dealing with a serial process here: all indices i come into play—see Fig. 3.13! On the other hand, one *can* exploit the analytic structure of expressions (4.31) and (4.32), one simply needs to replace all P_i by α_i. By a slightly tedious calculation (or by exploiting the linearity of $E[\cdot]$ and using formula (4.20)) we obtain very simple expressions for the average and variance:

$$E[X] = \overline{X} = \sum_{i=1}^{k} \frac{1}{\lambda_i}, \quad \mathrm{var}[X] = \sigma_X^2 = \sum_{i=1}^{k} \frac{1}{\lambda_i^2}.$$

It is easy to see—Pythagoras's theorem comes in handy—that one always has $\sigma_X/\overline{X} < 1$. The root "hypo" in the name of the distribution expresses precisely this property.

4.10.5 Gaussian Noise in an Electric Circuit

The noise in electric circuits is frequently of Gaussian nature. Assume that the noise (random variable X) is normally distributed, with average $\overline{X} = 0\,\mathrm{V}$ and variance $\sigma_X^2 = 10^{-8}\,\mathrm{V}^2$. ① Calculate the probability that the noise exceeds the value $10^{-4}\,\mathrm{V}$ and the probability that its value is on the interval between $-2 \cdot 10^{-4}\,\mathrm{V}$ and $10^{-4}\,\mathrm{V}$! ② What is the probability that the noise exceeds $10^{-4}\,\mathrm{V}$, given that it is positive? ③ Calculate the expected value of $|X|$.

✎ It is worthwhile to convert the variable $X \sim N(\overline{X}, \sigma_X^2)$ to the standardized form

$$Z = \frac{X - \overline{X}}{\sigma_X} = \frac{X - 0\,\mathrm{V}}{10^{-4}\,\mathrm{V}} = 10^4 X,$$

so that $Z \sim N(0, 1)$. The required probabilities ① are then

$$P\big(X > 10^{-4}\,\mathrm{V}\big) = P(Z > 1) = 0.5 - \int_0^1 f_Z(z)\,\mathrm{d}z \approx 0.5 - 0.3413 = 0.1587$$

and

$$\begin{aligned} P\big(-2 \times 10^{-4}\,\mathrm{V} < X < 10^{-4}\,\mathrm{V}\big) &= P(-2 < Z < 1) = P(0 \le Z < 1) + P(0 \le Z < 2) \\ &= \int_0^1 f_Z(z)\,\mathrm{d}z + \int_0^2 f_Z(z)\,\mathrm{d}z \\ &\approx 0.3413 + 0.4772 = 0.8185, \end{aligned}$$

where the probability density f_Z is given by (3.10). We have read off the numerical values of the integrals from Table D.1.

② The required conditional probability is

$$\begin{aligned} P\big(X > 10^{-4}\,\mathrm{V} | X > 0\,\mathrm{V}\big) &= P(Z > 1 | Z > 0) \\ &= \frac{P(Z > 1 \cap Z > 0)}{P(Z > 0)} = \frac{P(Z > 1)}{P(Z > 0)} = \frac{P(Z > 1)}{0.5} \approx 0.3174. \end{aligned}$$

③ Since $Z = 10^4 X$, we also have $E[|Z|] = E\big[10^4|X|\big] = 10^4 E[|X|]$, so we need to compute

$$E[|Z|] = \int_{-\infty}^{\infty} |z| f_Z(z)\,\mathrm{d}z = 2\int_0^{\infty} z f_Z(z)\,\mathrm{d}z = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} z\,\mathrm{e}^{-z^2/2}\,\mathrm{d}z = \sqrt{\frac{2}{\pi}} \int_0^{\infty} \mathrm{d}\left(\mathrm{e}^{-x}\right) = \sqrt{\frac{2}{\pi}}$$

and revert to the old variable, hence $E[|X|] = 10^{-4}\sqrt{2/\pi}\,\mathrm{V}$.

4.10.6 *Error Propagation in a Measurement of the Momentum Vector* ⋆

We are measuring the time t in which a non-relativistic particle of mass m and momentum p traverses a distance L (that is, $t = L/v = mL/p$), and the spherical angles θ and ϕ of the vector $\mathbf{p}$ relative to the z-axis. Suppose that we have measured the average values $1/p = 5\,(\mathrm{GeV}/c)^{-1}$, $\theta = 75^\circ$ and $\phi = 110^\circ$, but all measurements contain one-percent uncertainties $\Delta(1/p) \equiv \sigma_p = 0.05\,(\mathrm{GeV}/c)^{-1}$, $\Delta\theta \equiv \sigma_\theta = 0.75^\circ$ and $\Delta\phi \equiv \sigma_\phi = 1.1^\circ$, which are uncorrelated. Determine the uncertainties of the quantities

$$p_x = p\sin\theta\cos\phi, \quad p_y = p\sin\theta\sin\phi, \quad p_z = p\cos\theta!$$

✎ In the notation of Sect. 4.9 we are dealing with the variables

$$X_1 = 1/p, \quad X_2 = \theta, \quad X_3 = \phi,$$

with the averages $\mu_1 = 5\,(\mathrm{GeV}/c)^{-1}$, $\mu_2 = 75^\circ$ and $\mu_3 = 110^\circ$. The corresponding covariance matrix (omitting the units for clarity) is

$$\Sigma(\boldsymbol{X}) = \begin{pmatrix} \sigma_p^2 & 0 & 0 \\ 0 & \sigma_\theta^2 & 0 \\ 0 & 0 & \sigma_\phi^2 \end{pmatrix} \approx \begin{pmatrix} 0.0025 & 0 & 0 \\ 0 & 0.000171 & 0 \\ 0 & 0 & 0.000369 \end{pmatrix}.$$

We need to calculate the covariance matrix of the variables

$$Y_1 = p_x = \frac{1}{X_1}\sin X_2\cos X_3, \quad Y_2 = p_y = \frac{1}{X_1}\sin X_2\sin X_3, \quad Y_3 = p_z = \frac{1}{X_1}\cos X_2,$$

and we need the derivatives (4.26) to do that:

$$\begin{aligned}
\frac{\partial Y_1}{\partial X_1} &= -\frac{1}{X_1^2}\sin X_2\cos X_3, & \frac{\partial Y_1}{\partial X_2} &= \frac{1}{X_1}\cos X_2\cos X_3, & \frac{\partial Y_1}{\partial X_3} &= -\frac{1}{X_1}\sin X_2\sin X_3, \\
\frac{\partial Y_2}{\partial X_1} &= -\frac{1}{X_1^2}\sin X_2\sin X_3, & \frac{\partial Y_2}{\partial X_2} &= \frac{1}{X_1}\cos X_2\sin X_3, & \frac{\partial Y_2}{\partial X_3} &= \frac{1}{X_1}\sin X_2\cos X_3, \\
\frac{\partial Y_3}{\partial X_1} &= -\frac{1}{X_1^2}\cos X_2, & \frac{\partial Y_3}{\partial X_2} &= -\frac{1}{X_1}\sin X_2, & \frac{\partial Y_3}{\partial X_3} &= 0.
\end{aligned}$$

When these expressions are arranged in the 3×3 matrix D, (4.27) immediately yields

$$\Sigma(\boldsymbol{Y}) = D\Sigma(\boldsymbol{X})D^{\mathrm{T}} = \begin{pmatrix} \sigma_{p_x}^2 & \sigma_{p_xp_y} & \sigma_{p_xp_z} \\ \sigma_{p_yp_x} & \sigma_{p_y}^2 & \sigma_{p_yp_z} \\ \sigma_{p_zp_x} & \sigma_{p_zp_y} & \sigma_{p_z}^2 \end{pmatrix} \approx 10^{-7} \begin{pmatrix} 126.4 & 30.74 & 2.440 \\ 30.74 & 53.10 & -6.704 \\ 2.440 & -6.704 & 66.63 \end{pmatrix}.$$

The uncertainties of p_x, p_y and p_z then become

$$\sigma_{p_x} = \sqrt{\Sigma_{11}(\boldsymbol{Y})} \approx 0.00355, \quad \sigma_{p_y} = \sqrt{\Sigma_{22}(\boldsymbol{Y})} \approx 0.00230,$$
$$\sigma_{p_z} = \sqrt{\Sigma_{33}(\boldsymbol{Y})} \approx 0.00258.$$

The propagation of the one-percent errors on the variables $1/p$, θ and ϕ has therefore resulted in more than one-percent errors on the variables p_x, p_y and p_z:

$$\begin{aligned} p_x &= (-0.0661 \pm 0.0036)\,\mathrm{GeV}/c = -0.0661(1 \pm 0.054)\,\mathrm{GeV}/c,\\ p_y &= (0.182 \pm 0.0023)\,\mathrm{GeV}/c = 0.182(1 \pm 0.013)\,\mathrm{GeV}/c,\\ p_z &= (0.0518 \pm 0.0026)\,\mathrm{GeV}/c = 0.0518(1 \pm 0.050)\,\mathrm{GeV}/c. \end{aligned}$$

The error of p_x and $p_z = p\cos\theta$ has increased dramatically. A feeling for why this happens in p_z can be acquired by simple differentiation $\mathrm{d}p_z = \mathrm{d}p\cos\theta - p\sin\theta\,\mathrm{d}\theta$ or

$$\frac{\Delta p_z}{p\cos\theta} = \frac{\Delta p}{p} - \frac{\sin\theta}{\cos\theta}\,\Delta\theta.$$

The average value of θ is not very far from 90°, where $\sin\theta \approx 1$ and $\cos\theta \approx 0$. Any error in $\Delta\theta$ in this neighborhood, no matter how small, is amplified by the large factor $\tan\theta$ that even diverges as $\theta \to \pi/2$.

In addition, the covariances $\sigma_{p_xp_y} = \sigma_{p_yp_x}$, $\sigma_{p_xp_z} = \sigma_{p_zp_x}$ and $\sigma_{p_yp_z} = \sigma_{p_zp_y}$ are all non-zero, and the corresponding correlation coefficients are

$$\rho_{p_xp_y} = \frac{\sigma_{p_xp_y}}{\sigma_{p_x}\sigma_{p_y}} \approx 0.375, \quad \rho_{p_xp_z} = \frac{\sigma_{p_xp_z}}{\sigma_{p_x}\sigma_{p_z}} \approx 0.027, \quad \rho_{p_yp_z} = \frac{\sigma_{p_yp_z}}{\sigma_{p_y}\sigma_{p_z}} \approx -0.113.$$

References

1. E. Brynjolfsson, Y.J. Hu, D. Simester, Goodbye Pareto principle, hello long tail: the effect of search costs on the concentration of product sales. Manage. Sci. **57**, 1373 (2011)
2. E. Brynjolfsson, Y.J. Hu, M.D. Smith, The longer tail: the changing shape of Amazon's sales distribution curve. http://dx.doi.org/10.2139/ssrn.1679991. 20 Sep 2010
3. C. Anderson, *The Long Tail: Why the Future of Business is Selling Less of More* (Hyperion, New York, 2006)
4. F. James, *Statistical Methods in Experimental Physics*, 2nd edn. (World Scientific, Singapore, 2010)
5. Y. Viniotis, *Probability and Random Processes for Electrical Engineers* (WCB McGraw-Hill, Singapore, 1998)

Chapter 5
Special Discrete Probability Distributions

Abstract The binomial (Bernoulli), multinomial, negative binomial (Pascal), and Poisson distributions are presented as the most frequently occurring discrete probability distributions in practice. The normal approximation of the binomial distribution is introduced as an example of the Laplace limit theorem, and the Poisson distribution is shown to represent a special limiting case of the binomial.

In this chapter we discuss distributions of discrete random variables, of which the binomial and the Poisson distributions are the most important.

5.1 Binomial Distribution

We are dealing with the binomial (Bernoulli) distribution whenever many random, mutually independent ("Bernoulli") trials yield only two kinds of outcomes—something *occurs* (probability p) or *does not occur* (probability $q = 1 - p$). Tossing a coin results in heads or tails; a girl or a boy is born; the weather prediction for tomorrow is rainy or dry. The probability that in N trials we encounter n outcomes of "type p" and $N - n$ outcomes of "type q" counted by the random variable X, is given by a two-parameter distribution

$$P(X = n; N, p) = \binom{N}{n} p^n q^{N-n}, \qquad n = 0, 1, 2, \ldots, N, \tag{5.1}$$

with parameters N and $p = 1 - q$. The structure $p^n q^{N-n}$ is obvious: since the "p" events and "q" events are mutually independent, we simply multiply the probability of "p" occurring n-times and "q" occurring $(N - n)$-times. We just need to figure out the number of ways such a combination can be accomplished: it is given by the binomial symbol (1.5). Of course, the distribution is normalized,

$$\sum_{n=0}^{N} P(X = n; N, p) = \sum_{n=0}^{N} \binom{N}{n} p^n q^{N-n} = (q + p)^N = 1, \tag{5.2}$$

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_5

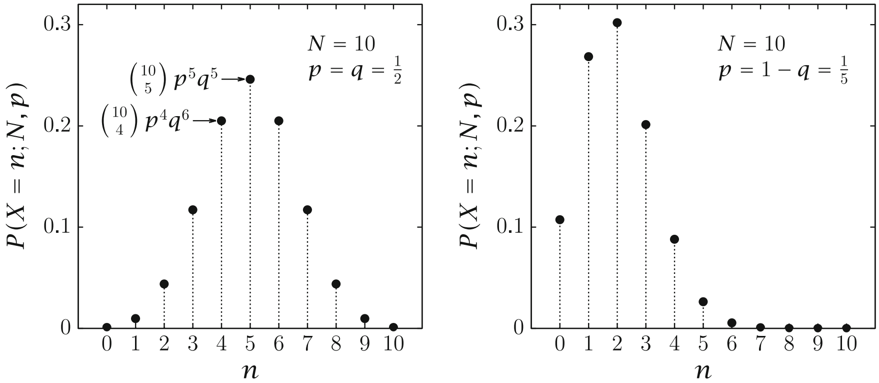


Fig. 5.1 Binomial (Bernoulli) distribution for n outcomes of "type p" with $N = 10$ trials. [Left] The distribution with parameter $p = \frac{1}{2}$ is symmetric around $N/2$. [Right] A distribution with parameter $p < \frac{1}{2}$ (in this case $p = \frac{1}{5}$) is squeezed towards the origin; for $p > \frac{1}{2}$ it is pushed towards N

where we have used the binomial expansion (1.6). The examples of the binomial distribution with parameters $N = 10$, $p = 1/2$ and $N = 10$, $p = 1/5$ are shown in Fig. 5.1. The distribution with $p = 1/2$ is symmetric about its average value; the trend of its values (filled circles) vaguely reminds us of the normal distribution; this will be exploited later on (Sect. 5.4).

Example A six-sided fair die is thrown five times. What is the probability of obtaining 3 dots exactly twice ($X = 2$) in these five trials ($N = 5$)? The probability of the outcome "3 dots" in each throw is $p = 1/6$, while the probability for any other outcome is $q = 1 - p = 5/6$. Hence the probability is

$$P(X = 2; N, p) = \binom{5}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{5-2} = \frac{625}{3888} \approx 0.161.$$

What is the probability that three dots appear *at most once* ($X \leq 1$)? There are only two possibilities: they do not appear at all ($X = 0$) or precisely once ($X = 1$). These events are mutually exclusive, so

$$P(X \leq 1) = P(X = 0) + P(X = 1) = \underbrace{\binom{5}{0}\left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^5}_{3125/7776} + \underbrace{\binom{5}{1}\left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^4}_{3125/7776} \approx 0.804.$$

What is the probability of having three dots *at least twice* ($X \geq 2$)? The strenuous path to answering this is

$$\begin{aligned}P(X \geq 2) &= P(X=2)+P(X=3)+P(X=4)+P(X=5)\\ &= \underbrace{\binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3}_{625/3888}+\underbrace{\binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2}_{125/3888}+\underbrace{\binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^1}_{25/7776}+\underbrace{\binom{5}{5}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^0}_{1/7776}\\ &\approx 0.196,\end{aligned}$$

while an easy one (complementarity of events!) is $P(X \geq 2) = 1 - P(X \leq 1)$. ◁

Mini-example The unstable meson η can decay in a variety of decay modes: $\eta \to 2\gamma$ (mode 1), $\eta \to 3\pi^0$ (mode 2), $\eta \to \pi^+\pi^-\pi^0$ (mode 3), ... with branching fractions (see definition (3.27)) $\mathrm{Br}_1 \approx 39.4\%$, $\mathrm{Br}_2 \approx 32.5\%$, $\mathrm{Br}_3 \approx 22.6\%$, ... Suppose we observe N decays, x of which are of type 2: then X is a binomially distributed random variable—as we are going to obtain a different x at fixed N in any experiment—with parameters N and $p = \mathrm{Br}_2$. ◁

Example The waiting time in a cafeteria is an exponentially distributed random variable with the average of 4 min. What is the probability that the student will be served in less than 3 min on at least four of the following six days?

The probability that the student (on any day) *has not been served* in 4 min decays exponentially: $P(t) = \mathrm{e}^{-t/\tau}$, where $\tau = 4$ min. The probability of him being served in less than 3 min is $p = 1 - \mathrm{e}^{-3/4}$. We must consider all possibilities on consecutive days, which is accounted for by the binomial distribution. On each day there are only two options: he is served in less than 3 min (probability p) or later than that (probability $1-p$). Thus the probability we are looking for is

$$\sum_{n=4}^{6}\binom{6}{n}p^n(1-p)^{6-n} = \sum_{n=4}^{6}\binom{6}{n}\left(1-\mathrm{e}^{-3/4}\right)^n\left(\mathrm{e}^{-3/4}\right)^{6-n} \approx 0.3969.$$

Calculate the probability that the student is served quickly (in less than 3 min) on at least one day out of six and the probability of him being served quickly precisely on day six! How does the latter result change when τ is modified? (Hint: expand the exponential up to the linear term.) ◁

Example The reliability of an airplane engine (the probability of it functioning flawlessly) is p. The airplane is able to fly if at least half of its engines operate. For which values of p a four-engine airplane is safer than a two-engine airplane?

A two-engine airplane can fly if at least one engine is operational, i.e. with probability

$$P_2 = \binom{2}{1}p(1-p) + \binom{2}{2}p^2 = 2p - p^2.$$

A four-engine airplane can fly if at least two engines operate, i.e. with probability

$$P_4 = \binom{4}{2} p^2(1-p)^2 + \binom{4}{3} p^3(1-p) + \binom{4}{4} p^4 = 6p^2(1-p)^2 + 4p^3(1-p) + p^4.$$

We are seeking values of p satisfying the inequality $P_4 > P_2$. After rearranging the terms we obtain

$$(p-1)^2(3p-2) > 0.$$

Since $0 \leq p \leq 1$ only the second factor is relevant: thus four engines are safer than two if $2/3 < p \leq 1$. A consolation for potentially frightened passengers: if $p = 0.9995$, which is a poor engine by modern engineering standards, one still has $P_2 = 0.99999975$ and $P_4 = 0.9999999995$. ◁

5.1.1 Expected Value and Variance

The expected value (average) and the variance of a binomially distributed random variable X can be calculated by substituting $p \to \lambda p$ in (5.2), computing the first and second derivative with respect to λ, and finally resetting $\lambda \to 1$. Thus

$$\sum_{n=0}^{N} \binom{N}{n} (\lambda p)^n q^{N-n} = (\lambda p + q)^N,$$

of which the first derivative with respect to λ yields

$$\sum_{n=0}^{N} n\lambda^{n-1} \binom{N}{n} p^n q^{N-n} = N(\lambda p + q)^{N-1} p, \tag{5.3}$$

and the second derivative gives

$$\sum_{n=0}^{N} n(n-1)\lambda^{n-2} \binom{N}{n} p^n q^{N-n} = N(N-1)(\lambda p + q)^{N-2} p^2. \tag{5.4}$$

When $\lambda = 1$ is restored, the left-hand side of (5.3) is precisely the expression for the expected value of X, while the left-hand side of (5.4) is the expected value of its function $X(X-1)$. The first equation therefore gives

$$E[X] = \overline{X} = N(p+q)^{N-1} p = Np,$$

while the second equation yields

$$E\big[X(X-1)\big] = \overline{X(X-1)} = \overline{X^2} - \overline{X} = N(N-1)(p+q)^{N-2}p^2 = N(N-1)p^2.$$

(According to the convention (4.2) we denote the expected values by a line over the corresponding random quantity.) Finally, both results can be combined to calculate the variance:

$$\begin{aligned}\sigma_X^2 = \mathrm{var}[X] &= \overline{\big(X-\overline{X}\big)^2} = \overline{X^2} - \overline{2X\overline{X}} + \overline{X}^2 = \overline{X^2} - \overline{X}^2 = \overline{X(X-1)} + \overline{X} - \overline{X}^2 \\ &= N(N-1)p^2 + Np - N^2p^2 = Np\big[(N-1)p + 1 - Np\big] = Npq.\end{aligned}$$

Let us summarize:

$$\overline{X} = Np, \qquad \sigma_X^2 = Npq. \tag{5.5}$$

If we interpret σ_X as the uncertainty of the measured number of events—do not confuse it with the σ parameter of the normal distribution!—we have

$$X_{\mathrm{meas}} = \overline{X} \pm \sigma_X = Np \pm \sqrt{Npq}. \tag{5.6}$$

What does that mean for the empirical determination of probabilities in Bernoulli trials? If in N trials we observe X_{meas} "good" and $N - X_{\mathrm{meas}}$ "bad" outcomes, the ratios $\widetilde{p} = X_{\mathrm{meas}}/N$ and $\widetilde{q} = (N - X_{\mathrm{meas}})/N$ at large enough N become good approximations to the unknown probabilities p and q. In this case we may write

$$\sqrt{Npq} \approx \sqrt{N\widetilde{p}\widetilde{q}} = \sqrt{N\widetilde{p}(1-\widetilde{p})}$$

and use (5.6) to express p:

$$p = \widetilde{p} \pm \sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{N}} = \widetilde{p}\left[1 \pm \sqrt{\frac{1-\widetilde{p}}{X_{\mathrm{meas}}}}\right]. \tag{5.7}$$

A method to calculate arbitrary moments of discrete random variables directly by means of probability generating functions is discussed in Appendix B.1.

Example (Adapted from [1].) Initially we have $N = 100$ radioactive nuclei, $X_{\mathrm{meas}} = n = 15$ of which remain "alive" after $t = 10$ s. How accurately can one determine the half-time ($t_{1/2} = \tau \log 2$) based on this information? We use (5.7) to calculate the probability p (and its uncertainty) of having n undecayed nuclei at time t:

$$p = \widetilde{p} \pm \sigma_{\widetilde{p}} = \frac{n}{N} \pm \sqrt{\frac{n(N-n)}{N^3}} = 2^{-t/t_{1/2}}.$$

It follows that

$$t_{1/2} = -\frac{t\log 2}{\log(\widetilde{p} \pm \sigma_{\widetilde{p}})} \approx -\frac{t\log 2}{\log \widetilde{p}}\left(1 \pm \frac{\sigma_{\widetilde{p}}}{\widetilde{p}\log \widetilde{p}}\right) = 3.65(1 \pm 0.29)\,\mathrm{s},$$

where we have used the small-x expansion of the function $1/\log(a+x)$ and inserted $\widetilde{p} = 0.15$ and $\sigma_{\widetilde{p}} = 0.0357$. ◁

5.2 Multinomial Distribution

The binomial distribution can be generalized by considering not just two kinds of outcomes with probabilities p and $q = 1 - p$ in N trials, where $n_p + n_q = N$, but having k types of outcomes with probabilities $p_1, p_2, \ldots, p_k$ and requiring

$$\sum\nolimits_{i=1}^{k} p_i = 1, \qquad \sum\nolimits_{i=1}^{k} n_i = N.$$

The probability that in N trials we obtain precisely n_1 outcomes of type 1, n_2 outcomes of type 2 and so on, is given by the *multinomial distribution*

$$P(X = n_1, \ldots, X = n_k;\ N, p_1, \ldots, p_k) = \binom{N}{n_1, n_2, \ldots, n_k} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k},$$

where the combinatorial factor at the right is the multinomial symbol (1.4). Let us assume that the ith outcome is a “good” event while all other outcomes are “bad”. This means that every random variable X_i by itself (with values n_i) is distributed binomially with parameters N and p_i. By (5.5) the expected value and variance of individual X_i's are

$$E[X_i] = \overline{X_i} = Np_i, \qquad \mathrm{var}[X_i] = \overline{X_i^2} - \overline{X_i}^2 = Np_i(1 - p_i),$$

while the covariances of the (X_i, X_j) pairs are

$$\mathrm{cov}[X_i, X_j] = \overline{\big(X_i - \overline{X_i}\big)\big(X_j - \overline{X_j}\big)} = -Np_ip_j, \qquad i \neq j.$$

Mini-example We measure the velocity distribution of molecules (we expect a result similar to Fig. 4.1 (right)) and arrange the data in a histogram with $k = 15$ equidistant bins [0, 100] m/s, [100, 200] m/s, and so on, up to [1400, 1500] m/s. Individual bins contain n_i molecules; all bins are mutually independent. In total we count $N = n_1 + n_2 + \cdots + n_{15}$ molecules. Such a histogram—it will change at each new measurement—represents a multinomial distribution. ◁

5.3 Negative Binomial (Pascal) Distribution

Suppose we observe a sequence of independent Bernoulli trials with probability p for a type-A outcome (e.g. an electronic circuit says "success") and probability $q = 1-p$ for a type-B outcome (circuit reports "failure"), as shown below:

$$\underbrace{\mathrm{AAAAAA}}_{6}\overbrace{\mathrm{B}}^{1}\underbrace{\mathrm{AAAAA}}_{5}\overbrace{\mathrm{BB}}^{2}\underbrace{\mathrm{AAAA}}_{4}\overbrace{\mathrm{B}}^{1}\underbrace{\mathrm{AAAAAAAA}}_{8}\overbrace{\mathrm{B}}^{1}$$

How long must we wait for r failures to occur? The probability to count n successes ($n = 6 + 5 + 4 + 8 = 23$ in the above sequence) before accumulating r failures ($r \geq 1$, $r = 5$ above) is given by the *negative binomial*[1] random variable with the distribution

$$P(X = n; r, p) = \binom{n + r - 1}{n} p^n (1 - p)^r, \qquad n = 0, 1, 2, \ldots \tag{5.8}$$

What is the probability of having n outcomes of *any kind* (A or B, variable Y), before encountering r failures? Because the sum of "good" and "bad" events, $n + r$, is constant, we just need to replace $n \to n - r$ in the definition, thus

$$P(Y = n; r, p) = \binom{n - 1}{n - r} p^{n-r} (1 - p)^r, \qquad n = r, r + 1, r + 2, \ldots$$

Both forms of the distribution are normalized, which one can check by using the formula

$$\sum_{n=0}^{\infty} \binom{n + r}{n} p^n = \frac{1}{(1 - p)^{r+1}}, \qquad 0 \leq p < 1.$$

5.3.1 Negative Binomial Distribution of Order k

Here is a tougher nut to crack: how long must we wait for k *consecutive* type-B outcomes or, more generally still, how long must we wait for r appearances of k consecutive type-B outcomes? One may imagine a device exposed to strong radiation that causes errors in its memory. The device is able to recover from these errors (B) and remains operational (A) until the radiation damage is so large that k consecutive

[1] The 'negative' attribute in the name of the distribution originates in the property

$$\binom{n + r - 1}{n} = (-1)^n \frac{(-r)(-r - 1)(-r - 2)\ldots(-r - n + 1)}{n!} = (-1)^n \binom{-r}{n}.$$

.

errors occur. For example, in the sequence

$$\underbrace{\text{BABAAABABABABAAABABAA}\,\overbrace{\text{BBB}}^{k}}_{n},$$

we have had $k = 3$ consecutive failures (B) after $n = 21$ outcomes of type A or B, while in the sequence

$$\underbrace{\text{BABAAABABBBABAAABABAA}\,\overbrace{\text{BBBB}}^{k}\,\text{AABABAABAAAABBAABA}\,\overbrace{\text{BBBB}}^{k}}_{n}$$

we have had two ($r = 2$) occurrences of a four-fold ($k = 4$) consecutive error after $n = 47$ outcomes. The probability for arbitrary k and r is given by the *negative binomial distribution of order k* [2]:

$$P(X\!=\!n;\,k,r,p) = \sum_{n_1,n_2,\ldots,n_k} \binom{n_1+n_2+\cdots+n_k+r-1}{n_1,\ n_2,\ \ldots,\ n_k,\ r-1} p^n \left(\frac{1-p}{p}\right)^{n_1+n_2+\cdots+n_k},$$

where $n \geq kr$ and where we sum over all non-negative integers $n_1, n_2, \ldots, n_k$, such that $n_1 + 2n_2 + \cdots + kn_k = n - kr$. An example of how this distribution is used can be found in Sect. 5.6.3.

5.4 Normal Approximation of the Binomial Distribution

If N is large and neither p nor q are too close to zero, the binomial distribution can be approximated by the normal distribution, although this appears to be something preposterous as the former is discrete, while the latter is continuous! This approximation is embodied in the *Laplace limit theorem*

$$\lim_{N\to\infty} P\left(a \leq \frac{X-Np}{\sqrt{Npq}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-x^2/2}\,\mathrm{d}x, \tag{5.9}$$

proven in Appendix B.3.1. In other words, the standardized binomial variable

$$\frac{X-\overline{X}}{\sigma_X} = \frac{X-Np}{\sqrt{Npq}}$$

is asymptotically normal. In practice, this applies already when $Np, Nq \gtrsim 5$.

Example A die is thrown 120 times. What is the probability of observing 4 dots no more than 18 times? The distribution of all $N = 120$ events is binomial, with probabilities $p = 1/6$ and $q = 1 - p = 5/6$. The exact answer—requiring us to calculate 19 terms and an overwhelming amount of factorials—is

$$\sum_{n=0}^{18} P(X = n; N, p) = \sum_{n=0}^{18} \binom{120}{n} \left(\frac{1}{6}\right)^n \left(\frac{5}{6}\right)^{120-n} \approx 0.3657. \tag{5.10}$$

By resorting to (5.9) we can obtain an approximate answer with much less effort. The formula requires us to calculate the average value and the effective deviation

$$\overline{X} = Np = 120\frac{1}{6} = 20, \qquad \sigma_X = \sqrt{Npq} = \sqrt{120\frac{1}{6}\frac{5}{6}} \approx 4.08,$$

and then calculate the standardized variables corresponding to the original (binomial) variables, i.e. the lower ($X = 0$) and upper ($X = 18$) summation index (see Fig. 5.2 (left)). One usually takes an 0.5 *smaller* lower value and an 0.5 *larger* upper value (see Fig. 5.2 (right)): this is the easiest way to approximate any discrete value $P(X = n)$ by the area under the curve of the probability density f_X on the interval $[n - 1/2, n + 1/2]$—and ensure that even by approximating a single point of a discrete distribution one obtains a non-zero result.

The boundary values of the standardized variables are $z_1 = (-0.5 - 20)/4.08 \approx -5.02$ and $z_2 = (18.5 - 20)/4.08 \approx -0.37$. By using Table D.1 we calculate

$$P(X \leq 18) \approx \Phi(-0.37) - \Phi(-5.02) \approx 0.3557,$$

where Φ is the distribution function of the standardized normal distribution. Compared to (5.10), this approximate probability is off by less than 3%. ◁

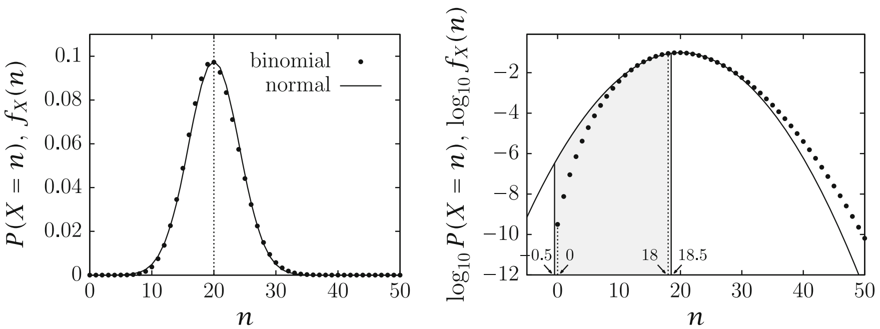


Fig. 5.2 [Left] Approximating the binomial distribution by a normal distribution in the case $N = 120$, $p = 1/6$ ($\overline{X} = Np = 20$, $\sigma_X = \sqrt{Npq} \approx 4.08$).] [Right] Same figure in logarithmic scale showing the boundary values (0 and 18) of the binomial distribution and the corresponding integration boundaries of the normal distribution

5.5 Poisson Distribution

The Poisson distribution is the limit of the binomial in which the probability p of an individual outcome becomes very small ($p \to 0$) and the number of trials very large ($N \to \infty$), such that the average $\overline{X} = Np$ remains unchanged. In each term of (5.1) we therefore write $p = \overline{X}/N$ and $q = 1 - p = 1 - \overline{X}/N$:

$$\binom{N}{n}\left(\frac{\overline{X}}{N}\right)^n\left(1-\frac{\overline{X}}{N}\right)^{N-n} = \frac{N(N-1)(N-2)\ldots(N-n+1)}{n!\,N^n}\,\overline{X}^n\left(1-\frac{\overline{X}}{N}\right)^{N-n}$$
$$= \frac{1}{n!}\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{n-1}{N}\right)\overline{X}^n\left(1-\frac{\overline{X}}{N}\right)^{N-n}.$$

In the limit $N \to \infty$, the last factor is just

$$\lim_{N\to\infty}\left(1-\frac{\overline{X}}{N}\right)^{N-n} = \lim_{N\to\infty}\left(1-\frac{\overline{X}}{N}\right)^{N}\left(1-\frac{\overline{X}}{N}\right)^{-n} = \mathrm{e}^{-\overline{X}},$$

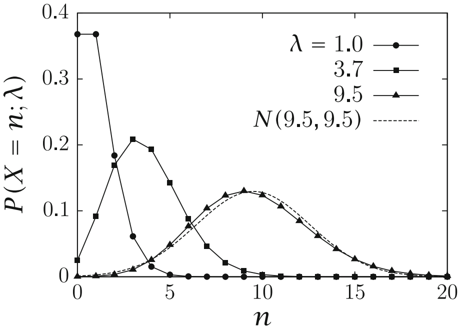
therefore $P(X = n; \overline{X}) = \overline{X}^n \mathrm{e}^{-\overline{X}}/n!$. We have obtained a *single-parameter* distribution with parameter $\overline{X}$, its expected value. It is traditionally denoted by $\overline{X} = \lambda$, so this important distribution, illustrated in Fig. 5.3, is usually written as

$$P(X = n; \lambda) = \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!}, \qquad n = 0, 1, 2, \ldots \tag{5.11}$$

No approximations have been made by taking the $p \to 0$ and $N \to \infty$ limits, just one parameter has evaporated. At its heart, then, the Poisson distribution is still "binomial", whence it also inherits the expressions for its average and variance; but $p \to 0$ implies $q \to 1$, thus

$$E[X] = \overline{X} = Np = \lambda, \qquad \mathrm{var}[X] = \sigma_X^2 = Npq = \overline{X} = \lambda. \tag{5.12}$$

Fig. 5.3 Three examples of the Poisson distribution with $\lambda = 1.0$, 3.7 and 9.5. For comparison, the density of the normal distribution $N(\mu = 9.5, \sigma^2 = 9.5)$ is also shown



Instead of (5.6) we may therefore write

$$X_{\text{meas}} = \overline{X} \pm \sqrt{\overline{X}}\,.$$

Example A total of $N = 2000$ people are vaccinated. The probability p for side effects is small, $p = 0.001$, therefore, on average, only $\lambda = Np = 2$ people will experience them. What is the probability that the number of people experiencing unwanted effects will be greater than two? The probability that precisely n people experience a side effect, is

$$P(X = n; \lambda) = \frac{\lambda^n \, \mathrm{e}^{-\lambda}}{n!}.$$

The probability we are seeking is therefore $P(>2) = P(3) + P(4) + \cdots + P(2000)$. The calculation of these 1998 values can be avoided by considering the complementary event: $P(>2) = 1 - P(0) - P(1) - P(2) \approx 0.323$. ◁

Classical example We count X nuclear decays in time t; let $X \gg 1$ (at least a few times ten). The estimate of the true source activity (decays per unit time) a is $\widehat{a} = X/t$. The measured X fluctuates about $\overline{X} = at$ by $\pm\sigma_X = \pm\sqrt{\overline{X}}$. But the true value $\overline{X}$ is unknown, hence we approximate $\sigma_X \approx \sqrt{X}$ and write

$$X = \overline{X} \pm \sqrt{\overline{X}} \approx \overline{X} \pm \sqrt{X} \quad \text{or} \quad \overline{X} \approx X \pm \sqrt{X}.$$

Dividing the second equation by t we obtain the relation between the true activity a and the measured value $\widehat{a}$:

$$a = \widehat{a}\left(1 \pm \frac{1}{\sqrt{X}}\right).$$

Therefore, if we wish to measure the source activity to a precision of 1%, we must count 10^4 decays. This obstacle awaits us in all experiments where anything is being "counted". A k-fold reduction in statistical uncertainty requires us to count k^2 times more events, i.e. measure k^2 times as long. ◁

Example (Adapted from [1].) On an average day the surface of the Earth (radius $R = 6400\,\mathrm{km}$) is hit by 25 meteorites. What is the probability that in 10 years at least one of its $N = 7 \cdot 10^9$ inhabitants will be hit by a meteorite?

The probability of an individual being hit is proportional to the ratio of surface areas S_1/S, where $S_1 \approx 0.2\,\mathrm{m}^2$ is the average surface area of a human being and $S = 4\pi R^2$ is the surface area of the Earth. In ten years the Earth receives $M = 10 \cdot 365 \cdot 25 = 91250$ meteorites, thus the expected number of hit people in this period of time is $\lambda = NMS_1/S = 0.248$. The probability that a meteorite hits *at least one person*, is therefore $1 - \lambda^0 \mathrm{e}^{-\lambda}/0! = 1 - \mathrm{e}^{-\lambda} \approx 0.22$. ◁

Example Let us stay with the dangers from the sky! When London was bombarded with German "flying bombs" during World War II, some people thought that hits tend to form clusters, looking as if the probabilities of certain areas being hit were relatively higher [3]. Can this assumption be justified?

There were 537 hits on the surface area of 144 km^2, divided into $24 \times 24 = 576$ quadrants with an area of 0.25 km^2 each, so the average number of hits in any quadrant was $\lambda = 537/576 \approx 0.9323$. If the points of impact were completely random, the probability that a chosen quadrant has been hit $n = 0, 1, 2, \ldots$ times, is given by the Poisson distribution

$$P(X = n) = \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!}, \qquad n = 0, 1, 2, \ldots$$

The expected number of quadrants with precisely n hits should therefore be 576 $P(X = n)$, for example, $576\,P(X = 0) = 576 \cdot \mathrm{e}^{-0.9323} \approx 226.74$ quadrants with no hits at all. The expected numbers of quadrants with n hits and the corresponding observed numbers are shown in the Table 5.1.

If the projectiles "preferred" specific quadrants, one should be able to see this primarily as a decrease of the number of quadrants with no hits and an increase in the middle portion of the distribution. But the excellent agreement of the expected and observed numbers proves that the distribution of hits is consistent with a random—Poisson—distribution. We shall put this statement on a more quantitative footing in Sect. 10.3. ◁

Example (Adapted from [1].) A gas mixture contains $10^9/\mathrm{cm}^3$ molecules of CH_4 endowed with the radioactive $^{14}\mathrm{C}$ isotope. A sample of $V = 1\,\mathrm{mm}^3$ is taken for analysis. What is the probability that the concentration of the radioactive admixture in the sample will exceed its average concentration by more than 0.2%?

On average, a sample will contain $\lambda = (10^9/\mathrm{cm}^3)V = 10^6$ radioactive molecules. The probability we wish to compute is the sum of probabilities that the sample contains $1.002\,\lambda$ or $(1.002\,\lambda + 1)$ or $(1.002\,\lambda + 2)$ molecules, and so on, thus

$$P = \sum_{n=n_{\mathrm{min}}}^{n_{\mathrm{max}}} \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!},$$

Table 5.1 The distribution of hits in World War II bombing raids over London

Number of bombs in quadrant	Expected number of quadrants	Observed number of quadrants
0	226.74	229
1	211.39	211
2	98.54	93
3	30.62	35
4	7.14	7
≥ 5	1.57	1

where $n_{\min} = 1.002\,\lambda$. How do we determine $n_{\max}$? Assume that the mixture is at standard conditions, where a mole of gas ($N_{\mathrm{A}} \approx 6 \cdot 10^{23}$ molecules) has a volume of $V_0 \approx 22.4\,\mathrm{dm}^3$. Therefore, a sample with volume V contains $n_{\max} = N_{\mathrm{A}} V / V_0 \approx 3 \cdot 10^{16}$ molecules, which by far exceeds the number of radioactive molecules in it, so we can safely set $n_{\max} = \infty$. For such high n we can approximate the Poisson distribution by the normal (see Sect. 5.4) and replace the sum of millions of billions of Poissonian contributions by an integral of the normal density with average $\mu = \lambda$ and variance $\sigma^2 = \lambda$ with the integration boundaries $1.002\,n$ and ∞, i.e. the standardized normal distribution with the boundaries

$$a = \frac{1.002\,\lambda - \lambda}{\sqrt{\lambda}} = \frac{0.002 \cdot 10^6}{\sqrt{10^6}} = 2, \qquad b = \infty.$$

By using (3.9), (3.12) and Table D.2 we obtain $P \approx \frac{1}{2}\bigl[1 - \mathrm{erf}(2/\sqrt{2})\bigr] \approx 0.02275$. A direct calculation of the sum by MATHEMATICA yields $P \approx 0.02280$. ◁

5.6 Problems

5.6.1 Detection Efficiency

(Adapted from [4].) Galactic sources of gamma radiation are measured by specially designed gamma-ray spectrometers. Assume that for a certain region of the sky we have used two different detectors (different electronics, varying atmospheric conditions and so on) and determined the following numbers of sources:

N_{12}	sources seen by both detectors,
$N_{12} + N_1$	sources seen by the first detector,
$N_{12} + N_2$	sources seen by the second detector,
$N - (N_{12} + N_1 + N_2)$	sources not seen by any of them.

Calculate the total detection efficiency—the ratio between the number of observed rays and the number of all incident rays—and its uncertainty! Note that the measurement has a binomial nature: any source is either detected (probability p) or missed (probability $q = 1 - p$).

✎ The efficiencies (which are the estimates of the true values based on repeated samples) for detection by a single spectrometer and for simultaneous detection are

$$\widehat{P}(1) = \varepsilon_1 = \frac{N_{12} + N_1}{N}, \qquad \widehat{P}(2) = \varepsilon_2 = \frac{N_{12} + N_2}{N}, \qquad \widehat{P}(1 \cap 2) = \frac{N_{12}}{N},$$

where N is the true number of the sources. Of course, because the values N_{12}, N_1 and N_2 fluctuate statistically, one can only obtain an estimate $\widehat{N}$ for N. The measurements with two spectrometers are mutually independent, $\widehat{P}(1 \cap 2) = \widehat{P}(1)\widehat{P}(2)$, and therefore

$$\widehat{N} = \frac{(N_{12} + N_1)(N_{12} + N_2)}{N_{12}}.$$

But the measurements with two devices ("event 1" and "event 2") are not mutually exclusive, implying $\widehat{P}(1 \cup 2) = \widehat{P}(1) + \widehat{P}(2) - \widehat{P}(1 \cap 2) = \widehat{P}(1) + \widehat{P}(2) - \widehat{P}(1)\widehat{P}(2)$, thus the total efficiency is

$$\varepsilon = \varepsilon_1 + \varepsilon_2 - \varepsilon_1\varepsilon_2 = \widehat{P}(1 \cup 2) = \frac{N_{12} + N_1 + N_2}{\widehat{N}} = 1 - \frac{N_1 N_2}{(N_{12} + N_1)(N_{12} + N_2)}.$$

The random variable X that "counts" good events (detected sources) is binomially distributed, the minimum and maximum numbers of detected sources being 0 and N, respectively. The relative number of the detected sources X/N therefore has the variance

$$\mathrm{var}\left[\frac{X}{N}\right] = \frac{1}{N^2}\,\mathrm{var}[X] = \frac{Npq}{N^2} = \frac{p(1-p)}{N}$$

(recall (4.12)). The variance of the efficiency of an individual detector is then

$$\sigma^2(\varepsilon_i) \approx \frac{\varepsilon_i(1-\varepsilon_i)}{\widehat{N}}, \qquad i = 1, 2.$$

The variance of the total efficiency $\varepsilon = \varepsilon_1 + \varepsilon_2 - \varepsilon_1\varepsilon_2$ is calculated by using (4.25):

$$\sigma^2(\varepsilon) \approx \left(\frac{\partial\varepsilon}{\partial\varepsilon_1}\right)^2 \sigma^2(\varepsilon_1) + \left(\frac{\partial\varepsilon}{\partial\varepsilon_2}\right)^2 \sigma^2(\varepsilon_2) = \frac{(1-\varepsilon_1)(1-\varepsilon_2)(\varepsilon_1 + \varepsilon_2 - 2\varepsilon_1\varepsilon_2)}{\widehat{N}}.$$

The total efficiency ε as a function of ε_1 and ε_2 is shown in Fig. 5.4 (left), while its uncertainty, multiplied by $\widehat{N}^{1/2}$, is shown in Fig. 5.4 (right).

5.6.2 The Newsboy Problem ★

A newsboy purchases his newspapers at a price of $a = 1.00$ € and sells them at $b = 1.50$ €. While he must purchase all copies from his supplier at once, his actual sales depend on the fluctuating daily demand. The long-term average of the daily demand is 100 copies, and he buys just as many from the supplier. ① Calculate the newsboy's daily profit by assuming that the demand is described by a random variable with a Poisson distribution with average $\lambda = 100$. ② How many copies should he purchase in order to maximize his profit?

✎ If the variable X represents the demand and n is the number of copies purchased by the newsboy, the variable $Y_n = \min\{X, n\}$ describes the number of copies sold

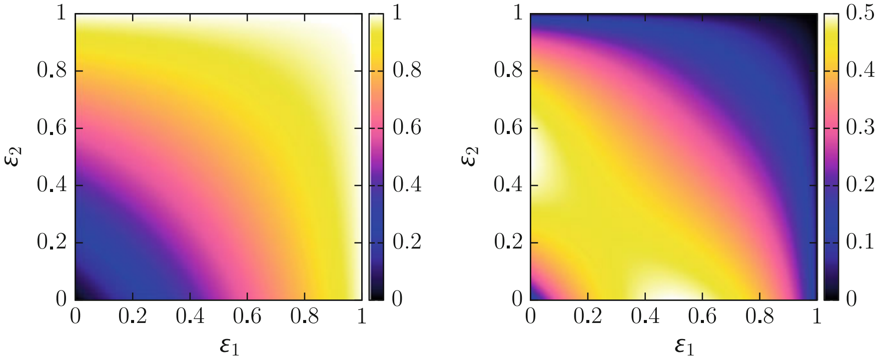


Fig. 5.4 [Left] The total efficiency ε for the detection of gamma-rays with two spectrometers as a function of individual detector efficiencies ε_1 and ε_2. [Right] The uncertainty of the total efficiency, multiplied by $\widehat{N}^{1/2}$. The uncertainty is smallest in the limit $\varepsilon_1, \varepsilon_2 \to 1$, but also when $\varepsilon_1, \varepsilon_2 \approx 0$, although in that corner, one also has $\varepsilon \approx 0$.

to the customers: if the demand is below the number of copies he has purchased, he sells X, while in the opposite case he sells n (since he has none left). ① His profit with Y_n sold copies (fluctuating by the day) is therefore measured by the random variable

$$\Pi_n = aY_n - bn,$$

and its expected value is $E[\Pi_n] = aE[Y_n] - bn$. It holds that

$$\begin{aligned}
E[Y_n] &= \sum_{k=0}^{n} k\,P\big(Y_n = k\big) = \sum_{k=1}^{n-1} k\,\underbrace{P\big(\min\{X,n\} = k\big)}_{P(X=k)} + n\,\underbrace{P\big(\min\{X,n\} = n\big)}_{P(X\geq n)=1-P(X<n)} \\
&= \sum_{k=1}^{n-1} k\,P(X=k) + n - n\sum_{k=0}^{n-1} P(X=k) \\
&= n + \sum_{k=1}^{n-1} k\,\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} - n\sum_{k=0}^{n-1}\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} \\
&= n + \lambda\,\mathrm{e}^{-\lambda}\sum_{k=0}^{n-2}\frac{\lambda^k}{k!} - n\,\mathrm{e}^{-\lambda}\sum_{k=0}^{n-2}\frac{\lambda^k}{k!} - n\,\mathrm{e}^{-\lambda}\frac{\lambda^{n-1}}{(n-1)!} \\
&= n + (\lambda - n)\,\mathrm{e}^{-\lambda}\sum_{k=0}^{n-2}\frac{\lambda^k}{k!} - n\,\mathrm{e}^{-\lambda}\frac{\lambda^{n-1}}{(n-1)!}.
\end{aligned}$$

The second term in the final expression is zero, since $\lambda = n = 100$ (the newsboy's daily purchase equals the average demand), therefore

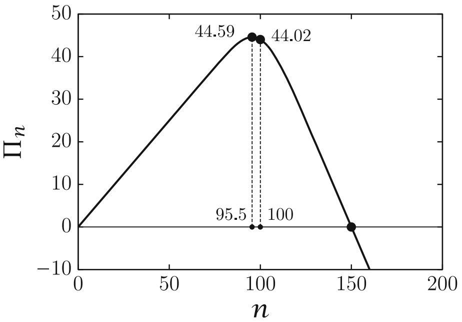


Fig. 5.5 The profit of the newsboy purchasing n newspaper copies $a = 1.00$ € from the supplier and selling them at $b = 1.50$ € to customers whose demand is modeled by a Poisson distribution with average 100

$$E[Y_n] = n\left[1 - \mathrm{e}^{-\lambda}\frac{\lambda^{n-1}}{(n-1)!}\right] \approx 96.01$$

or $E[\Pi_n] = aE[Y_n] - bn = 1.50\,€ \cdot 96.01 - 1.00\,€ \cdot 100 \approx 44.02\,€$.

② As shown in Fig. 5.5, the newsboy's daily purchase of 100 copies is not optimal: he could count on a maximum profit of 44.59 € by purchasing 95 or 96 copies per day (the maximum of Π_n is at $n \approx 95.5$); in all other cases his profit will be smaller than that. If he buys more than 150 copies per day, he will even lose money.

5.6.3 Time to Critical Error

A computer memory constantly reviews its "sanity": it may detect flawless operation (A) or error (B). The memory operates until encountering four consecutive errors (BBBB, "critical error"). The probability of a single error is $p = 0.5$. Calculate the probability that after $N = 40$ sanity checks the memory still operates. Use ① your knowledge of combinatorics and ② the negative binomial distribution of order k.

✎ ① In a random sequence of outcomes A and B we await the subsequence BBBB, which may occur at the very beginning, or the subsequence ABBBB that, however, should not be preceded by the critical BBBB quartet. The options are:

$N=4$	$N=5$	$N=6$	$N=7$	$N=8$
BBBB	ABBBB	AABBBB	AAABBBB	AAAABBBB
		BABBBB	ABABBBB	AABABBBB
			BAABBBB	ABAABBBB
			BBABBBB	BAAABBBB
				ABBABBBB
				BABABBBB
				BBAABBBB
				BBBABBBB

The probability of the critical error occurring already in the first $N = 4$ steps (immediate BBBB combination), is $P(4) = p^4 = 0.0625$. The probability of it occurring in $N = 5$ steps (subsequence ABBBB) is $P(5) = p^5 = 0.03125$, while in $N = 6$ steps (combinations AABBBB or BABBBB), it is $P(6) = 2p^1p^5 = 0.03125$. From the table above we further see that $P(7) = 2^2p^2p^5 = 0.03125$ and $P(8) = 2^3p^3p^5 = 0.03215$, but to conclude $P(9) = 2^4p^4p^5$ would be a mistake: indeed there would be $2^4 = 16$ possibilities before the critical part ABBBB of the whole sequence, but one of them has the form BBBBABBBB, which should be discarded as it already contains the terminating sequence BBBB at the very beginning, which brings us to the $N = 4$ case considered before. For $N = 9$, we therefore have $P(9) = (2^4 - 1)p^4p^5 = 0.0292968$. Analogously, the subsequent $P(N)$ are

$$P(N) = \left(1 - \frac{N_{\mathrm{BBBB}}}{2^{N-5}}\right)p^5,$$

where N_{BBBB} is the number of quartets of the form BBBB that precede the terminal quintet ABBBB in the sequence and should therefore *not* be considered. This number becomes increasingly difficult to determine at high N: for example, we have $N_{\mathrm{BBBB}} = 3$ for $N = 10$, $N_{\mathrm{BBBB}} = 8$ for $N = 11$, $N_{\mathrm{BBBB}} = 20$ for $N = 12$, and so on.

There is a more elegant solution. Define the probability $P(i)$ that the critical error occurs precisely at the ith place in the sequence. The first few values (from $i = 0$ to $i = 8$) are known from the previous discussion:

$$P(1) = P(2) = P(3) = 0, \qquad P(4) = \left(\tfrac{1}{2}\right)^4, \qquad P(5) = P(6) = P(7) = P(8) = \left(\tfrac{1}{2}\right)^5.$$

Now define the probability $\mathcal{P}(n)$ of the critical error occurring *anywhere* up to (including) the nth place,

$$\mathcal{P}(n) = \sum_{i=1}^{n} P(i).$$

From $N = 9$ onwards we therefore simply await the ABBBB subsequence which occurs with probability p^5, while no critical error should occur before that, leading to the recurrence

$$P(N) = \left[1 - \mathcal{P}(N-5)\right]p^5, \qquad N > 9.$$

The solution of the problem—the sum up to $N = 40$ can be evaluated by some symbolic computation program, e.g. Mathematica—is therefore

$$1 - \mathcal{P}(40) = 1 - \sum_{N=0}^{40} P(N) \approx 0.2496.$$

② Undoubtedly the solution by using the negative binomial distribution of order k is the simplest (we can use Mathematica again), but it is bereft of any insight into the heart of the problem:

$$1 - \sum_{N=4}^{40} P(N;\, k = 4, r = 1, p = 0.5) \approx 0.2496.$$

5.6.4 Counting Events with an Inefficient Detector

(Adapted from [4].) Charged particles are counted by a detector with a non-ideal efficiency: the probability to detect a particle is $p < 1$. Assume that the number X of particles traversing the detector in fixed time t is Poisson-distributed with average λ. What is the probability of detecting precisely r particles in time t?

✎ If we are supposed to count r particles, *at least* r particles should actually fly through the detector. The desired probability is therefore the sum of probabilities of detecting r particles while $n = r, r+1, r+2, \ldots$ particles have actually flown through. For given n the number r of actually detected particles is given by the binomial distribution; from the total probability formula (1.15) it follows that

$$\begin{aligned}
P(X = r) &= \sum_{n=r}^{\infty} P(r \text{ detected} \,|\, n \text{ traversed})\, P(n \text{ traversed}) \\
&= \sum_{n=r}^{\infty} P_{\text{binom}}(X = r;\, n, p)\, P_{\text{Poisson}}(X = n;\, \lambda) \\
&= \sum_{n=r}^{\infty} \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r} \cdot \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!} \\
&= \frac{1}{r!} \left(p\lambda\right)^r \mathrm{e}^{-\lambda} \sum_{n=r}^{\infty} \frac{\left[(1-p)\lambda\right]^{n-r}}{(n-r)!} = \frac{1}{r!} \left(p\lambda\right)^r \mathrm{e}^{-\lambda} \mathrm{e}^{(1-p)\lambda}
\end{aligned}$$

or

$$P(X = r) = \frac{1}{r!} \left(p\lambda\right)^r \mathrm{e}^{-p\lambda}, \qquad r = 0, 1, 2, \ldots,$$

which is nothing but the Poisson distribution with the expected value $p\lambda$. (In the theory of Poisson processes this $\lambda \to p\lambda$ effect is suggestively called *thinning*.)

5.6.5 Influence of Primary Ionization on Spatial Resolution ⋆

Charged particles flying through gas ionize its atoms and molecules. In this (so-called primary) ionization, a few times ten electron–ion pairs are generated per centimeter of the particle's path length through the gas at normal conditions, depending on

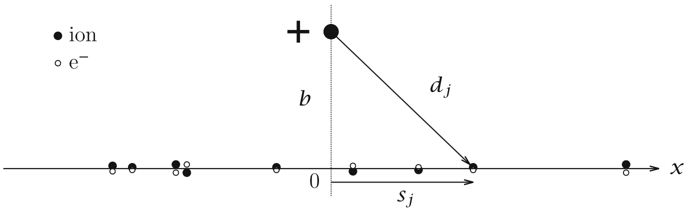


Fig. 5.6 A charged particle passing near the anode wire in a gas ionization detector. The jth electron–ion pair is formed at s_j. The electrons are attracted by the anode, and their drift time towards it is a measure of the impact distance of the original particle

the average atomic number of the gas, $\overline{Z}$. The average number of ionizations on distance x is κx, where $\kappa \approx 1.5\overline{Z}\,\mathrm{cm}^{-1}$ [5]. ① Determine the distribution of locations where the electron–ion pairs are formed! ② How does the discrete nature of the primary ionization (see Fig. 5.6) influence the spatial resolution of such an elementary detector?

✎ Ionizations are rare, independent events, so they obey the Poisson distribution: if κL is the average number of ionizations on distance L, the probability for n ionizations is

$$P\big(X(\kappa) = n\big) = \frac{(\kappa L)^n}{n!}\,\mathrm{e}^{-\kappa L}.$$

① For these n ionizations the distribution of each (jth) pair ($1 \le j \le n$) along x is

$$D_{nj}(x) = \frac{n!}{(n-j)!(j-1)!}\,(L-x)^{n-j}x^{j-1}\frac{1}{L^n}.$$

The function $D_{73}(x)$, for example, describes the point of creation of the third pair if seven pairs were created in total. (To understand the structure of this expression, plot D_{11}, then D_{21}, D_{22} and their sum—it is 2—then D_{31}, D_{32}, D_{33} and their sum—it is 3—and so on!) The distribution for the point of creation of the jth pair, if κ pairs per unit length were created, is then

$$A_j^{(\kappa)}(x) = \sum_{n=j}^{\infty} P\big(X(\kappa) = n\big)\,D_{nj}(x) = \kappa\frac{(\kappa x)^{j-1}}{(j-1)!}\,\mathrm{e}^{-\kappa x}, \qquad 0 \le x \le L.$$

These distributions are shown in Fig. 5.7 (left) for $\kappa = 10\,\mathrm{cm}^{-1}$, which approximately corresponds to neon at a pressure of 1.2 bar. Each function is normalized on $[0, \infty)$.

② The first factor, relevant for the spatial resolution of an ionization detector, is the non-uniformity of primary ionizations. The jth ionization occurs on distance s_j from the origin (see Fig. 5.6), but in both directions (negative or positive x), which can be absorbed in computing the moments of s_j by replacing $\kappa \to 2\kappa$. The average distance of the jth ionization and the average of its square are therefore

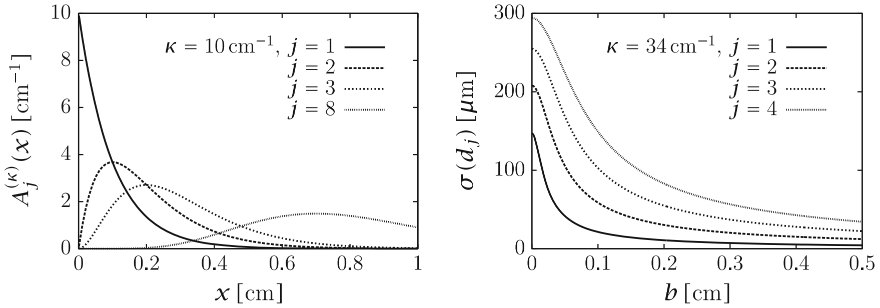


Fig. 5.7 [Left] The distribution of points of creation of the jth electron–ion pair. [Right] The influence of primary statistics on the spatial resolution

$$\overline{s}_j = \int_0^\infty x A_j^{(2\kappa)}(x)\,\mathrm{d}x = \frac{j}{2\kappa}, \qquad \overline{s_j^2} = \int_0^\infty x^2 A_j^{(2\kappa)}(x)\,\mathrm{d}x = \frac{j(j+1)}{4\kappa^2},$$

so the corresponding variance is

$$\sigma^2(s_j) = \overline{s_j^2} - \left(\overline{s}_j\right)^2 = \frac{j}{4\kappa^2}.$$

From Fig. 5.6 we see that $d_j = \sqrt{b^2 + s_j^2}$, thus, by error-propagation (4.25), we find

$$\sigma^2(d_j) = \left(\frac{\partial d_j}{\partial s_j}\right)^2 \bigg|_{s_j = \overline{s}_j} \sigma^2(s_j) = \frac{j^3}{4\kappa^2(j^2 + 4\kappa^2 b^2)}.$$

Figure 5.7 (right) shows the uncertainties $\sigma(d_j)$ for various j with $\kappa = 34\,\mathrm{cm}^{-1}$ (70% : 30% mixture of argon and isobutane at normal conditions). The spatial resolution therefore *deteriorates* with increasing number of primary ionizations.

References

1. I. Kuščer, A. Kodre, *Mathematik in Physik und Technik* (Springer, Berlin, 1993)
2. A.N. Philippou, The negative binomial distribution of order k and some of its properties. Biom. J. **26**, 789 (1984)
3. R.D. Clarke, An application of the Poisson distribution. J. Inst. Actuaries **72**, 481 (1946)
4. A.G. Frodesen, O. Skjeggestad, H. Tøfte, *Probability and Statistics in Particle Physics* (Universitetsforlaget, Bergen, 1979)
5. F. Sauli, *Principles of Operation of Multiwire Proportional and Drift Chambers* (CERN Reprint 77–09, Geneva, 1977)

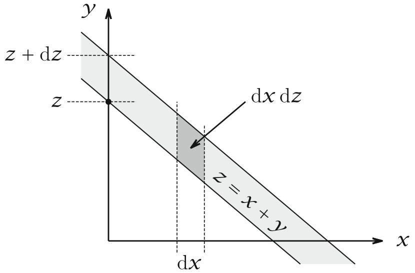
Chapter 6
Stable Distributions and Random Walks

Abstract Stable distributions are special types of probability distributions whose origin is a particular limit regime of other types of distributions. They are closely related to the simple convolution process, which is introduced first for continuous and then for discrete random variables. This leads to the central limit theorem as one of the most important results of probability theory, as well as to its generalized version which is useful in the analysis of random walks. Extreme-value distributions are also presented, as they possess a limit theorem of their own (Fisher–Tippett–Gnedenko). The last part is devoted to the discussion of discrete-time and continuous-time random walks, together with their characteristic diffusion properties.

In this chapter we sum independent random variables X_i and discuss what happens to the distribution of their sum, $Y = \sum_i X_i$. We shall see that the distribution of Y is given by the convolution of distributions of individual X_i's, and that in the case $i \to \infty$—under certain conditions—the distributions of Y tend to *stable distributions*, relevant for the processes of *random walks*.

6.1 Convolution of Continuous Distributions

What is the distribution of $Z = X + Y$ if continuous random variables X and Y correspond to densities $f_X(x)$ and $f_Y(y)$? We are interested in the probability that the sum $x + y$ falls within the interval $[z, z + \mathrm{d}z]$, where x and y are arbitrary within their own definition domains. All points fulfilling this requirement are represented by the oblique shaded area in the figure.



S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_6

One must add up all contributions to the probability within this band. The infinitesimal area $\mathrm{d}x\,\mathrm{d}z$ (shaded rhomboid) corresponds to the infinitesimal probability $f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}z$. By integrating over x we obtain the probability $f_Z(z)\,\mathrm{d}z$. Let us write only its density and insert $y = z - x$:

$$f_Z(z) = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x. \tag{6.1}$$

This operation is called the *convolution of distributions* and we denote it by the symbol $*$. If you do not trust this geometric argument, one can also reason as follows:

$$\begin{aligned} f_Z(z)\,\mathrm{d}z &= P(z \le Z \le z + \mathrm{d}z) = P(z \le X + Y \le z + \mathrm{d}z) \\ &= \int_{-\infty}^{\infty} \mathrm{d}x \int_{z-x}^{z-x+\mathrm{d}z} f_X(x) f_Y(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} f_X(x)\mathrm{d}x \underbrace{\int_{z-x}^{z-x+\mathrm{d}z} f_Y(y)\,\mathrm{d}y}_{f_Y(z-x)\,\mathrm{d}z}, \end{aligned}$$

whence (6.1) follows immediately. Convolution is a symmetric operation:

$$\big(f_X * f_Y\big)(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\,\mathrm{d}y = \big(f_Y * f_X\big)(z).$$

A convolution of three probability distributions is calculated as follows:

$$\big(f_1 * f_2 * f_3\big)(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_1(x_1) f_2(x_2) f_3(z - x_1 - x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2,$$

and generalizations of higher order are obvious.

Example What do we obtain after two consecutive convolutions of a symmetric uniform distribution $U(-1/2, 1/2)$, corresponding to the "box" probability density f shown in Fig. 6.1? The first convolution yields

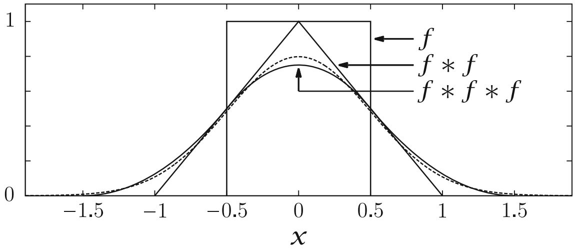


Fig. 6.1 Twofold consecutive convolution of $U(-1/2, 1/2)$ with itself

$$g(x) = \bigl(f * f\bigr)(x) = \int_{-\infty}^{\infty} f(x') f(x - x')\,\mathrm{d}x' = \begin{cases} \displaystyle\int_{-1/2}^{x+1/2} \mathrm{d}x' = 1 + x\,; & -1 \le x \le 0, \\[2ex] \displaystyle\int_{x-1/2}^{1/2} \mathrm{d}x' = 1 - x\,; & 0 \le x \le 1, \end{cases}$$

which is a triangular distribution ($f * f$ in the figure). The second convolution gives

$$\begin{aligned} \bigl(f * f * f\bigr)(x) &= \int_{-\infty}^{\infty} f(x')\,g(x - x')\,\mathrm{d}x' \\ &= \begin{cases} \displaystyle\int_{-1/2}^{x+1} [\bullet]\,\mathrm{d}x' = \frac{9}{8} + \frac{3x}{2} + \frac{x^2}{2} & ;\ -\tfrac{3}{2} \le x \le -\tfrac{1}{2}, \\[2ex] \displaystyle\int_{-1/2}^{x} [\bullet]\,\mathrm{d}x' + \int_{x}^{1/2} [\bullet]\,\mathrm{d}x' = -\frac{x^2}{2} + \frac{3}{4}\ ; & |x| \le \tfrac{1}{2}, \\[2ex] \displaystyle\int_{x-1}^{1/2} [\bullet]\,\mathrm{d}x' = \frac{9}{8} - \frac{3x}{2} + \frac{x^2}{2} & ;\ \tfrac{1}{2} \le x \le \tfrac{3}{2}, \end{cases} \end{aligned}$$

where $\bullet = 1 + (x - x')$. This density is denoted by $f * f * f$ in the figure. Try to proceed yet another step and calculate $f * f * f * f$! (You shall see in an instant where this is leading.) ◁

Example What about the convolution of an asymmetric distribution? For instance, what is the distribution of the variable $Y = X_1 + X_2 + \cdots + X_n$ if all X_i are uniformly distributed on [0, 1], i.e. $X_i \sim U(0, 1)$?

The density of the variable Y for arbitrary $n \ge 1$ is

$$f_Y(x) = \frac{1}{(n-1)!} \sum_{h=0}^{\lfloor x \rfloor} \binom{n}{h} (-1)^h (x - h)^{n-1}, \qquad n \ge 1, \tag{6.2}$$

and is shown in Fig. 6.2 for $n = 1$ (original distribution), $n = 2$ (single convolution), $n = 3$, $n = 6$ and $n = 12$. As in the previous example, the density after several convolutions reminds one of something "bell-shaped", one could suspect, the normal distribution. Besides, the distribution of the sum variable creeps away from the origin: this is a cue for the following subsection. ◁

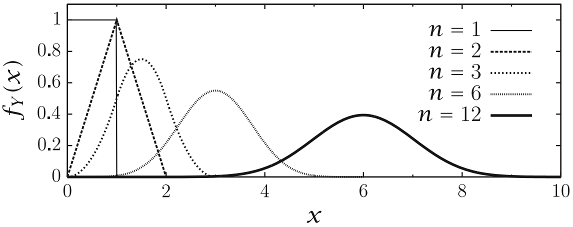


Fig. 6.2 Multiple convolutions of the $U(0, 1)$ distribution with itself

6.1.1 The Effect of Convolution on Distribution Moments

First consider what happens to the average of the sum of two random variables:

$$\begin{aligned}\overline{Z} &= \int_{-\infty}^{\infty} z f_Z(z)\,\mathrm{d}z = \int_{-\infty}^{\infty} z \left[\int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x\right] \mathrm{d}z \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} z f_Y(z-x)\,\mathrm{d}z\right] f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}(x+y) f_Y(y)\,\mathrm{d}y\right] f_X(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f_Y(y)\,\mathrm{d}y\right] f_X(x)\,\mathrm{d}x + \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y = \overline{X} + \overline{Y},\end{aligned}$$

thus

$$\overline{X+Y} = \overline{X} + \overline{Y}$$

or $E[X+Y] = E[X] + E[Y]$, which we already know from (4.6). Now let us also calculate the variance of Z! We must average the expression

$$\left(Z-\overline{Z}\right)^2 = \left[\left(X-\overline{X}\right) + \left(Y-\overline{Y}\right)\right]^2 = \left(X-\overline{X}\right)^2 + 2\left(X-\overline{X}\right)\left(Y-\overline{Y}\right) + \left(Y-\overline{Y}\right)^2.$$

Because X and Y are independent, the expected value of the second term is zero, so we are left with only

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$

or $\mathrm{var}[X+Y] = \mathrm{var}[X] + \mathrm{var}[Y]$. We know that too, namely, from (4.20), in a slightly different garb also from (4.25) if one sets $Y = X_1 + X_2$. As an exercise, check what happens to the third and fourth moment upon convolution: you will find out that $M_{3,X+Y} = M_{3,X} + M_{3,Y}$, so the third moments of distributions are additive. By taking into account the definition of skewness $\rho = M_3/\sigma^3$ (see (4.18)) this can also be written as

$$\rho_{X+Y}\sigma_{X+Y}^3 = \rho_X\sigma_X^3 + \rho_Y\sigma_Y^3.$$

The fourth moments are not additive, since $M_{4,X+Y} = M_{4,X} + M_{4,Y} + 6M_{2,X}M_{2,Y}$, but by using (4.19) this can be simplified to

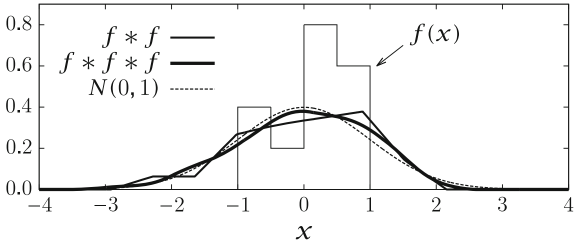


Fig. 6.3 As few as two convolutions may be needed to turn a relatively irregular distribution into a distribution that looks almost like the standardized normal. (We have subtracted the expected value of the distribution obtained at each step and rescaled the variance)

$$\varepsilon_{X+Y}\sigma_{X+Y}^4 = \varepsilon_X\sigma_X^4 + \varepsilon_Y\sigma_Y^4.$$

Example It appears as if even the most "weird" distribution evolves into something "bell-shaped" when it is convoluted with itself a couple of times. Figure 6.3 shows an example in which upon just two convolutions a rather irregular density (fulfilling all requirements for a probability density) turns into a density closely resembling the standardized normal distribution. ◁

Example Still, convolution does not perform miracles. Let us calculate the n-fold convolution of the Cauchy distribution with itself! We obtain

$$\begin{aligned}
f^{(1)}(x) &= f(x) = \frac{1}{\pi}\frac{1}{1+x^2},\\
f^{(2)}(x) &= \big(f*f\big)(x) = \frac{1}{\pi}\frac{2}{4+x^2},\\
f^{(3)}(x) &= \big(f*f*f\big)(x) = \frac{1}{\pi}\frac{3}{9+x^2},\\
&\vdots\\
f^{(n)}(x) &= \big(\underbrace{f*f*\cdots*f}_{n}\big)(x) = \frac{1}{\pi}\frac{n}{n^2+x^2}.
\end{aligned} \tag{6.3}$$

Certainly $f^{(n)}$ does not approach the density of the normal distribution; rather, it remains faithful to its ancestry. Consecutive convolutions yield just further Cauchy distributions! We say that the Cauchy distribution is *stable with respect to convolution*. The reasons for this behaviour will be discussed below. ◁

6.2 Convolution of Discrete Distributions

The discrete analog of the continuous convolution formula (6.1) for the summation of independent discrete random variables X and Y is at hand: if X takes the value i, then Y must be $n-i$ if their sum is to be n. Since X and Y are independent, the probabilities for such an "event" should be multiplied, thus

$$P(X+Y=n)=\sum_i P(X=i,Y=n-i)=\sum_i P(X=i)P(Y=n-i) \qquad (6.4)$$

or

$$f_{X+Y}(n)=\sum_i f_X(i)f_Y(n-i).$$

Example Let us demonstrate that the convolution of two Poisson distributions is still a Poisson distribution! Let $X \sim \mathrm{Poisson}(\lambda)$ and $Y \sim \mathrm{Poisson}(\mu)$ be mutually independent Poisson variables with parameters λ and μ. For their sum $Z = X + Y$ one then has

$$\begin{aligned} P(Z=n) &= \sum_{i=0}^{n} P(X=i,Y=n-i)=\sum_{i=0}^{n} P(X=i)P(Y=n-i) \\ &= \sum_{i=0}^{n} \frac{\lambda^i \mathrm{e}^{-\lambda}}{i!}\frac{\mu^{(n-i)}\mathrm{e}^{-\mu}}{(n-i)!}=\frac{\mathrm{e}^{-(\lambda+\mu)}}{n!}\sum_{i=0}^{n}\frac{n!}{i!(n-i)!}\lambda^i\mu^{n-i}=\frac{\mathrm{e}^{-(\lambda+\mu)}(\lambda+\mu)^n}{n!}, \end{aligned}$$

thus indeed $Z \sim \mathrm{Poisson}(\lambda+\mu)$. A more elegant solution of this problem will be given by the Example on p. 369 in Appendix B.1. ◁

Example Let us compute the probability distribution of the sum $Z = X + Y$ of independent discrete random variables X and Y, distributed according to

$$f_n = P(X=n) = \begin{cases} 0.15 & ;\ n=-3, \\ 0.25 & ;\ n=-1, \\ 0.1 & ;\ n=2, \\ 0.3 & ;\ n=6, \\ 0.2 & ;\ n=8, \\ 0 & ;\ \text{otherwise}, \end{cases} \qquad g_n = P(Y=n) = \begin{cases} 0.2 & ;\ n=-2, \\ 0.1 & ;\ n=1, \\ 0.3 & ;\ n=5, \\ 0.4 & ;\ n=8, \\ 0 & ;\ \text{otherwise}. \end{cases}$$

The distributions are shown in Fig. 6.4 (left) [1].

In principle we are supposed to find all values $P(Z=z)$, so we must compute the convolution sum $\{h\} = \{f\} * \{g\}$ for each n separately:

$$h_n = P(Z=n) = \sum_{j=-\infty}^{\infty} f_j g_{n-j}.$$

To make the point, let us just calculate the probability that $X + Y = 4$. We need

$$\begin{aligned} h_4 = P(Z=4) = {} & f_{-3}\cancel{g}_7 + f_{-2}\cancel{g}_6 + \underline{f_{-1}g_5} + \cancel{f}_0\cancel{g}_4 + \cancel{f}_1\cancel{g}_3 + f_2\cancel{g}_2 \\ & + \cancel{f}_3 g_1 + \cancel{f}_4\cancel{g}_0 + \cancel{f}_5\cancel{g}_{-1} + \underline{f_6 g_{-2}} + \cancel{f}_7\cancel{g}_{-3} + f_8\cancel{g}_{-4} \\ = {} & 0.25\cdot 0.3 + 0.3\cdot 0.2 = 0.135. \end{aligned}$$

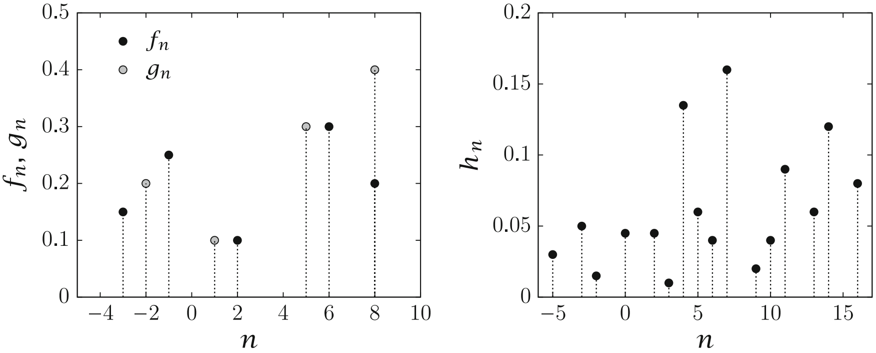


Fig. 6.4 Discrete convolution in the case when the distributions have different supports. [Left] Distributions f and g. [Right] Convolution of f and g

When n and j indices are combed through, many $f_j g_{n-j}$ terms vanish (crossed-out terms above); only the underlined bilinears survive. Such a procedure must be repeated for each n: a direct calculation of convolutions may become somewhat tedious. The problem can also be solved by using generating functions, as demonstrated by the Example on p. 374 in Appendix B.2. ◁

6.3 Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be real, independent and identically distributed random variables with probability density f_X, whose expected value $\mu_X = E[X_i]$ and variance $\sigma_X^2 = E[(X_i - \mu_X)^2]$ are bounded. Define the sum of random variables $Y_n = \sum_{i=1}^n X_i$. By (4.6) and (4.22), the expected value and variance of Y_n are $E[Y_n] = \mu_{Y_n} = n\mu_X$ and $\sigma_{Y_n}^2 = n\sigma_X^2$, respectively. The probability density f_Y of the sum variable Y_n is given by the n-fold convolution of the densities of the X_i's,

$$f_{Y_n} = \underbrace{f_X * f_X * \cdots * f_X}_{n}.$$

The example in Fig. 6.2 has revealed that the average of the probability density, calculated by consecutive convolutions of the original density, kept on increasing: in that case, the average in the limit $n \to \infty$ even diverges! One sees that the variance keeps on growing as well. Both problems can be avoided by defining a rescaled variable

$$Z_n = \frac{Y_n - \mu_{Y_n}}{\sigma_{Y_n}} = \frac{Y_n - n\mu_X}{\sqrt{n}\sigma_X}.$$

This ensures that upon subsequent convolutions, the average of the currently obtained density is subtracted and its variance is rescaled: see Fig. 6.3. In the limit $n \to \infty$ the distribution function of the variable Z_n then converges to the distribution function of the standardized normal distribution $N(0, 1)$,

$$\lim_{n\to\infty} P(Z_n \le z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-t^2/2}\,\mathrm{d}t,$$

or, in the language of probability densities,

$$\lim_{n\to\infty} \sigma_{Y_n} f_{Y_n}\left(\sigma_{Y_n} z + \mu_{Y_n}\right) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-z^2/2}.$$

In other words, the dimensionless probability density $\sigma_{Y_n} f_{Y_n}$ converges to the standardized normal probability density in the limit $n \to \infty$, which is known as the *central limit theorem* (CLT).

6.3.1 Proof of the Central Limit Theorem

The central limit theorem can be proven in many ways: one way is to exploit our knowledge on momentum-generating functions from Appendix B.2. Suppose that the momentum-generating function of the variables X_i exists and is finite for all t in some neighborhood of $t = 0$. Then for each standardized variable $U_i = (X_i - \mu_X)/\sigma_X$, for which $E[U_i] = 0$ and $\mathrm{var}[U_i] = 1$ (thus also $E[U_i^2] = 1$), there exists a corresponding momentum-generating function

$$M_{U_i}(t) = E\big[\mathrm{e}^{tU_i}\big],$$

which is the same for all U_i. Its Taylor expansion in the vicinity of $t = 0$ is

$$M_U(t) = 1 + t\,\underbrace{E[U]}_{0} + \frac{t^2}{2!}\,\underbrace{E[U^2]}_{1} + \frac{t^3}{3!}\,E[U^3] + \cdots = 1 + \frac{t^2}{2} + \mathcal{O}\big(t^2\big). \tag{6.5}$$

Let us introduce the standardized variable

$$Z_n = (U_1 + U_2 + \cdots + U_n)/\sqrt{n} = (X_1 + X_2 + \cdots + X_n - n\mu_X)/(\sigma_X\sqrt{n}).$$

Its momentum-generating function is $M_{Z_n}(t) = E\big[\mathrm{e}^{tZ_n}\big]$. Since the variables X_i are mutually independent, this also applies to the rescaled variables U_i, therefore, by formula (B.16), we get

$$E\big[\mathrm{e}^{tZ_n}\big] = E\big[\mathrm{e}^{t(U_1+U_2+\cdots+U_n)/\sqrt{n}}\big] = E\big[\mathrm{e}^{(t/\sqrt{n})U_1}\big]E\big[\mathrm{e}^{(t/\sqrt{n})U_2}\big]\cdots E\big[\mathrm{e}^{(t/\sqrt{n})U_n}\big]$$

or

$$M_{Z_n}(t) = \big[M_U\big(t/\sqrt{n}\big)\big]^n, \qquad n = 1, 2, \ldots$$

By using the expansion of M_U, truncated at second order, we get

$$M_{Z_n}(t) = \left[1 + \frac{t^2}{2n} + \mathcal{O}(t^2/n)\right]^n, \qquad n = 1, 2, \ldots$$

Hence

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty}\left(1 + \frac{t^2}{2n}\right)^n = \mathrm{e}^{t^2/2}.$$

We know from (B.13) that this is precisely the momentum-generating function corresponding to the normal distribution $N(0, 1)$, so indeed

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-z^2/2},$$

which we set out to prove. A direct proof (avoiding the use of generating functions) can be found in [2]; it proceeds along the same lines as the proof of the Laplace limit theorem in Appendix B.3.1.

The speed of convergence to the standardized normal distribution $N(0, 1)$ with the distribution function $\Phi(z)$ is quantified by the Berry–Esséen theorem [2]. If the third moment of $|X - \mu_X|$ is bounded ($\rho = E[|X - \mu_X|^3] < \infty$), it holds that

$$|\,P(Z_n \le z) - \Phi(z)\,| \le \frac{C\rho}{\sqrt{n}\sigma_X^3},$$

where $0.4097 \lesssim C \lesssim 0.4748$ [3]. Now we also realize why consecutive convolutions in (6.3) have not led us to the normal distribution: no moments exist for the Cauchy distribution (see Sect. 4.7.1), so the condition $\rho < \infty$ is not fulfilled. Moreover, one should not truncate the Taylor expansion (6.5).

The central limit theorem and the form of the bound on the speed of convergence remain valid when summing variables X_i distributed according to different (non-identical) probability distributions, if the variables are not too "dispersed" (Lindeberg criterion, see [2]). An excellent (and critical) commentary on "why normal distributions are normal" is also given by [4].

Example Let us revisit the convolution of the uniform distribution in Fig. 6.2. We sum twelve mutually independent variables $X_i \sim U(0, 1)$ and subtract 6,

$$Y = \sum_{i=1}^{12} X_i - 6. \tag{6.6}$$

What are we supposed to get? The averages of all X_i are $1/2$, $E[X_i] = 1/2$, while their variances are $\mathrm{var}[X_i] = 1/12$ (see Table 4.1). Hence, Y should also have an average of zero and a variance of $\mathrm{var}[Y] = \mathrm{var}[X_1] + \cdots + \mathrm{var}[X_{12}] = 12/12 = 1$. By the central limit theorem, Y should be almost normally distributed, if we believe that $12 \approx \infty$. How well this holds is shown in Fig. 6.5.

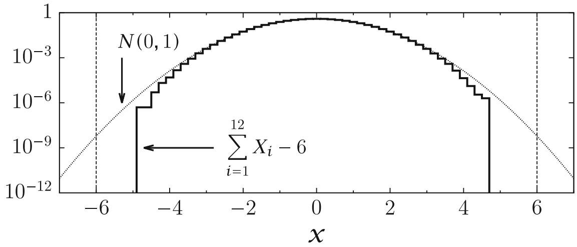


Fig. 6.5 Histogram of 10^7 values Y, randomly generated according to (6.6), compared to the density of the standardized normal distribution (3.10). In effect, the figure also shows the deviation of (6.2) from the normal density. The sharp cut-offs at ≈−5 and ≈4.7 are random: by drawing a larger number of values the histogram would fill the whole interval [−6, 6]

We have thus created a primitive "convolution" generator of approximately normally distributed numbers, but with its tails cut off since Y can never exceed 6 and can never drop below −6. It is a practical generator—which does not mean that it is good. How a "decent" generator of normally distributed random numbers can be devised will be discussed in Sect. C.2.5. ◁

Example (Adapted from [5].) The mass M of granules of a pharmaceutical ingredient is a random variable, distributed according to the probability density

$$f_M(m) = \frac{1}{24m_0^5}\, m^4 \mathrm{e}^{-m/m_0}, \quad m \geq 0, \quad m_0 = 40\,\mathrm{mg}. \tag{6.7}$$

To analyze the granulate, we acquire a sample of 30 granules. What is the probability that the total mass of the granules in the sample exceeds its average value by more than 10%?

The average mass of a single granule and its variance are

$$\overline{M} = \int_0^\infty m f_M(m)\,\mathrm{d}m = 5m_0, \qquad \sigma_M^2 = \int_0^\infty \left(m - \overline{M}^2\right) f_M(m)\,\mathrm{d}m = 5m_0^2.$$

The probability density f_X of the total sample mass X, which is also a random variable, is a convolution of thirty densities of the form (6.7); this number is large enough to invoke the central limit theorem, so the density f_X is almost normal, with average $\overline{X} = 30\,\overline{M} = 150\,m_0$ and variance $\sigma_X^2 = 30\,\sigma_M^2 = 150\,m_0^2$:

$$f_X(x) \approx f_{\mathrm{norm}}\left(x; \overline{X}, \sigma_X^2\right) = f_{\mathrm{norm}}\left(x; 150\,m_0, 150\,m_0^2\right).$$

The desired probability is then

$$P(X > 165\,m_0) \approx \int\limits_{165\,m_0}^{\infty} f_{\mathrm{norm}}\left(x; \overline{X}, \sigma_X^2\right)\mathrm{d}x = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{(165-150)m_0}{\sqrt{2}\sqrt{150}\,m_0}\right)\right] \approx 11\%,$$

where we have used Table D.2. ◁

6.4 Stable Distributions ⋆

The normal distribution as the limit distribution of the sum of independent random variables can be generalized by the concept of *stable distributions* [6, 7].

Suppose we are dealing with independent random variables X_1, X_2 and X_3 with the same distribution over the sample space Ω. We say that such a distribution is *stable*, if for each pair of numbers a and b there exists a pair c and d such that the distribution of the linear combination $aX_1 + bX_2$ is equal to the distribution of $cX_3 + d$, that is,

$$P\big(aX_1 + bX_2 \in A\big) = P\big(cX_3 + d \in A\big) \qquad \forall A \subset \Omega .$$

Such random variables are also called 'stable'; a superposition of stable random variables is a linear function of a stable random variable with the same distribution.

Stable distributions are most commonly described by their characteristic functions (see Appendix B.3). Among many possible notations we follow [6]. We say that a random variable X has a stable distribution $f_{\mathrm{stab}}(x; \alpha, \beta, \gamma, \delta)$, if the logarithm of its characteristic function (B.17) has the form

$$\log \phi_X(t) = \mathrm{i}\delta t - \gamma^\alpha |t|^\alpha \big[1 - \mathrm{i}\beta\Phi_\alpha(t)\big],$$

where

$$\Phi_\alpha(t) = \begin{cases} \mathrm{sign}(t)\ \tan(\pi\alpha/2) \ ; \ \alpha \neq 1, \\ -\frac{2}{\pi}\,\mathrm{sign}(t) \log |t| \ ; \ \alpha = 1. \end{cases}$$

The parameter $\alpha \in (0, 2]$ is the *stability index* or *characteristic exponent*, the parameter $\beta \in [-1, 1]$ describes the skewness of the distribution, and two further parameters $\gamma > 0$ and $\delta \in \mathbb{R}$ correspond to the distribution scale and location, respectively. For $\alpha \in (1, 2]$ the expected value exists and is equal to $E[X] = \delta$. For general $\alpha \in (0, 2]$ there exist moments $E[|X|^p]$, where $p \in [0, \alpha)$.

It is convenient to express X by another random variable Z,

$$X = \begin{cases} \gamma Z + \delta & ; \ \alpha \neq 1, \\ \gamma\left(Z + \frac{2}{\pi}\beta \log \gamma\right) + \delta \ ; & \alpha = 1. \end{cases}$$

Namely, the characteristic function of Z is somewhat simpler,

$$\log \phi_Z(t) = -|t|^\alpha \big[1 - \mathrm{i}\beta\Phi_\alpha(t)\big],$$

as it depends only on two parameters, α and β. The probability density f_Z of the variable Z is calculated by the inverse Fourier transformation of the characteristic function ϕ_Z:

$$f_Z(z; \alpha, \beta) = \frac{1}{\pi}\int_0^\infty \exp(-t^\alpha) \cos\big(zt - t^\alpha \beta \Phi_\alpha(t)\big)\, \mathrm{d}t,$$

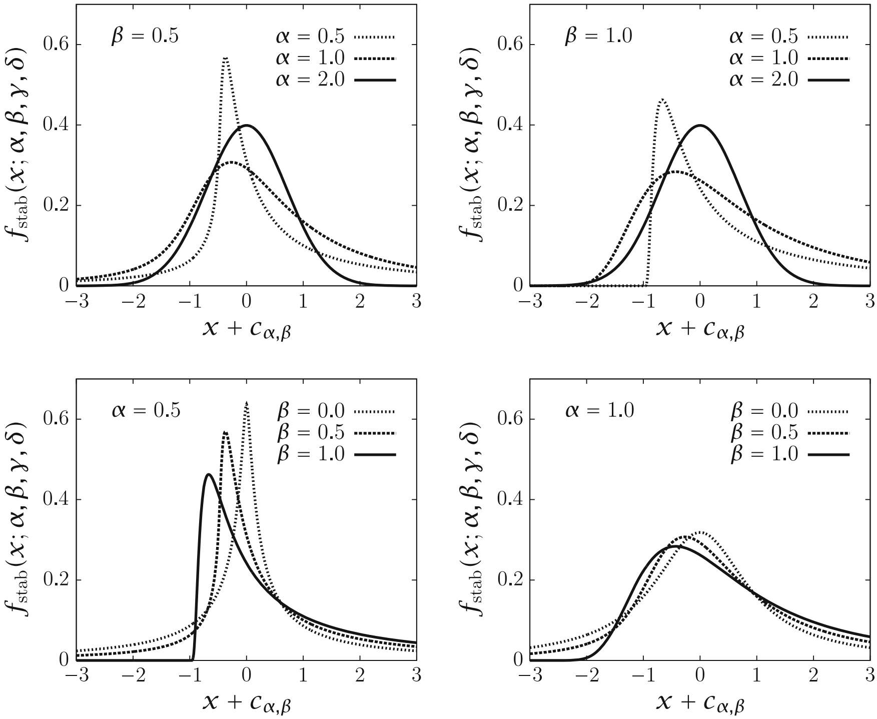


Fig. 6.6 Stable distributions $f_{\mathrm{stab}}(x;\, \alpha,\, \beta,\, \gamma,\, \delta)$. [Top left and right] Dependence on parameter α at $\beta = 0.5$ and 1.0. [Bottom left and right] Dependence on parameter β at $\alpha = 0.5$ and 1.0. At $\alpha \neq 1$ the independent variable is shifted by $c_{\alpha,\beta} = \beta \tan(\pi\alpha/2)$

where $f_Z(-z;\, \alpha,\, \beta) = f_Z(z;\, \alpha,\, -\beta)$. The values of f_Z and f_X can be computed by using integrators tailored to rapidly oscillating integrands: see [8], p. 660; a modest software support for stable distributions can also be found in [9]. With respect to α and β, the definition domains of f_Z are

$$z \in \begin{cases} (-\infty, 0] & ;\ \alpha < 1,\ \ \beta = -1, \\ [0, \infty) & ;\ \alpha < 1,\ \ \beta = 1, \\ \mathbb{R} & ;\ \text{otherwise}. \end{cases}$$

The dependence of f_{stab} (f_X or f_Z with appropriate scaling) on the parameter α is shown in Fig. 6.6 (top left and right), while the dependence on β is shown in the same figure at bottom left and right.

By a suitable choice of parameters such a general formulation allows for all possible stable distributions. The most relevant ones are

$$\begin{aligned}
\text{normal}: &\ \alpha = 2,\ \beta = 0,\ f_X(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}, && x \in \mathbb{R}\,;\\
\text{Cauchy}: &\ \alpha = 1,\ \beta = 0,\ f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}, && x \in \mathbb{R}\,;\\
\text{Lévy}: &\ \alpha = \tfrac{1}{2},\ \beta = 1,\ f_X(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-1/(2x)}\,x^{-3/2}, && x \in \mathbb{R}_+.
\end{aligned}$$

Stable distributions with $\alpha \in (0, 2)$ have a characteristic behaviour of probability densities known as *power* or *fat tails*. The cumulative probabilities satisfy the asymptotic relations

$$\begin{aligned}
\beta \in (-1, 1]: &\ \int_x^\infty f_Z(z; \alpha, \beta)\,\mathrm{d}z \sim \frac{1}{2}\,c_\alpha(1+\beta)x^{-\alpha}, && x \to \infty,\\
\beta \in [-1, 1): &\ \int_{-\infty}^x f_Z(z; \alpha, \beta)\,\mathrm{d}z \sim \frac{1}{2}\,c_\alpha(1-\beta)(-x)^{-\alpha}, && x \to -\infty,
\end{aligned} \tag{6.8}$$

where $c_\alpha = 2\sin(\pi\alpha/2)\Gamma(\alpha)/\pi$. For $\beta \in (-1, 1)$ such asymptotic behaviour is valid in both limits, $x \to \pm\infty$. Note that the probability density has the asymptotics $\mathcal{O}(|x|^{-\alpha-1})$ if the cumulative probability goes as $\mathcal{O}(|x|^{-\alpha})$.

6.5 Generalized Central Limit Theorem ⋆

Having introduced stable distributions (Sect. 6.4) one can formulate the generalized central (or Lévy's) limit theorem, elaborated more closely in [2]. Here we just convey its essence.

Suppose we have a sequence of independent, identically distributed random variables $\{X_i\}_{i\in\mathbb{N}}$, from which we form the partial sum

$$Y_n = X_1 + X_2 + \cdots + X_n.$$

Assume that their distribution has power tails, so that for $\alpha \in (0, 2]$ the following limits exist:

$$\lim_{x\to\infty} |x|^\alpha\, P(X > x) = d_+, \qquad \lim_{x\to-\infty} |x|^\alpha\, P(X < x) = d_-,$$

and $d = d_+ + d_- > 0$. Then real coefficients $a_n > 0$ and b_n exist such that the rescaled partial sum

$$Z_n = \frac{Y_n - nb_n}{a_n}$$

in the limit $n \to \infty$ is stable, and its probability density is $f_{\mathrm{stab}}(x; \alpha, \beta, 1, 0)$. Its skewness is given by $\beta = (d_+ - d_-)/(d_+ + d_-)$, while a_n and b_n are

$$a_n = \begin{cases} (d\,n/c_\alpha)^{1/\alpha} & ;\ \alpha \in (0,2), \\ \sqrt{(d\,n \log n)/2} & ;\ \alpha = 2, \end{cases}$$

$$b_n = \begin{cases} E\big[X_i\big] & ;\ \alpha \in (1,2], \\ E\big[X_i H\big(|X_i| - a_n\big)\big] & ;\ \text{otherwise}, \end{cases}$$

where H is the Heaviside function. The constant c_α is defined next to (6.8). The coefficient a_n for $\alpha < 2$ diverges with increasing n as $\mathcal{O}(n^{1/\alpha})$.

The generalized central limit theorem is useful in analyzing the process of random walk, which is analogous to extending the partial sum of random numbers Y_n. Such processes are discussed in Sects. 6.7 and 6.8. The convergence to the stable distribution when $n \to \infty$ is becoming more and more "capricious" when α decreases.

6.6 Extreme-Value Distributions ⋆

In Sects. 6.3 and 6.4 we have discussed the distributions of values obtained in summing independent, identically distributed random variables $\{X_i\}_{i=1}^n$. Now we are interested in statistical properties of their *maximal* and *minimal* values, i.e. the behaviour of the quantities

$$\begin{aligned} M_n &= \max\{X_1, X_2, \ldots, X_n\}, \\ \widetilde{M}_n &= \min\{X_1, X_2, \ldots, X_n\}, \end{aligned}$$

when $n \to \infty$. We thereby learn something about the probability of extreme events, as exceptionally strong earthquakes, unusual extent of floods or inconceivably large amounts of precipitation: "*It rained for four years, eleven months, and two days.*" (See [10], p. 315.) The variables X_i are the values of the process, usually recorded at constant time intervals—for example, $n = 365$ daily temperature averages on Mt. Blanc—while M_n is the corresponding annual maximum. We are interested in, say, the probability that on top of Mt. Blanc, the temperature of $+42\,^\circ\mathrm{C}$ will be exceeded on any one day in the next ten years.

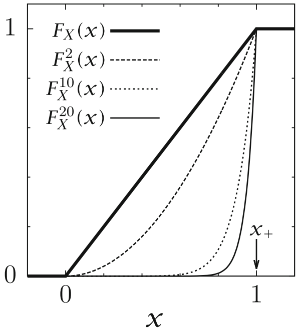
In principle, we have already answered these questions—about both the maximal and minimal value—in Problem 2.11.6: if F_X is the distribution function of the individual X_i's, the maximal values M_n are distributed according to

$$F_{M_n}(x) = P\big(M_n \le x\big) = \big[F_X(x)\big]^n, \tag{6.9}$$

and the minimal as

$$1 - F_{\widetilde{M}_n}(x) = 1 - P\big(\widetilde{M}_n \le x\big) = P\big(\widetilde{M}_n > x\big) = \big[1 - F_X(x)\big]^n.$$

But this does not help much, as F_X is usually not known! A statistical analysis of the observations may result in an approximate functional form of F_X, but even small errors in F_X (particularly in its tails) may imply large deviations in F_X^n. We therefore accept the fact that F_X is unknown and try to find families of functions F_X^n, by which extreme data can be modeled directly [11, 12].



There is another problem. Define x_+ as the smallest value x, for which $F_X(x) = 1$. Then for any $x < x_+$ we get $F_X^n(x) \to 0$, when $n \to \infty$, so that the distribution function of M_n degenerates into a "step" at x_+. The figure above shows this in the case of uniformly distributed variables $X_i \sim U(0, 1)$ with probability density $f_X(x) = 1$ $(0 \le x \le 1)$ and distribution function $F_X(x) = x$ $(0 \le x \le x_+ = 1)$. When $n \to \infty$, the distribution function F_X^n tends to the step (Heaviside) function at $x = 1$, while its derivative (probability density) resembles the delta "function" at the same point. Our goal is to find a non-degenerate distribution function. We will show that this can be accomplished by a rescaling of the variable M_n,

$$M_n^* = \frac{M_n - b_n}{a_n}, \tag{6.10}$$

where $a_n > 0$ and b_n are constants. Illustrations of a suitable choice of these constants or of their calculation are given by the following Example and Exercise in Sect. 6.9.5. A general method to determine these constants is discussed in [13, 14].

Example Let $X_1, X_2, \ldots, X_n$ be a sequence of independent, exponentially distributed variables, thus $F_X(x) = 1 - \mathrm{e}^{-x}$. Let $a_n = 1$ and $b_n = \log n$. Then

$$\begin{aligned} P\left(\frac{M_n - b_n}{a_n} \le x\right) &= P\big(M_n \le a_n x + b_n\big) = P\big(M_n \le x + \log n\big) \\ &= \big[F_X\big(x + \log n\big)\big]^n = \big[1 - \mathrm{e}^{-(x+\log n)}\big]^n = \left[1 - \frac{1}{n}\,\mathrm{e}^{-x}\right]^n \\ &\to \exp\big(-\exp(-x)\big), \qquad x \in \mathbb{R}, \end{aligned}$$

when $n \to \infty$. By a suitable choice of a_n and b_n we have therefore stabilized the location and scale of the distributions of M_n^* in the limit $n \to \infty$.

Let us repeat this calculation for independent variables with the distribution function $F_X(x) = \mathrm{e}^{-1/x}$ and for uniformly distributed variables, $F_X(x) = x$! In the first case we set $a_n = n$ and $b_n = 0$, and get $P(M_n^* \le x) = \mathrm{e}^{-1/x}$ $(x > 0)$. In the second case a good choice is $a_n = 1/n$ and $b_n = 1$, yielding $P(M_n^* \le x) \to \mathrm{e}^x$ $(x < 0)$ in the limit $n \to \infty$. Plot all three functions $F_X(x)$ of this Example and elaborate why one or the other are more or less sensible for the actual physical description of extreme phenomena! ◁

6.6.1 Fisher–Tippett–Gnedenko Theorem

Apparently the choice of constants a_n and b_n is crucial if we wish the distribution of M_n^* in the limit $n \to \infty$ to be non-trivial (not degenerated into a point); the basic formalism for a correct determination of these constants is discussed e.g. in [14]. In the following we assume that such constants can be found; one can then namely invoke the Fisher–Tippett–Gnedenko theorem [15, 16], which is the extreme-value analog of the central limit theorem of Sect. 6.3: if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that in the limit $n \to \infty$ we have

$$P\left(\frac{M_n - b_n}{a_n} \le x\right) \to G(x)$$

for a non-degenerate distribution function G, then G belongs to the family

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \qquad (6.11)$$

defined on the set of points $\{x : 1 + \xi(x-\mu)/\sigma) > 0\}$, where $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$. Formula (6.11) defines the family of *generalized extreme-value distributions* (GEV). An individual distribution is described by the location parameter μ (a sort of average of extreme values), the scale parameter σ (their dispersion), and the shape parameter ξ. The value of ξ characterizes three sub-families of the GEV set—Fréchet $(\xi > 0)$, Gumbel $(\xi = 0)$ and Weibull $(\xi < 0)$—differing by the location of the point x_+ and asymptotics. The Gumbel-type distributions must be understood as the $\xi \to 0$ limit of (6.11):

$$G(x) = \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}, \qquad -\infty < x < \infty. \qquad (6.12)$$

The corresponding probability density in the case $\xi \ne 0$ is

$$g(x) = G'(x) = \frac{1}{\sigma}\big[t(x)\big]^{1+\xi}\,\mathrm{e}^{-t(x)}, \qquad t(x) = \left[1 + \xi\,\frac{x-\mu}{\sigma}\right]^{-1/\xi} \qquad (6.13)$$

while for $\xi = 0$ it is

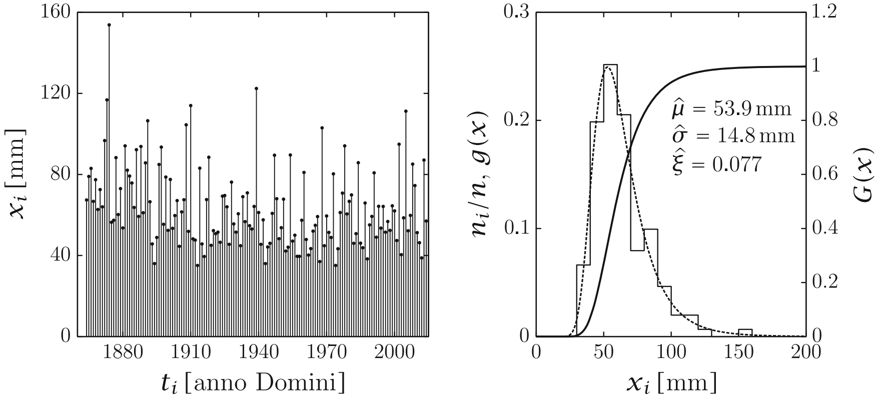


Fig. 6.7 Rainfall in Engelberg (1864–2014). [Left] Time series of annual extreme values. [Right] Histogram of extremes, the corresponding probability density g (*dashed curve*) and the GEV distribution function (*full curve*). The optimal parameters $\widehat{\mu}$, $\widehat{\sigma}$ and $\widehat{\xi}$ have been determined by fitting g to the histogram

$$g(x) = \frac{1}{\sigma}\exp\left[-\frac{x-\mu}{\sigma} - \exp\left(-\frac{x-\mu}{\sigma}\right)\right].$$

The predictive power of the Fisher–Tippett–Gnedenko theorem does not lag behind the one of the central limit theorem: if one is able to find suitable $\{a_n\}$ and $\{b_n\}$, the limiting extreme-value distribution is always of the type (6.11), *regardless of the parent distribution* F_X that generated these extreme values in the first place! Different choices of $\{a_n\}$ and $\{b_n\}$ lead to GEV-type distributions with different μ and σ, but with the same shape parameter ξ, which is the essential parameter of the distribution.

Example Figure 6.7 (left) shows the annual rainfall maxima, measured over 151 years (1864–2014) in the Swiss town of Engelberg [17]. Each data point represents the extreme one-day total (the wettest day in the year): we are therefore already looking at the extreme values and we are interested in *their* distribution, not the distribution of *all* non-zero daily quantities: *that* is most likely normal!

Figure 6.7 (right) shows the histogram of 151 extreme one-day totals, normalized such that the sum over all bins is equal to one. It can therefore be directly fitted by the density (6.13) (dashed curve), resulting in the distribution parameters $\widehat{\mu} = 53.9$ mm, $\widehat{\sigma} = 14.8$ mm, $\widehat{\xi} = 0.077$ (Fréchet family). The corresponding distribution function is shown by the full curve. ◁

6.6.2 Return Values and Return Periods

The extreme-value distribution and its asymptotic behaviour can be nicely illustrated by a *return-level plot*. Suppose that we have measured $n = 365$ daily rainfall amounts

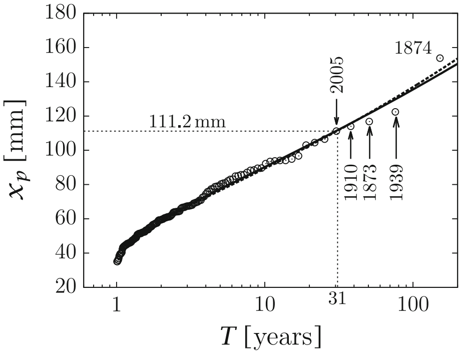


Fig. 6.8 Return values for extreme rainfall in Engelberg (period 1864–2014). The *full curve* is the model prediction with parameters from Fig. 6.7, and the *dashed curve* is the model with parameters obtained by the maximum likelihood method

x_i over a period of N consecutive years, so that their annual maxima are also available:

$$\underbrace{x_1, x_2, \ldots, x_n}_{M_{n,1}}, \ \underbrace{x_{n+1}, x_{n+2}, \ldots, x_{2n}}_{M_{n,2}}, \ \ldots, \ \underbrace{x_{(N-1)n+1}, x_{(N-1)n+2}, \ldots, x_{Nn}}_{M_{n,N}} .$$

The quantiles of the annual extremes distribution are obtained by inverting (6.11):

$$x_p = \begin{cases} \mu - \dfrac{\sigma}{\xi}\left[1 - \bigl(-\log(1-p)\bigr)^{-\xi}\right] & ; \ \xi \neq 0, \\ \mu - \sigma \log\bigl(-\log(1-p)\bigr) & ; \ \xi = 0, \end{cases}$$

where $G(x_p) = 1 - p$. We call x_p the *return level* corresponding to the *return period* $T = 1/p$. One may namely expect that the value x_p will be exceeded once every $1/p$ years or that the annual maximum will exceed the value x_p in any year with a probability of $p = 1/T$. From these definitions it follows that

$$T = \frac{1}{p} = \frac{1}{1 - G(x_p)}. \qquad (6.14)$$

The model dependence of x_p on T in the case of Engelberg rainfall is shown in Fig. 6.8 by the full curve. On the abscissa one usually uses a logarithmic scale; one thereby shrinks the region of "extreme extreme" values and obtains a clearer picture of the asymptotics in terms of ξ. We must also plot the actually measured extreme observations $M_{n,1}, M_{n,2}, \ldots, M_{n,N}$. In general, these are not sorted, so—in the spirit of (6.14)—individual extremes $M_{n,i}$ are mapped to their return periods:

$$T_i = \frac{N}{N + 1 - \mathrm{rank}(M_{n,i})}, \qquad i = 1, 2, \ldots, N.$$

The points $(T_i, M_{n,i})$ are denoted by circles in the figure. The maximum one-day total of 111.2 mm, recorded in 2005, has an expected return period of 31 years, while the deluge witnessed in 1874 may reoccur every ≈150 years on the average.

The fitting of the probability density to the data as in the previous Example depends on the number of bins in the histogram (see Sect. 9.2), so this is not the best way to pin down the optimal parameters. In Problem 8.8.3 the parameters of the GEV distribution and their uncertainties will be determined for the same data set by the method of maximum likelihood. In Fig. 6.7 (right) this distribution is shown by the dashed line.

6.6.3 Asymptotics of Minimal Values

So far we have only discussed the distributions of maximal values, most frequently occurring in practice. On the other hand, the distributions of extremely *small* values, i.e. the asymptotic behaviour of the quantities

$$\widetilde{M}_n = \min\{X_1, X_2, \ldots, X_n\}$$

when $n \to \infty$, are also important, in particular in modeling critical errors in systems, where the lifetime of the whole system, $\widetilde{M}_n$, is equal to the minimal lifetime of one of its components $\{X_i\}$.

There is no need to derive new formulas for minimal values; we can simply use the maximal-value results. Define $Y_i = -X_i$ for $i = 1, 2, \ldots, n$, so that small values of X_i correspond to large values of Y_i. Thus if $\widetilde{M}_n = \min\{X_1, X_2, \ldots, X_n\}$ and $M_n = \max\{Y_1, Y_2, \ldots, Y_n\}$, we also have

$$\widetilde{M}_n = -M_n.$$

In the limit $n \to \infty$ we therefore obtain

$$\begin{aligned} P\big(\widetilde{M}_n \le x\big) &= P\big(-M_n \le x\big) = P\big(M_n \ge -x\big) = 1 - P\big(M_n \le -x\big) \\ &\to 1 - \exp\left\{-\left[1 + \xi\left(\frac{-x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \\ &= 1 - \exp\left\{-\left[1 - \xi\left(\frac{x-\widetilde{\mu}}{\sigma}\right)\right]^{-1/\xi}\right\} \qquad (6.15) \end{aligned}$$

on $\{x : 1 - \xi(x - \widetilde{\mu})/\sigma > 0\}$, where $\widetilde{\mu} = -\mu$. This means that a minimal-value distribution can be modeled either by directly fitting (6.15) to the observations or by using (6.11) and considering the symmetry exposed above: if, for example, we wish to model the data $x_1, x_2, \ldots, x_n$ by a minimal-value distribution (parameters $\widetilde{\mu}$, σ, ξ), this is equivalent to modeling the data $-x_1, -x_2, \ldots, -x_n$ by a maximal-value distribution with the same σ and ξ, but with $\widetilde{\mu} = -\mu$.

6.7 Discrete-Time Random Walks ⋆

Random walks are non-stationary *random processes* used to model a variety of physical processes. A random or *stochastic* process is a generalization of the concept of a random variable: instead of drawing a single value, one "draws" a whole time series (signal), representing one possible *realization* of the random process or its *sample path*. The non-stationarity of the process means that its statistical properties change with time. (A detailed classification of random process can be found in [8].) In this subsection we discuss discrete-time random walks [2, 18, 19], while the next subsection is devoted to their continuous-time counterparts [18–21]. See also Chap. 12.

Imagine a discrete-time random process X, observed as a sequence of random variables $\{X(t)\}_{t\in\mathbb{N}}$. The partial sums of this sequence are

$$Y(t) = Y(0) + \sum_{i=1}^{t} X(i) = Y(t-1) + X(t) \tag{6.16}$$

and represent a new discrete-time random process Y, i.e. a sequence of random variables $\{Y(t)\}_{t\in\mathbb{N}_0}$. The process Y is called a *random walk*, whose individual step is the process $X(t)$. Let the sample space Ω of X and Y be continuous. We are interested in the time evolution of the probability density $f_{Y(t)}$ of the random variable Y if the initial density $f_{Y(0)}$ is known.

If we assume that Y is a process in which the state of each point depends only on the state of the previous point, the time evolution of $f_{Y(t)}$ is determined by

$$f_{Y(t)}(y) = \int_\Omega f\big(Y(t) = y \,|\, Y(t-1) = x\big) f_{Y(t-1)}(x)\,\mathrm{d}x,$$

where $f\big(Y(t) = y \,|\, Y(t-1) = x\big)$ is the conditional probability density that Y goes from value x at time $t-1$ to value y at time t. Let us also assume that the process X is independent of the previous states, so that $f\big(X(t) = x \,|\, Y(t-1) = y - x\big) = f_{X(t)}(x)$. By considering (6.16) and substituting $z = y - x$ we get

$$f_{Y(t)}(y) = \int_\Omega f_{X(t)}(z) f_{Y(t-1)}(y-z)\,\mathrm{d}z = \big(f_{X(t)} * f_{Y(t-1)}\big)(y).$$

By using this formula $f_{Y(t)}$ can be expressed as a convolution of the initial distribution $f_{Y(0)}$ and the distribution of the sum of steps until time t, f_X^{*t}:

$$f_{Y(t)} = f_{Y(0)} * f_X^{*t}, \qquad f_X^{*t} = f_{X(1)} * f_{X(2)} * \cdots * f_{X(t)}.$$

The time evolution $f_{Y(t)}(y)$ is most easily realized in Fourier space, where it is given by the product of Fourier transforms $\mathcal{F}$ of the probability densities,

$$\mathcal{F}\left[f_{Y(t)}\right] = \mathcal{F}\left[f_{Y(0)}\right] \prod_{i=1}^{t} \mathcal{F}\left[f_{X(i)}\right].$$

One often assumes that at time zero the value of the process Y is zero and that $f_{Y(0)}(y) = \delta(y)$. This assumption is useful in particular when one is interested in the qualitative behaviour of $f_{Y(t)}$ at long times.

6.7.1 Asymptotics

To understand the time asymptotics of the distribution $f_{Y(t)}$ is is sufficient to study one-dimensional random walks. Assume that the steps are identically distributed, with the density $f_{X(t)} = f_X$, and $Y(0) = 0$. The distribution corresponding to the process Y is therefore determined by the formula

$$f_{Y(t)} = \mathcal{F}^{-1}\left[\left(\mathcal{F}\left[f_X\right]\right)^t\right]$$

for all times t. The behaviour of $f_{Y(t)}$ in the limit $t \to \infty$ is determined by the central limit theorem (Sect. 6.3) and its generalization (Sect. 6.5). The theorems tell us that at large t, $f_{Y(t)}$ converges to the *limiting* (or *asymptotic*) distribution which can be expressed by one of the stable distributions f_{stab}, such that

$$f_{Y(t)}(y) \sim L(t) f_{\mathrm{stab}}\big(L(t)y + t\mu(t)\big)$$

with suitably chosen functions L and μ. The function L represents the effective width of the central part of the distribution $f_{Y(t)}$, where the bulk of the probability is concentrated, and is called the *characteristic spatial scale* of the distribution. The function μ has the role of the distribution average.

Furthermore, if the distribution of steps, f_X, has a bounded variance, $\sigma_X^2 < \infty$, the central limit theorem tells us that $f_{Y(t)}$ tends to the normal distribution with a standard deviation of

$$L = \sigma_{Y(t)} \sim t^{1/2}.$$

Such asymptotic dependence of the spatial scale on time defines *normal diffusion*, and this regime of random walks is named accordingly (Fig. 6.9 (left)).

If the probability density f_X asymptotically behaves as

$$f_X(x) \sim \frac{C_\pm}{|x|^{\alpha+1}}, \qquad x \to \pm\infty,$$

where $C_\pm$ are constants, we say that the distribution has *power* or *fat tails*, a concept familiar from Sect. 6.4. For $\alpha \in (0, 2)$, the second moment of the distribution no longer exists, and $f_{Y(t)}$ at large t tends to a distribution with scale

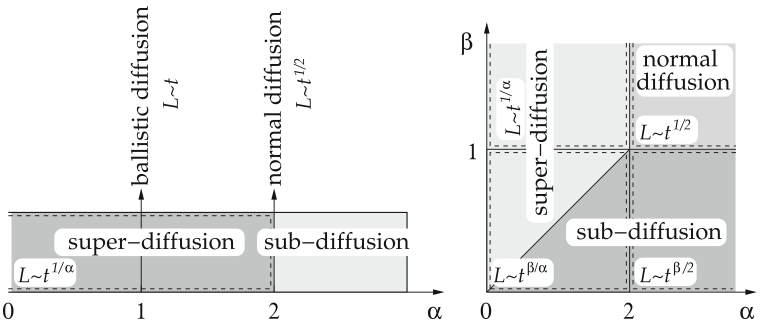


Fig. 6.9 Dependence of the characteristic spatial scale L on time t. [Left] Discrete-time random walks. [Right] Continuous-time random walks

$$L \sim t^{1/\alpha}.$$

Because in this case the characteristic scale changes faster than in the case of normal diffusion, we are referring to *super-diffusion*. The dynamics of the process Y in this regime is known as *Lévy flights*. The diffusion with $\alpha = 1$ is called *ballistic*: particles propagate with no restrictions with given velocities, so that

$$L \sim t.$$

Near $\alpha = 2$ we have $L(t) \sim (n \log n)^{1/2}$, a regime called *log-normal diffusion*.

These properties can be easily generalized to multiple space dimensions. We observe the projection of the walk, $\hat{n}^{\mathrm{T}} Y(t)$, along the direction defined by the unit vector $\hat{n}$, and its probability density $f_{\hat{n}^{\mathrm{T}} Y(t)}$. For each $\hat{n}$ we apply the central limit theorem or its generalization and determine the scale $L_{\hat{n}}$. A random walk possesses a particular direction $\hat{n}^*$ along which the scale is largest or increases fastest with time. We may take $L_{\hat{n}^*}$ to be the characteristic scale of the distribution $f_{Y(t)}$. An example of a simulation of a two-dimensional random walk where the steps in x and y directions are independent, is shown in Fig. 6.10.

If the densities $f_{X(t)}$ have power tails, $f_{Y(t)}$ also has them. This applies regardless of the central limit theorem or its generalization. Suppose that in the limit $t \to \infty$ we have $f_{X(t)}(x) \sim C_{\pm,t}|x|^{-\alpha-1}$. When the walk "generates" the density $f_{Y(t)}$, the tails add up, so $f_{Y(t)}(x) \sim \left(\sum_{i=1}^{t} C_{\pm,i}\right)|x|^{-\alpha-1}$ when $x \to \pm\infty$. This means that the probability of extreme events in $Y(t)$ increases with time, since

$$P\big(|Y(t)| > y\big) \sim \sum_{i=1}^{t} P\big(|X(i)| > y\big), \qquad \text{when } y \to \infty.$$

To estimate the variance of such processes we therefore apply methods of robust statistics (Sect. 7.4). Instead of calculating the standard deviation $\sigma_{Y(t)}$ in sub-diffusive random walks, for example, one is better off using MAD (7.23).

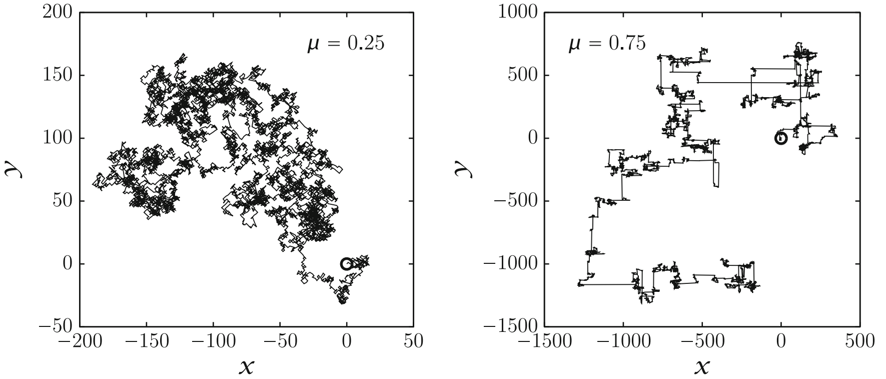


Fig. 6.10 Examples of random walks (x_t, y_t) with 10^4 steps, generated according to $x_{t+1} = x_t + \mathrm{sign}(X)|X|^{-\mu}$ and $y_{t+1} = y_t + \mathrm{sign}(Y)|Y|^{-\mu}$, where X and Y are independent random variables, uniformly distributed on $[-1, 1]$. [Left] $\mu = 0.25$. [Right] $\mu = 0.75$. The *circles* denote the initial position of the walks, $x = y = 0$

6.8 Continuous-Time Random Walks ⋆

In continuous-time random walks [18–21] the number of steps $N(t)$ taken until time t becomes a continuous random variable. The definition of a discrete-time random walk (6.16) should therefore be rewritten as

$$Y(t) = Y(0) + \sum_{i=1}^{N(t)} X(i).$$

The expression for $Y(t)$ can not be cast in iterative form $Y(t) = Y(t-1) + \cdots$ as in (6.16). The number of steps $N(t)$ has a probability distribution $F_{N(t)}$. Suppose that $N(t)$ and $X(i)$ are independent processes—which is not always true, as it is not possible to take arbitrary many steps within given time [22, 23]. If $X(i)$ at different times are independent and correspond to probability densities $f_{X(i)}$, the probability density of the random variable $Y(t)$ is

$$f_{Y(t)}(y) = \sum_{n=0}^{\infty} F_{N(t)}(n) \big(f_{Y(0)} * f_X^{*n} \big)(y),$$

where

$$f_X^{*t} = f_{X(1)} * f_{X(2)} * \cdots * f_{X(t)}.$$

In the interpretation of such random walks and the choice of distribution $F_{N(t)}$ we follow [20]. A walk is envisioned as a sequence of steps whose lengths $X(i)$ and *waiting time* $T(i)$ between the steps are randomly drawn. After N steps the walk makes it to the point $\mathcal{X}(N)$ and the elapsed time is $\mathcal{T}(N)$, so that

$$\mathcal{X}(N) = \sum_{i=1}^{N} X(i), \quad \mathcal{T}(N) = \sum_{i=1}^{N} T(i), \quad \mathcal{X}(0) = \mathcal{T}(0) = 0.$$

Within given time, a specific point can be reached in different numbers of steps N. If the step lengths $X(i)$ and waiting times $T(i)$ are independent, the number of steps $N(t)$ taken until time t is determined by the process of drawing the waiting times. Let us introduce the probability that the ith step does not occur before time t,

$$F_{T(i)}(t) = \int_t^\infty f_{T(i)}(t')\,\mathrm{d}t',$$

where $f_{T(i)}$ is the probability density corresponding to the distribution of waiting times. The probability of making n steps within the time interval $[0, t]$ is then

$$F_{N(t)}(n) = \int_0^t f_T^{*n}(t') F_{T(n+1)}(t - t')\,\mathrm{d}t' = \left(f_T^{*n} * F_{T(n+1)}\right)(t),$$

where

$$f_T^{*n} = f_{T(1)} * f_{T(2)} * \cdots * f_{T(n)}.$$

The distribution $F_{N(t)}$ can be calculated by using the Laplace transformation in time domain and the Fourier transformation in spatial domain: this allows one to avoid convolutions and operate with products of functions in transform spaces. The procedure, which we can not discuss in detail, leads to the Montroll–Weiss equation [20], helping us to identify four parameter regions corresponding to distributions of step lengths (density f_X) and waiting times (density f_T) with different dependencies of the scale L on time t, which determine the diffusion properties of the random walk. These regions are shown in Fig. 6.9 (right) and quantified below. We assume that the distributions of step lengths and waiting times do not change during the walk, so that $f_{X(i)} = f_X$ and $f_{T(i)} = f_T$.

Normal diffusion with spatial scale $L \sim t^{1/2}$ is obtained when $E[T] < \infty$, $\sigma_X < \infty$.

Sub-diffusion with scale $L \sim t^{\beta/2}$ is obtained with $E[T] = \infty$, $\sigma_X < \infty$ and distribution of waiting times

$$f_T(t) \sim \frac{1}{t^{1+\beta}}, \qquad \beta \in (0, 1).$$

Super-diffusion with $L \sim t^{1/\alpha}$ is obtained with $E[T] < \infty$, $\sigma_X = \infty$ and distribution of step lengths

$$f_X(x) \sim \frac{1}{x^{1+\alpha}}, \qquad \alpha \in (0, 2).$$

When $E[T] = \infty$ and $\sigma_X = \infty$, and when

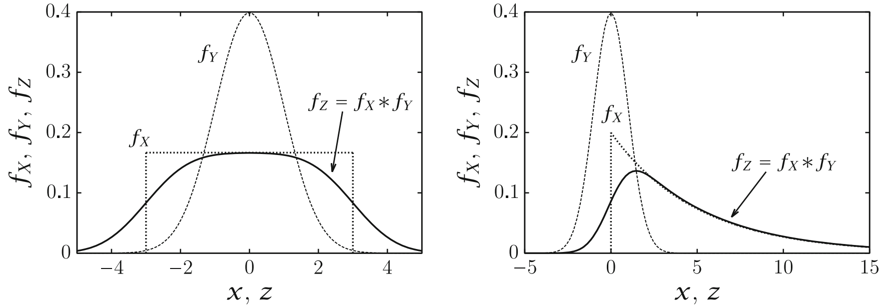


Fig. 6.11 [Left] Convolution of the uniform distribution $U(-3, 3)$ and the standardized normal distribution $N(0, 1)$. [Right] Convolution of the exponential distribution with parameter $\lambda = 0.2$ and the standardized normal distribution

$$f_X(x) \sim \frac{1}{x^{1+\alpha}}, \quad f_T(t) \sim \frac{1}{t^{1+\beta}}, \qquad \alpha \in (0, 2), \quad \beta \in (0, 1),$$

the scale is $L \sim t^{\beta/\alpha}$. The walks are super-diffusive if $2\beta > \alpha$ and sub-diffusive otherwise. Processes for which $E[T] = \infty$ are deeply non-Markovian: this means that the values of the process at given time depend on its whole history, not just on the immediately preceding state. Further reading can be found in [19, 21].

6.9 Problems

6.9.1 Convolutions with the Normal Distribution

Calculate the convolution of the normal distribution with the ① uniform, ② normal and ③ exponential distributions!

✎ ① The convolution of the uniform distribution with the density $f_X(x) = 1/(b - a)$ (see (3.1)) and the normal distribution with the density f_Y (Definition (3.7)) is

$$f_Z(z) = \int_a^b f_X(x) f_Y(z-x)\,\mathrm{d}x = \frac{1}{b-a}\frac{1}{\sqrt{2\pi}\,\sigma}\int_a^b \exp\left[-\frac{(z-x)^2}{2\sigma^2}\right]\mathrm{d}x$$

$$= \frac{1}{b-a}\frac{1}{\sqrt{2\pi}}\int_{(a-z)/\sigma}^{(b-z)/\sigma} \mathrm{e}^{-u^2/2}\,\mathrm{d}u = \frac{1}{2(b-a)}\left[\mathrm{erf}\left(\frac{b-z}{\sqrt{2}\,\sigma}\right) - \mathrm{erf}\left(\frac{a-z}{\sqrt{2}\,\sigma}\right)\right].$$

This function is shown in Fig. 6.11 (left).

Part ② is easily solved by using characteristic functions (B.20) and property (B.22):

$$\phi_Z(t) = \phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) = \mathrm{e}^{\mathrm{i}\,(\mu_X+\mu_Y)t}\mathrm{e}^{-(\sigma_X^2+\sigma_Y^2)t^2/2} = \mathrm{e}^{\mathrm{i}\,\mu_Z t}\mathrm{e}^{-\sigma_Z^2 t^2/2}.$$

From this it is clear that the convolution of two normal distributions with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2 is also a normal distribution, with mean $\mu_Z = \mu_X + \mu_Y$ and variance $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$.

Problem ③ requires us to convolute the distribution with the probability density $f_X(x) = \lambda \exp(-\lambda x)$ (see (3.4)) and the normal distribution, where we set $\mu = 0$:

$$f_Z(z) = \frac{\lambda}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{z} \mathrm{e}^{-\lambda(z-y)}\mathrm{e}^{-y^2/(2\sigma^2)}\,\mathrm{d}y.$$

Upon rearranging the exponent,

$$\begin{aligned} -\lambda(z-y) - \frac{y^2}{2\sigma^2} &= -\frac{\lambda}{2\sigma^2}\left(2\sigma^2(z-y) + \frac{y^2}{\lambda} + \lambda\sigma^4 - \lambda\sigma^4\right) \\ &= -\lambda z + \frac{\lambda^2\sigma^2}{2} - \frac{1}{2\sigma^2}\left(y - \lambda\sigma^2\right)^2, \end{aligned}$$

it follows that

$$\begin{aligned} f_Z(z) &= \frac{\lambda}{\sqrt{2\pi}\,\sigma} \exp\left(-\lambda z + \frac{\lambda^2\sigma^2}{2}\right) \int_{-\infty}^{(z-\lambda\sigma^2)/\sigma} \mathrm{e}^{-u^2/2}\,\mathrm{d}u \\ &= \lambda \exp\left(-\lambda z + \frac{\lambda^2\sigma^2}{2}\right) \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{z - \lambda\sigma^2}{\sqrt{2}\,\sigma}\right)\right]. \end{aligned}$$

This function is shown in Fig. 6.11 (right).

6.9.2 Spectral Line Width

The lines in emission spectra of atoms and molecules have finite widths [24]. Line broadening has three contributions: the *natural width* (N), *collisional broadening* (C) due to inelastic collisions of radiating particles, and *Doppler broadening* (D). Calculate a realistic spectral line profile by convoluting these three distributions.

✎ The natural width of the line in spontaneous emission—usually the smallest contribution to broadening—has a Lorentz (Cauchy) profile with a width of $\Delta\nu_\mathrm{N}$,

$$\phi_\mathrm{N}(\nu) = \frac{1}{\pi}\frac{\Delta\nu_\mathrm{N}/2}{(\nu-\nu_0)^2 + (\Delta\nu_\mathrm{N}/2)^2}.$$

As noted in the discussion of (3.19), such a distribution embodies a Fourier transformation of the exponential time dependence of the decays into Fourier space.

The broadening due to inelastic collisions depends on pressure and temperature—approximately one has $\Delta\nu_{\mathrm{C}} \propto p/\sqrt{T}$—and has a Cauchy profile as well:

$$\phi_{\mathrm{C}}(\nu) = \frac{1}{\pi}\frac{\Delta\nu_{\mathrm{C}}/2}{(\nu-\nu_0)^2 + (\Delta\nu_{\mathrm{C}}/2)^2}. \tag{6.17}$$

A convolution of two Cauchy distributions is again a Cauchy distribution,

$$\phi_{\mathrm{N+C}}(\nu) = \int_{-\infty}^{\infty} \phi_{\mathrm{N}}(\rho)\phi_{\mathrm{C}}(\nu-\rho)\,\mathrm{d}\rho = \frac{1}{\pi}\frac{(\Delta\nu_{\mathrm{N}} + \Delta\nu_{\mathrm{C}})/2}{(\nu-\nu_0)^2 + (\Delta\nu_{\mathrm{N}} + \Delta\nu_{\mathrm{C}})^2/4},$$

where we have shifted the origin of ϕ_{C} before integrating by setting $\nu_0 = 0$ in (6.17). If we had failed to do that, the peak of the convoluted distribution $\phi_{\mathrm{N+C}}$ would shift from ν_0 to $2\nu_0$—see Sect. 6.1.1!

The Doppler effect is proportional to the velocity of the radiating objects, which is normally distributed (see (3.14) for a single velocity component), hence the corresponding contribution to the line profile has the form

$$\phi_{\mathrm{D}}(\nu) = \frac{2\sqrt{\log 2}}{\sqrt{\pi}\Delta\nu_{\mathrm{D}}}\exp\left\{-\left(\frac{2\sqrt{\log 2}}{\Delta\nu_{\mathrm{D}}}(\nu-\nu_0)\right)^2\right\}.$$

The final spectral line shape is calculated by the convolution of the distributions $\phi_{\mathrm{N+C}}$ and ϕ_{D}, where again the origin must be shifted. We obtain

$$\phi_{\mathrm{V}}(\nu) = \int_{-\infty}^{\infty} \phi_{\mathrm{N+C}}(\rho)\phi_{\mathrm{D}}(\nu-\rho)\,\mathrm{d}\rho = \frac{2\sqrt{\log 2}}{\sqrt{\pi}\Delta\nu_{\mathrm{D}}}\left\{\frac{a}{\pi}\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-x^2}}{(w-x)^2 + a^2}\,\mathrm{d}x\right\},$$

where

$$a = \sqrt{\log 2}\,\frac{\Delta\nu_{\mathrm{N}} + \Delta\nu_{\mathrm{C}}}{\Delta\nu_{\mathrm{D}}}, \qquad w = 2\sqrt{\log 2}\,\frac{\nu-\nu_0}{\Delta\nu_{\mathrm{D}}}.$$

This is called the *Voigt* distribution. The natural width is usually neglected because $\Delta\nu_{\mathrm{N}} \ll \Delta\nu_{\mathrm{C}}, \Delta\nu_{\mathrm{D}}$. How well ϕ_{V} describes an actual line shape (as compared to the Cauchy and Gaussian profiles) is shown in Fig. 6.12.

6.9.3 Random Structure of Polymer Molecules

(Adapted from [5].) A polymer molecule can be envisioned as a chain consisting of a large number of equal, rigid, thin segments of length L. Starting at the origin, a molecule grows by attaching to the current terminal point further and further segments in arbitrary directions in space. ① What is the probability distribution for the position of the terminal point? ② Calculate the expected distance $\overline{R}$ between the initial and terminal point of the chain and $\overline{R^2}$!

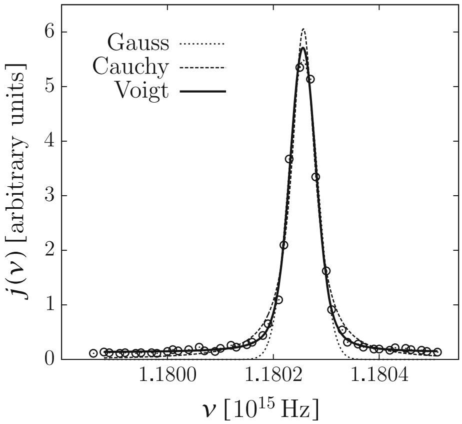


Fig. 6.12 Description of the Si (III) emission line at the wave-length of 254.182 nm (compare to Fig. 3.6) by a Gaussian (normal), Cauchy (Lorentz) and Voigt distribution with added constant background

✎ ① When a new segment is attached to the chain, it "chooses" its orientation at random: the directional distribution is therefore isotropic, $f_\Theta(\cos\theta) = \mathrm{d}F_\Theta/\mathrm{d}(\cos\theta) = 1/2$. For a projection of a single segment onto an arbitrary direction (e.g. x) we have

$$\overline{X_1} = L\,\overline{\cos\Theta} = L\int_0^\pi \cos\theta f_\Theta(\cos\theta)\,\sin\theta\,\mathrm{d}\theta = 0,$$

$$\sigma^2_{X_1} = \overline{X_1^2} = L^2\,\overline{\cos^2\Theta} = L^2\int_0^\pi \cos^2\theta f_\Theta(\cos\theta)\,\sin\theta\,\mathrm{d}\theta = \frac{L^2}{3}. \qquad (6.18)$$

The X-coordinate of the terminal point of an N-segment chain is the sum of independent and identically distributed random variables X_1 so, by the central limit theorem, it is nearly normally distributed at large N, with expected value $\overline{X} = N\overline{X_1} = 0$ and variance $\sigma_X^2 = \overline{X^2} = N\sigma^2_{X_1} = NL^2/3$. The corresponding probability density is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma_X}\exp\left(-\frac{x^2}{2\sigma_X^2}\right) = \sqrt{\frac{3}{2\pi NL^2}}\exp\left(-\frac{3x^2}{2NL^2}\right).$$

The x, y and z projections are *not* independent when a single segment is attached, but they are independent on average (after many attachments), so the same reasoning applies to Y and Z coordinates. Since $R^2 = X^2 + Y^2 + Z^2$, the probability density corresponding to the radial distribution of the terminal point of the chain is

$$f_R(r) = f_X(x)f_X(y)f_X(z) = \left(\frac{3}{2\pi NL^2}\right)^{3/2}\exp\left(-\frac{3r^2}{2NL^2}\right). \qquad (6.19)$$

② This can be used to calculate the expected values of R and R^2:

$$\overline{R} = \int_0^\infty r f_R(r)\,4\pi r^2\,\mathrm{d}r = L\sqrt{\frac{8N}{3\pi}}, \qquad \overline{R^2} = \int_0^\infty r^2 f_R(r)\,4\pi r^2\,\mathrm{d}r = NL^2.$$

The latter can also be derived by recalling (6.18), since

$$\overline{R^2} = \overline{X^2 + Y^2 + Z^2} = \overline{X^2} + \overline{Y^2} + \overline{Z^2} = 3\,\frac{NL^2}{3} = NL^2.$$

There is yet another path to the same result. Each segment ($n = 1, 2, \ldots, N$) is defined by a vector $\boldsymbol{r}_n = (x_n, y_n, z_n)^{\mathrm{T}}$. We are interested in the average square of the sum vector,

$$R^2 = |\boldsymbol{R}|^2 = \left(\sum_{m=1}^{N} \boldsymbol{r}_m\right)^{\mathrm{T}} \left(\sum_{n=1}^{N} \boldsymbol{r}_n\right) = \sum_{n} \boldsymbol{r}_n^2 + \sum_{m\neq n} \boldsymbol{r}_m^{\mathrm{T}} \boldsymbol{r}_n.$$

Averaging the second sum yields zero due to random orientations, $\overline{\boldsymbol{r}_m^{\mathrm{T}} \boldsymbol{r}_n} = 0$, hence

$$\overline{R^2} = \sum_{n=1}^{N} \overline{\boldsymbol{r}_n^2} = N\overline{\boldsymbol{r}_1^2} = NL^2.$$

6.9.4 Scattering of Thermal Neutrons in Lead

(Adapted from [5].) A neutron moves with velocity v in lead and scatters elastically off lead nuclei. The average time between collisions is τ, corresponding to the mean free path $\lambda = v\tau$. The times between consecutive collisions are mutually independent, and each scattering is isotropic. ① What is the (spatial) probability distribution of neutrons at long times? Calculate the average distance $\overline{R}$ of neutrons from the origin and $\overline{R^2}$! ② Demonstrate that $\overline{R^2}$ is proportional to time, so the process has the usual diffusive nature! The diffusion coefficient D is defined by the relation $\overline{R^2} = 6Dt$. How does D depend on λ and v?

✎ ① Isotropic scattering implies $f_\Theta(\cos\theta) = \mathrm{d}F_\Theta/\mathrm{d}(\cos\theta) = 1/2$. But we must also take into account the times between collisions or the distances l traversed by the neutron between collisions, $f_T(t) = \mathrm{d}F_T/\mathrm{d}t = \tau^{-1}\exp(-t/\tau)$, thus

$$f_L(l) = \frac{\mathrm{d}F_L}{\mathrm{d}l} = \frac{\mathrm{d}F_T}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}l} = \frac{1}{\tau}\,\mathrm{e}^{-t/\tau}\frac{1}{v} = \frac{1}{\lambda}\,\mathrm{e}^{-l/\lambda},$$

where $l = vt$. The joint probability density of the linear and angular variable, relevant to each collision, is therefore

$$f_{L,\Theta}(l, \cos\theta) = \frac{1}{2\lambda}\,\mathrm{e}^{-l/\lambda}.$$

The expected value of the projection of the neutron trajectory between two collisions onto the x-axis and the corresponding variance are

$$\overline{X_1} = \overline{L\cos\Theta} = \int_0^\infty \int_0^\pi l\cos\theta\, f_{L,\Theta}(l,\cos\theta)\,\mathrm{d}l\ \sin\theta\,\mathrm{d}\theta = 0,$$

$$\sigma_{X_1}^2 = \overline{X_1^2} = \overline{L^2\cos^2\Theta} = \int_0^\infty \int_0^\pi l^2\cos^2\theta\, f_{L,\Theta}(l,\cos\theta)\,\mathrm{d}l\ \sin\theta\,\mathrm{d}\theta = \frac{2\lambda^2}{3}.$$

Hence, as in Sect. 6.9.3, $\overline{X} = N\overline{X_1} = 0$ after N scatterings, while $\sigma_X^2 = N\sigma_{X_1}^2 = 2N\lambda^2/3$. Therefore the probability density for the distribution of R (distance from the origin to the current collision point) has the same functional form as in (6.19),

$$f_R(r) = \left(\frac{3}{4\pi N\lambda^2}\right)^{3/2} \exp\left(-\frac{3r^2}{4N\lambda^2}\right),$$

one only needs to insert the variance $2N\lambda^2/3$ instead of $NL^2/3$. It follows that

$$\overline{R} = \int_0^\infty r f_R(r)\, 4\pi r^2\,\mathrm{d}r = \lambda\sqrt{\frac{16N}{3\pi}}, \qquad \overline{R^2} = \int_0^\infty r^2 f_R(r)\, 4\pi r^2\,\mathrm{d}r = 2N\lambda^2.$$

② The elapsed time after N collisions is $t = N\lambda/v$, so that indeed $\overline{R^2}$ is proportional to time, $\overline{R^2} = 2N\lambda^2 = 2(vt/\lambda)\lambda^2 = 2vt\lambda$. From the definition $\overline{R^2} = 6Dt$ it follows that

$$D = \frac{\lambda v}{3}.$$

6.9.5 *Distribution of Extreme Values of Normal Variables* ★

Let continuous random variables X_i $(i = 1, 2, \ldots, n)$ be normally distributed, $X_i \sim N(0, 1)$, with the corresponding distribution function $F_X(x) = \Phi(x)$ and probability density $f_X(x) = \phi(x)$ for each variable:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-t^2/2}\,\mathrm{d}t, \qquad \phi(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}.$$

What is the distribution function F_{M_n} of the values $M_n = \max\{X_1, X_2, \ldots, X_n\}$? This Problem [13] is a continuation of the Example on p. 157 and represents a method to determine the parameters a_n and b_n for the scaling formula (6.10) such that the limiting distribution (6.9) is non-degenerate.

✎ Let $0 \le \tau \le \infty$ and let $\{u_n\}$ be a sequence of real numbers such that

$$1 - F_X(u_n) \to \frac{\tau}{n} \quad \text{when } n \to \infty. \tag{6.20}$$

By definition of the exponential function by a series we get

$$F_{M_n}(u_n) = P\big(M_n \le u_n\big) = F_X^n(u_n) = \big[1 - \big(1 - F_X(u_n)\big)\big]^n = \Big[1 - \frac{\tau}{n} + \mathcal{O}(1/n)\Big]^n \sim \mathrm{e}^{-\tau},$$

when $n \to \infty$. The leading dependence of the distribution function, $F_{M_n} \sim \exp(-\tau)$, follows without explicit reference to the parent distribution $F_X(x)$ being normal! A motivation for a specific form of τ can then be found in the asymptotic property of the normal distribution

$$1 - \Phi(z) \sim \frac{\phi(z)}{z}, \quad n \to \infty. \tag{6.21}$$

Let $\tau = \mathrm{e}^{-x}$. The reason for this choice, fully consistent with (6.20), will become clear in the following: this is the only way to obtain in the final expression a *linear* dependence on x in the rescaled argument of the distribution function. By comparing (6.20) to (6.21) we obtain

$$1 - \Phi(u_n) \sim \frac{\mathrm{e}^{-x}}{n} \sim \frac{\phi(u_n)}{u_n} \quad \Rightarrow \quad \frac{1}{n}\,\mathrm{e}^{-x}\,\frac{u_n}{\phi(u_n)} \to 1.$$

Taking the logarithm we get $-\log n - x + \log u_n - \log \phi(u_n) \to 0$ or

$$-\log n - x + \log u_n - \tfrac{1}{2}\log 2\pi + \tfrac{1}{2}\,u_n^2 \to 0. \tag{6.22}$$

For fixed x in the limit $n \to \infty$ one therefore has $u_n^2/(2\log n) \to 1$, so that taking the logarithm again yields $2\log u_n - \log 2 - \log\log n \to 0$ or

$$\log u_n = \tfrac{1}{2}\big(\log 2 + \log\log n\big) + \mathcal{O}(1).$$

Inserting this in (6.22), we get $\frac{1}{2}u_n^2 = x + \log n - \frac{1}{2}\log 4\pi - \frac{1}{2}\log\log n + \mathcal{O}(1)$, hence

$$u_n^2 = 2\log n\left[1 + \frac{x - \frac{1}{2}\log 4\pi - \frac{1}{2}\log\log n}{\log n} + \mathcal{O}\left(\frac{1}{\log n}\right)\right],$$

and finally, after taking the square root,

$$u_n = \sqrt{2\log n}\left[1 + \frac{x - \frac{1}{2}\log 4\pi - \frac{1}{2}\log\log n}{2\log n} + \mathcal{O}\left(\frac{1}{\log n}\right)\right].$$

This expression has the form

$$u_n = a_n x + b_n + \mathcal{O}\big((\log n)^{-1/2}\big) = a_n x + b_n + \mathcal{O}\big(a_n\big),$$

whence we read off the normalization constants a_n and b_n:

$$a_n = \frac{1}{\sqrt{2\log n}}, \qquad b_n = \sqrt{2\log n} - \frac{\log\log n + \log 4\pi}{2\sqrt{2\log n}}. \tag{6.23}$$

These constants imply $P\big(M_n \le a_n x + b_n + \mathcal{O}(a_n)\big) \to \exp\big(-\mathrm{e}^{-x}\big)$, that is,

$$F_{M_n}(x) = P\left(\frac{M_n - b_n}{a_n} + \mathcal{O}(1) \le x\right) \to \exp\left(-\mathrm{e}^{-x}\right).$$

The distribution of extreme values of normally distributed variables is therefore of the Gumbel type (6.12) with normalization constants (6.23).

References

1. D.L. Evans, L.M. Leemis, Algorithms for computing the distributions of sums of discrete random variables. Math. Comp. Model. **40**, 1429 (2004)
2. W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 2, 2nd edn. (Wiley, New York, 1971)
3. I.S. Tyurin, On the accuracy of the Gaussian approximation. Dokl. Math. **80**, 840 (2009)
4. A. Lyon, Why are normal distributions normal? Brit. J. Phil. Sci. **65**, 621 (2014)
5. I. Kuščer, A. Kodre, *Mathematik in Physik und Technik* (Springer, Berlin, 1993)
6. J.P. Nolan, *Stable Distributions—Models for Heavy Tailed Data* (Birkhäuser, Boston, 2010)
7. S. Borak, W. Härdle, R. Weron, *Stable Distributions, SFB 649 Discussion Paper 2005–008* (Humboldt University Berlin, Berlin, 2005)
8. S. Širca, M. Horvat, *Computational Methods for Physicists* (Springer, Berlin, 2012)
9. GSL (GNU Scientific Library), http://www.gnu.org/software/gsl
10. G.G. Márquez, *One Hundred Years of Solitude* (HarperCollins, New York, 2006)
11. E.J. Gumbel, *Statistics of Extremes* (Columbia University Press, New York, 1958)
12. S. Coles, *An Introduction to Statistical Modeling of Extreme Values* (Springer, Berlin, 2001)
13. M.R. Leadbetter, G. Lindgren, H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes* (Springer, New York, 1983)
14. S.I. Resnick, *Extreme Values, Regular Variation, and Point Processes* (Springer, New York, 1987)
15. R.A. Fisher, L.H.C. Tippett, On the estimation of the frequency distribution of the largest or smallest member of a sample. Proc. Camb. Phil. Soc. **24**, 180 (1928)
16. B.V. Gnedenko, Sur la distribution limite du terme maximum d'une série aléatoire. Ann. Math. **44**, 423 (1943)
17. MeteoSwiss IDAWEB, http://gate.meteoswiss.ch/idaweb/
18. B.D. Hughes, *Random Walks and Random Environments: Vol. 1: Random Walks* (Oxford University Press, New York, 1995)
19. M. Bazant, *Random walks and diffusion*, MIT OpenCourseWare, Course 18.366, http://ocw.mit.edu/courses/mathematics/
20. E.W. Montroll, G.H. Weiss, Random walks on lattices II. J. Math. Phys. **6**, 167 (1965)
21. R. Metzler, J. Klafter, The random walk's guide to anomalous diffusion: a fractional dynamics approach. Phys. Rep. **339**, 1 (2000)
22. M.F. Shlesinger, B.J. West, J. Klafter, Lévy dynamics of enhanced diffusion: application to turbulence. Phys. Rev. Lett. **58**, 1100 (1987)
23. V. Tejedor, R. Metzler, Anomalous diffusion in correlated continuous time random walks. J. Phys. A: Math. Theor. **43**, 082002 (2010)
24. J.J. Brehm, W.J. Mullin, *Introduction to the Structure of Matter* (Wiley, New York, 1989)

Part II
Determination of Distribution Parameters

Chapter 7
Statistical Inference from Samples

Abstract Any kind of empirical determination of probability distributions and their parameters amounts to statistical inference procedures based on samples randomly drawn from a population. The concepts of the statistic and the estimator are introduced, paying attention to their consistency and bias. Sample mean and sample variance are defined, and three most relevant sample distributions are investigated: distribution of sums and differences, distribution of variances, and distribution of variance ratios. Confidence intervals for the sample mean and sample variance are discussed. The problem of outliers is elucidated in the context of robust measures, and linear (Pearson) and non-parametric (Spearman) correlations are presented.

Chapters 7–10 are devoted to basics of statistics. The main task of statistics is the empirical determination of probability distributions and their parameters.

We start by introducing the concepts of population and sample. A *population* is a finite or infinite set of elements from which we acquire *samples*. By using statistical methods we strive to determine the properties of the entire population by analyzing only its sample, even though the sample may be much smaller than the population. If a quantity represented by a random variable X is measured (counted, realized, recorded) n-times, we obtain a set of values $\{x_i\}_{i=1}^n$, burdened with some error or *uncertainty*. Part of this uncertainty has random (*statistical*) nature: the values x_i are scattered because, in general, a sample contains new elements each time it is acquired. This part of the uncertainty can be reduced by increasing the sample size. The other part of the uncertainty has a *systematic* origin and can not be removed by augmenting the sample.

Example From a population of $N = 2 \times 10^6$ we acquire a sample of $n = 1000$ people and measure their heights. We would like to use the measured n values to determine the average height and its variance, and offer some kind of a statement on what these numbers mean in the context of the whole population. If we measure the height of 1000 randomly selected people today and 1000 randomly selected people tomorrow, we shall in general obtain two different averages and variances (statistical uncertainty). If we use a faulty instrument that constantly gives a height 1 cm too

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_7

short, we will obtain wrong heights regardless of the sample size and regardless of whether sampling is repeated multiple times (systematic uncertainty). ◁

A population can be finite or infinite ($N = \infty$). Tossing a coin many times, for example, yields an estimate of the probability of observing head or tail (which will be approximately 1/2, see Sect. 1.3), but in this case the population consists of the set of all possible tosses, which is infinite.

7.1 Statistics and Estimators

Our vantage point is the random sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ from a population characterized by an unknown parameter θ. We would like to estimate this parameter from the sample. The function

$$\hat{\theta} = T(\boldsymbol{X}) = T(X_1, X_2, \ldots, X_n) \tag{7.1}$$

of random variables X_i with values x_i, used to obtain the estimate for θ, is called the *estimator* of parameter θ. Conventionally we use the same notation for the estimator as a prescription for the variables, e.g. $\hat{\theta} = (X_1 + X_2)/2$, as for its concrete value or *estimate*, e.g. $\hat{\theta} = (x_1 + x_2)/2$. Any function of the form (7.1) is called a *sample statistic*, while probability distributions of such statistics are called *sample distributions*.

Of course, the functional form of a statistic is not arbitrary: above all, one wishes to devise the statistic $\hat{\theta} = T(\boldsymbol{X})$ so that it is *consistent*. This means that the estimate $\hat{\theta}$ converges to the true value θ when the number of observations n is increased: if a sample of size n results in an estimate $\hat{\theta}_n$, then for any positive ε and η there should exist m such that $P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) < \eta$ holds true for each $n > m$. This is approximately equivalent to the statement that the variance of an estimator goes to zero for infinite samples:

$$\lim_{n\to\infty} \mathrm{var}\big[\hat{\theta}_n\big] = 0. \tag{7.2}$$

Moreover, it is usually desirable that the estimator is *unbiased*. This means that for samples of arbitrary size n, not only infinite ones, the expected value of $\hat{\theta}$ is equal to the true parameter,

$$E\big[\hat{\theta}\big] = \theta, \quad \forall n. \tag{7.3}$$

If, on the contrary,

$$E\big[\hat{\theta}\big] = \theta + b\big(\hat{\theta}\big), \tag{7.4}$$

where $b \neq 0$, we say that the estimator is *biased*. For sensible estimators one expects $b(\hat{\theta}) \ll \theta$ and, say, $b(\hat{\theta}) \sim 1/n$ when $n \to \infty$.

7.1.1 Sample Mean and Sample Variance

The crucial parameters of interest for any distribution of a random variable are its mean and variance. Their values inferred from a given sample are generally different from their values for the whole population.

Suppose we acquire a sample of n values from a population so that any value can occur multiple times. In the case of body heights this means that a person is chosen at random, her height is measured—this is the value of X—and "returned" to the population, whence she can be randomly "drawn" again. We say that the sample has been obtained *by replacement:* if $N \gg n$, we should not have any second thoughts about that. If all values in the sample have equal weights, the estimator

$$T = \overline{X} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right) \tag{7.5}$$

gives the *sample mean* of body heights according to (4.1). Note that the line above the symbol now represents the sample mean, while in previous sections it has been used as an alternative notation for the expected value. In the following we use the notation

$$\begin{aligned} \mu,\ \sigma^2 &\Leftrightarrow \text{population}, \\ \overline{X}\ (\text{or } \overline{x}),\ s_X^2\ (\text{or } s_x^2) &\Leftrightarrow \text{sample}, \end{aligned}$$

while expected values will be strictly denoted by $E[\,\bullet\,]$. (As usual, lower-case letters imply concrete values of the corresponding statistics upon each sampling.) Our estimate for the unknown *population mean* $\theta = \mu$, which we wish to infer based on the concrete sample $\{x_i\}_{i=1}^n$, is therefore

$$\hat{\theta} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

The estimator (7.5) is clearly unbiased since, according to (7.3), its expected value is equal to the population mean,

$$E\big[\overline{X}\big] = \frac{1}{n}\Big(E[X_1] + E[X_2] + \cdots + E[X_n]\Big) = \frac{1}{n}\,n\mu = \mu. \tag{7.6}$$

It is also consistent, since its variance approaches zero when $n \to \infty$:

$$\mathrm{var}\big[\overline{X}\big] = \mathrm{var}\left[\frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right)\right] = n\frac{1}{n^2}\,\mathrm{var}[X] = \frac{\sigma^2}{n}. \tag{7.7}$$

This looks nice, but just as we do not know the true population mean μ, the true population variance $\mathrm{var}[X] = \sigma^2$ is also unknown. At best, we can resort to (4.14) to devise a formula for the *sample variance*,

$$s_X^2 \overset{?}{=} \frac{1}{n}\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2, \tag{7.8}$$

which undoubtedly is a kind of estimator for σ^2, but is it unbiased? If it were unbiased, its expected value should be equal to the population variance,

$$E\big[s_X^2\big] \overset{?}{=} \sigma^2.$$

Does this hold true? Let us focus on the first term in the sum and write it as

$$\begin{aligned} X_1 - \overline{X} &= X_1 - \frac{1}{n}\big(X_1 + X_2 + \cdots + X_n\big) \\ &= \frac{1}{n}\big[(n-1)X_1 - X_2 - \cdots - X_n\big] \\ &= \frac{1}{n}\big[(n-1)(X_1-\mu) - (X_2-\mu) - \cdots - (X_n-\mu)\big], \end{aligned}$$

then square it,

$$\big(X_1 - \overline{X}\big)^2 = \frac{1}{n^2}\Big[(n-1)^2\big(X_1-\mu\big)^2 + \big(X_2-\mu\big)^2 + \cdots + \big(X_n-\mu\big)^2 + \text{mixed terms}\Big].$$

The variables X_i and X_j ($i \neq j$) are mutually independent, so $E[(X_i-\mu)(X_j-\mu)] = 0$ and the mixed terms do not contribute to the expected value:

$$\begin{aligned} E\big[\big(X_1 - \overline{X}\big)^2\big] &= \frac{1}{n^2}\Big\{(n-1)^2 E\big[\big(X_1-\mu\big)^2\big] + E\big[\big(X_2-\mu\big)^2\big] + \cdots + E\big[\big(X_n-\mu\big)^2\big]\Big\} \\ &= \frac{1}{n^2}\left\{(n-1)^2\sigma^2 + \underbrace{\sigma^2 + \cdots + \sigma^2}_{(n-1)\sigma^2}\right\} = \frac{n-1}{n}\,\sigma^2. \end{aligned}$$

The sum (7.8) contains n such terms, and there is a factor $1/n$ up front, thus

$$E\big[s_X^2\big] = \frac{n-1}{n}\,\sigma^2. \tag{7.9}$$

Therefore s_X^2 is a *biased* estimator for the population variance: an unbiased estimator is obtained if the right-hand side of (7.8) is multiplied by $n/(n-1)$,

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2. \tag{7.10}$$

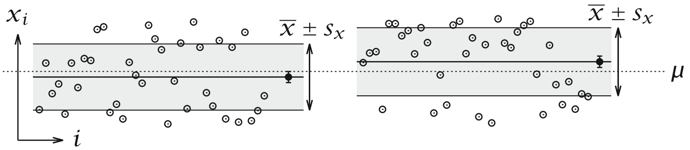


Fig. 7.1 Samples of size $n = 36$ with different means and almost equal effective deviations $\overline{x} = -0.052$, $s_x = 0.313$ (*left*) and $\overline{x} = 0.092$, $s_x = 0.322$ (*right*), taken from a population with mean $\mu = 0$ and effective deviation $\sigma = 0.288$. Full circles with error bars denote the sample means $\overline{x}$ and their uncertainties $s_x/\sqrt{n}$

Sometimes both (7.8) and (7.10) are invoked as formulas for sample variance, although by analogy to definitions (4.14) and (4.15) only the first form is correct. With respect to bias, the formulas are not equivalent, except for $n \gg 1$ when the difference is negligible. An illustration of two samples with different means $\overline{x}$ and almost equal variances s_x^2 is in Fig. 7.1. See also Problem 7.6.1.

When formula (7.7) is applied to the estimator (7.10), one obtains the *estimator of the variance of the sample mean*

$$s_{\overline{X}}^2 = \frac{s_X^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \big(X_i - \overline{X}\big)^2.$$

Taking the square root yields the uncertainty or *"error" of the mean*

$$s_{\overline{X}} = \frac{s_X}{\sqrt{n}}.$$

This result offers an important lesson whose importance can not be over-emphasized. Each sample from a population has a different sample mean and different sample variance. The observations will usually be scattered by approximately $\pm s_x$ (shaded areas in Fig. 7.1), but that does not mean that the average will also be scattered by $\pm s_x$—its uncertainty will only be $\pm s_x/\sqrt{n}$! We therefore simply write

$$\overline{x} = \mu \pm \frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \mu = \overline{x} \pm \frac{s_x}{\sqrt{n}}, \tag{7.11}$$

which is about the same. The expression on the left implies that the sample average approximately equals the true (population) average, and its uncertainty is $\pm\sigma/\sqrt{n}$. The formula on the right says that $\overline{x}$ is a good approximation for μ, the error is, perhaps, $\pm s_x/\sqrt{n}$. Regardless of the interpretation these formulas dictate the sample size needed for the desired precision: if, for example, we wish to determine μ to a precision of $0.01\,\sigma$, we need $n = 10^4$ observations.

Now let us assume that individual X_i are normally distributed. How can we estimate the scattering of the sample variance s_X^2 about the population variance σ^2? At large n, where the distinction between formulas (7.8) and (7.10) is immaterial, this "variance of variance" equals

$$E\left[\left(s_X^2-\sigma^2\right)^2\right]=E\left[\left(\frac{\left(X_1-\overline{X}\right)^2+\left(X_2-\overline{X}\right)^2+\cdots+\left(X_n-\overline{X}\right)^2}{n}-\sigma^2\right)^2\right].$$

Squaring the expression in square brackets first yields n quartic terms of the form $(X_i-\overline{X})^4$, $i=1,2,\ldots,n$, where $\overline{X}\approx E[X]=\mu$. The excess of a normally distributed continuous variable is zero, thus by (4.19) its fourth central moment is $M_4=E[(X_i-\mu)^4]=3\sigma^4$, contributing $3n\sigma^4/n^2$ to the final expression. Secondly, with the approximation $\overline{X}\approx\mu$ the factors in the products $2(X_i-\overline{X})^2(X_j-\overline{X})^2$, $i\neq j$, are independent, resulting in $n(n-1)/2$ terms of the form

$$2E[(X_i-\overline{X})^2(X_j-\overline{X})^2]\approx 2E[(X_i-\mu)^2]E[(X_j-\mu)^2]=2\sigma^2\sigma^2=2\sigma^4.$$

Their total contribution is $2\frac{1}{2}n(n-1)\sigma^4/n^2$. Thirdly, we are left with n mixed terms of the form $-2E[(X_i-\overline{X})^2]\sigma^2/n\approx -2E[(X_i-\mu)^2]\sigma^2/n=-2\sigma^4/n$ and the lone σ^4, thus at last

$$E\left[\left(s_X^2-\sigma^2\right)^2\right]\approx\frac{1}{n^2}\left(3n\sigma^4+2\frac{n(n-1)}{2}\sigma^4-2n^2\sigma^4\right)+\sigma^4=\frac{2\sigma^4}{n}.$$

We have obtained

$$s_X^2=\sigma^2\left(1\pm\sqrt{\frac{2}{n}}\right),$$

which in the case $n\gg 1$ implies

$$s_X=\sigma\left(1\pm\frac{1}{\sqrt{2n}}\right)\quad\text{or}\quad\sigma=s_X\left(1\pm\frac{1}{\sqrt{2n}}\right).$$

To determine σ to a precision of 1%, one therefore needs $n=5000$ observations.

Example (Adapted from [1].) The complete population consists of $N=5$ values $\{x_i\}_{i=1}^N=\{2,3,6,8,11\}$. The mean and variance of its elements are

$$\mu=E\big[X\big]=\frac{1}{N}\sum_{i=1}^N x_i=6,\qquad\sigma^2=\mathrm{var}[X]=\frac{1}{N}\sum_{i=1}^N\left(x_i-\mu\right)^2=10.8,$$

thus $\sigma=\sqrt{\mathrm{var}[X]}\approx 3.29$. From this population we draw all possible samples of size $n=2$ with replacement. There are $\mathcal{N}=N^2=25$ such samples:

$$\{2,2\},\ \{2,3\},\ \{2,6\},\ \{2,8\},\ \{2,11\},\ \{3,2\},\ \{3,3\},\ \ldots,\{11,11\}.$$

For each of these samples one can compute 25 sample means, (2.0, 2.5, 4.0, 5.0, 6.5, 2.5, 3.0, . . ., 11.0), denoted by $\overline{x}_k$ ($k = 1, 2, \ldots, \mathcal{N}$). Their expected value—i.e. the mean of the sample distribution of means—is

$$E\big[\overline{X}\big] = \frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} \overline{x}_k = \frac{150}{25} = 6 = \mu,$$

which is nothing but the true population mean, as expected according to (7.6). The variance of the sample distribution of means is

$$\mathrm{var}\big[\overline{X}\big] = \frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} \big(\overline{x}_k - \mu\big)^2 = \frac{135}{25} = 5.4,$$

which we also obtain from (7.7):

$$\mathrm{var}\big[\overline{X}\big] = \frac{\sigma^2}{n} = \frac{10.8}{2} = 5.4 \tag{7.12}$$

or $\sqrt{\mathrm{var}[\overline{X}]} \approx 2.32$.

If samples of size $n = 2$ are drawn *without replacement*, one can form only $\mathcal{N} = \binom{5}{2} = 10$ such samples:

$$\{2, 3\},\ \{2, 6\},\ \{2, 8\},\ \{2, 11\},\ \{3, 6\},\ \{3, 8\},\ \{3, 11\},\ \{6, 8\},\ \{6, 11\},\ \{8, 11\}.$$

The sample means $\overline{x}_k$ are now $2.5,\ 4.0,\ 5.0,\ 6.5,\ 4.5,\ \ldots,\ 9.5$. The expected value of the sample distribution of means is still $E[\overline{X}] = \mu = 6$, while the variance of the sample distribution of means is $\mathrm{var}[\overline{X}] = \frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} \big(\overline{x}_k - \mu\big)^2 = 4.05$. We have obtained a different result as in (7.12) because formula (7.7) is not applicable in the case of sampling without replacement. Instead, one should use

$$\mathrm{var}[\overline{X}] = \left(\frac{N - n}{N - 1}\right) \frac{\sigma^2}{n}.$$

Then indeed

$$\left(\frac{N - n}{N - 1}\right) \frac{\sigma^2}{n} = \left(\frac{5 - 2}{5 - 1}\right) \frac{10.8}{2} = 4.05.$$

We have used small N and n to convey the general idea, otherwise a set of five elements could hardly be identified with a “large” population ($N \gg n$) suitable for “proper” statistical analysis. ◁

7.2 Three Important Sample Distributions

7.2.1 Sample Distribution of Sums and Differences

Suppose we are dealing with *two* infinite populations with means μ_1 and μ_2 and variances σ_1 and σ_2. We draw a sample of size n_1 from the first population and a sample of size n_2 from the second, and compute the sample means $\overline{X}_1$ and $\overline{X}_2$. Referring to our previous findings—see in particular (4.6), (4.20) and (7.7)—we can write

$$E\big[\overline{X}_1 \pm \overline{X}_2\big] = E\big[\overline{X}_1\big] \pm E\big[\overline{X}_2\big] = \mu_1 \pm \mu_2 \tag{7.13}$$

and

$$\mathrm{var}\big[\overline{X}_1 \pm \overline{X}_2\big] = \mathrm{var}\big[\overline{X}_1\big] + \mathrm{var}\big[\overline{X}_2\big] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \tag{7.14}$$

If $n_1, n_2 \gtrsim 30$, the random variable

$$Z = \frac{\big(\overline{X}_1 \pm \overline{X}_2\big) - (\mu_1 \pm \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{7.15}$$

is distributed according to the standardized normal distribution to a very good approximation. The statement remains valid in the case of drawing with replacement from *finite* populations, as one can, in principle, draw an infinite sample from a population to which elements are restored after being drawn.

Example The lifetime of a circuit of type A is normally distributed, with mean $\mu_{\mathrm{A}} = 7.0\,\mathrm{yr}$ and standard deviation $\sigma_{\mathrm{A}} = 1.1\,\mathrm{yr}$. The circuits of type B have the mean lifetime $\mu_{\mathrm{B}} = 5.8\,\mathrm{yr}$ with the standard deviation $\sigma_{\mathrm{B}} = 0.9\,\mathrm{yr}$. We test $n_{\mathrm{A}} = 40$ of type-A circuits and $n_{\mathrm{B}} = 40$ of type-B circuits. What is the probability that A circuits will operate a year longer than B circuits?

By using (7.13) and (7.14) we obtain

$$\mu_{\mathrm{A}} - \mu_{\mathrm{B}} = 1.2\,\mathrm{yr}, \qquad \sqrt{\mathrm{var}\big[\overline{X}_{\mathrm{A}} - \overline{X}_{\mathrm{B}}\big]} = \sqrt{\frac{\sigma_{\mathrm{A}}^2}{n_{\mathrm{A}}} + \frac{\sigma_{\mathrm{B}}^2}{n_{\mathrm{B}}}} = 0.225\,\mathrm{yr}.$$

We are interested in the probability that the difference of mean lifetimes is larger than one year, $\overline{X}_{\mathrm{A}} - \overline{X}_{\mathrm{B}} > 1\,\mathrm{yr}$. The corresponding standardized variable (7.15) for the limit value $\overline{X}_{\mathrm{A}} - \overline{X}_{\mathrm{B}} = 1\,\mathrm{yr}$ is

$$Z = \frac{\big(\overline{X}_{\mathrm{A}} - \overline{X}_{\mathrm{B}}\big) - \big(\mu_A - \mu_{\mathrm{B}}\big)}{\sqrt{\mathrm{var}\big[\overline{X}_{\mathrm{A}} - \overline{X}_{\mathrm{B}}\big]}} = \frac{1 - 1.2}{0.225} \approx -0.89$$

and is normally distributed. So the probability we want is

$$P\bigl(\overline{X}_{\mathrm{A}} - \overline{X}_{\mathrm{B}} > 1\,\mathrm{yr}\bigr) = P\bigl(Z > -0.89\bigr) = \tfrac{1}{2} + P\bigl(0 \le Z \le 0.89\bigr) \approx 81.3\%,$$

where we have used Table D.1. ◁

7.2.2 Sample Distribution of Variances

Sample distributions of variances are obtained when one acquires all possible random samples of size n from the population and calculates the variance of each sample. From the population variance, σ^2, and the sample variance s_X^2 (in biased form (7.8)) we construct the random variable

$$\chi^2 = \frac{n s_X^2}{\sigma^2} = \sum_{i=1}^{n} \frac{\bigl(X_i - \overline{X}\bigr)^2}{\sigma^2}. \tag{7.16}$$

If random samples of size n are drawn from a *normally distributed* population, the statistic (7.16) is distributed according to the χ^2 distribution (3.21) with $n-1$ degrees of freedom.

Example Let us revisit the Example on p. 182, where we have drawn $\mathcal{N} = N^2 = 25$ samples of size $n = 2$ from a population of $N = 5$ elements. What are the mean and the variance of the corresponding sample variances, and what is the expected number of samples whose variance exceeds 7.15?

We first compute k ($k = 1, 2, \ldots, \mathcal{N}$) sample variances $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\bigl(x_i - \overline{x}\bigr)^2$ for each of the 25 samples. We obtain the variances

$$\begin{array}{lllllllll} 0, & 0.25, & 4, & \underline{9}, & \underline{20.25}, & 0.25, & 0, & 2.25, & \\ 6.25, & \underline{16}, & 4, & 2.25, & 0, & 1, & 6.25, & \underline{9}, & \\ 6.25, & 1, & 0, & 2.25, & \underline{20.25}, & \underline{16}, & 6.25, & 2.25, & 0. \end{array} \tag{7.17}$$

The mean of this sample distribution of variances is

$$E\bigl[s_X^2\bigr] = \frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} s_{x,k}^2 = \frac{135}{25} = 5.4.$$

This has been expected by (7.9), since $E\bigl[s_X^2\bigr] = (n-1)\sigma^2/n = 10.8/2 = 5.4$. The scattering of the sample variances is calculated by the usual formula:

$$\mathrm{var}\bigl[s_X^2\bigr] = \frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} \bigl(s_{x,k}^2 - E\bigl[s_X^2\bigr]\bigr)^2 = \frac{575.75}{25} = 23.03.$$

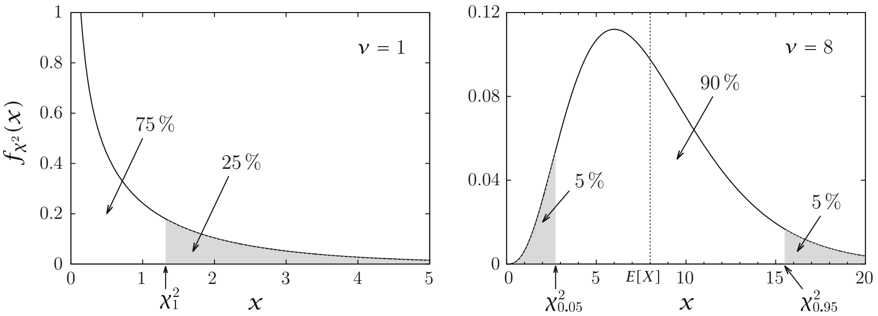


Fig. 7.2 The probability density of the χ^2 distribution for [Left] $\nu = 1$ and [Right] $\nu = 8$ degrees of freedom. The *dashed vertical line* indicates the distribution mean

The random variable χ^2 is given by (7.16). With the population variance, $\sigma^2 = 10.8$, and the prescribed sample variance, $s_1^2 = 7.15$, χ^2 takes the value

$$\chi_1^2 = \frac{ns_1^2}{\sigma^2} = \frac{2 \cdot 7.15}{10.8} \approx 1.32,$$

where the subscript 1 indicates that the variable is χ^2-distributed (see (3.21)) with $\nu = n - 1 = 1$ degree of freedom. The probability that the sample variance s_X^2 exceeds the prescribed variance s_1^2, is therefore equal to the probability that the value of χ^2 according to this distribution is larger than the critical value χ_1^2. From the first line of Table D.3 (column for $p = 0.75$ or $1 - p = 0.25$) and Fig. 7.2 (right) we find this probability to be

$$P\big(s_X^2 \geq s_1^2\big) = P\big(\chi^2 \geq \chi_1^2\big) = 0.25.$$

There are $\mathcal{N} = 25$ samples, $0.25 \cdot 25 = 6.25$ of which are expected to have a variance exceeding the prescribed one. In our Example there are indeed six: they correspond to the underlined elements in (7.17). ◁

7.2.3 *Sample Distribution of Variance Ratios*

From two *normally distributed* populations with variances σ_1^2 and σ_2^2 we draw two independent samples (one from each population) of sizes n_1 and n_2, respectively. Let the sample variances (in biased form (7.8)) be s_1^2 and s_2^2. Then the statistic

$$F = \left(\frac{n_1}{n_1 - 1}\right)\frac{s_1^2}{\sigma_1^2} \Big/ \left(\left(\frac{n_2}{n_2 - 1}\right)\frac{s_2^2}{\sigma_2^2}\right) \propto \left(\frac{s_1}{s_2}\right)^2$$

is distributed according to the F distribution (defined in (3.23)) with $(\nu_1, \nu_2) = (n_1 - 1, n_2 - 1)$ degrees of freedom.

Example From two normally distributed populations with variances $\sigma_1^2 = 5$ and $\sigma_2^2 = 9$ we draw two samples of size $n_1 = 8$ and $n_2 = 10$ from the first and second population, respectively. What is the probability that the variance of the first sample is at least twice the variance of the second sample? (A very small probability may be anticipated, as the variance of the first population is roughly twice smaller than the variance of the second, thus it seems highly unlikely that the situation would be nearly opposite in the *samples* from these populations.) We calculate the statistic

$$F = \left(\frac{n_1}{n_1 - 1}\right) \frac{s_1^2}{\sigma_1^2} \Big/ \left(\left(\frac{n_2}{n_2 - 1}\right) \frac{s_2^2}{\sigma_2^2}\right) = \frac{8}{7} \frac{s_1^2}{5} \Big/ \left(\frac{10}{9} \frac{s_2^2}{9}\right) \approx 1.85 \, \frac{s_1^2}{s_2^2}.$$

The Problem statement requires $s_1^2 > 2\, s_2^2$ or

$$F > 3.70.$$

The F statistic is distributed according to the F distribution (3.23) with $\nu_1 = n_1 - 1 = 7$ degrees of freedom (numerator) and $\nu_2 = n_2 - 1 = 9$ (denominator). From Tables D.5 and D.6 we read off the 95% and 99% quantiles $F_{0.95} = 3.29$ and $F_{0.99} = 5.61$, implying that the sought-after probability (that $s_1^2 > 2\, s_2^2$ and $F > 3.70$) is larger than 1% and smaller than 5% (Fig. 7.3 (left)). For a more precise answer, we integrate the density up to the specified bound (see Appendix D.1):

$$\int_{3.70}^{\infty} f_F(x; \nu_1 = 7, \nu_2 = 9) \, \mathrm{d}x = 1 - \int_0^{3.70} f_F(x; \nu_1 = 7, \nu_2 = 9) \, \mathrm{d}x \approx 0.036.$$

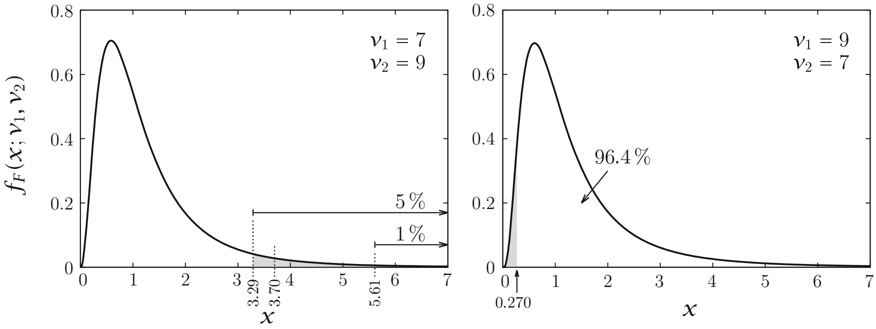


Fig. 7.3 Probability density of the F distribution for [Left] $\nu_1 = 7$, $\nu_2 = 9$ and [Right] $\nu_1 = 9$, $\nu_2 = 7$ degrees of freedom

What is in the numerator and what is in the denominator of the F ratio is irrelevant, one just needs consistent book-keeping. Let us switch the roles of the populations, so that $n_1 = 10$, $\sigma_1^2 = 9$, $n_2 = 8$ and $\sigma_2^2 = 5$. Seeking the probability that the sample variances satisfy the inequality $s_2^2 > 2\,s_1^2$ now means

$$F < \frac{1}{3.70} \approx 0.270.$$

As shown in Fig. 7.3 (right), this probability is also obtained by integrating the density of the F distribution, but with its degrees of freedom swapped:

$$\int_0^{0.270} f_F(x;\, \nu_1 = 9,\, \nu_2 = 7)\, \mathrm{d}x \approx 0.036,$$

which is the same as before. ◁

7.3 Confidence Intervals

Next to (7.11) we wrote: "The error is, perhaps, $\pm s_x/\sqrt{n}$." What exactly does that mean? The full circles in Fig. 7.1 denoting the sample averages are displaced by more than their uncertainty, i.e. by more than $\pm s_x/\sqrt{n}$! Obviously we need a more quantitative measure for "perhaps". It is offered by a criterion called the *confidence interval*.

7.3.1 Confidence Interval for Sample Mean

Let μ_T and σ_T^2 be the mean and variance of the sample distribution of some statistic T, e.g. $T = X$ or $T = \sum_i X_i$. If the sample statistic is approximately normal—which applies to many statistics if the sample size is at least a few times 10—we expect that the value of T will be on the interval $[\mu_T - \sigma_T, \mu_T + \sigma_T]$ approximately 68.3% of the time, on $[\mu_T - 2\sigma_T, \mu_T + 2\sigma_T]$ about 95.5% of the time, on $[\mu_T - 3\sigma_T, \mu_T + 3\sigma_T]$ roughly 99.7% of the time, and so on (see (3.13)). We say: with *confidence level* (CL) 68.3% we may be confident (we trust, believe, anticipate), that T will be found on the interval $[\mu_T - \sigma_T, \mu_T + \sigma_T]$, and analogously for the others. Such an interval is called the *confidence interval*.

Suppose we have a sample $\{x_i\}_{i=1}^n$ or n independent observations for which we have already determined the sample mean $\overline{x}$ and variance s_x^2 in *unbiased* form (7.10). To determine how well $\overline{x}$ estimates the true population mean, μ, we first form the statistic

$$T = \frac{\overline{X} - \mu}{s_X}\sqrt{n}. \tag{7.18}$$

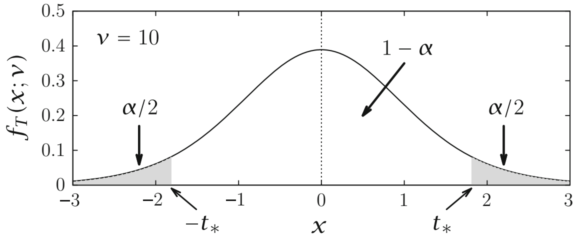


Fig. 7.4 The relation between the confidence level $1-\alpha$ and critical values $\pm t_*$ for the determination of the confidence interval $[-t_*, t_*]$ for the sample mean. (Example with $\nu = 10$.)

Table 7.1 Critical values t_* for the Student distribution with $\nu = 10$, 20 and 30 degrees of freedom for a few commonly used confidence levels CL $= 1-\alpha$

CL = $1-\alpha$	50%	68.26%	90%	95%	95.45%	99%	99.73%
t_* ($\nu = 10$)	0.700	1.053	1.812	2.228	2.290	3.169	3.892
t_* ($\nu = 20$)	0.687	1.026	1.724	2.086	2.139	2.845	3.376
t_* ($\nu = 30$)	0.683	1.018	1.697	2.042	2.092	2.750	3.230
z_*	0.675	1.000	1.645	1.960	2.000	2.576	3.000

The last line contains the critical values for the normal distribution, $z_* = t_*(\nu \to \infty)$

If X_i are normally distributed as $N(\mu, \sigma^2)$, the T statistic is distributed according to the Student distribution (3.22) with $\nu = n-1$ degrees of freedom. The integral of the density $f_T(x;\nu)$ determines the boundaries of the interval $[-t_*, t_*]$ on which the values of t or the corresponding mean μ are expected with the pre-specified probability (confidence level) $1-\alpha$, while there is a probability (risk level) α that t will be outside of it: see Fig. 7.4.

Then

$$-t_* \le \frac{\overline{x}-\mu}{s_x}\sqrt{n} \le t_*$$

or

$$\mu \in \left[\overline{x} - \frac{t_* s_x}{\sqrt{n}}, \overline{x} + \frac{t_* s_x}{\sqrt{n}}\right], \tag{7.19}$$

meaning: “The true mean of a large population, from which a sample $\{x_i\}_{i=1}^n$ has been obtained, is estimated as $\mu = \overline{x}$, and the confidence interval (7.19) contains μ with probability $1-\alpha$.” For large samples ($n \gtrsim 30$) the Student distribution is practically identical to the standardized normal, and the corresponding bounds t_* are simply the bounds in the Gauss curve (Table 7.1). Understandably, t_* increases with increasing confidence level: a broader interval implies less “risk”.

Example An eleven-fold ($n = 11$) measurement of a particle's mass yielded a mean of $\overline{m} = 4.180\,\mathrm{GeV}/c^2$ and an unbiased estimate for the standard deviation $s_m = 0.060\,\mathrm{GeV}/c^2$. What is the confidence interval on which the true mass of the particle μ may be expected with a confidence level of $1 - \alpha = 0.90$?

If the observations are normally distributed, the variable $T = (\overline{M} - \mu)\sqrt{n}/s_M$ is Student-distributed, with $\nu = n - 1 = 10$ degrees of freedom. Table 7.1, first row, $1 - \alpha = 0.90$, gives the critical value $t_* = 1.812$, shown also in Fig. 7.4. The requested confidence interval for μ is therefore

$$\left[\overline{m} - \frac{t_* s_m}{\sqrt{n}}, \overline{m} + \frac{t_* s_m}{\sqrt{n}}\right] = \big[4.147, 4.213\big]\,\mathrm{GeV}/c^2.$$

Because the Student distribution is symmetric about the origin, the equation

$$P\big(-t_* \le T \le t_*\big) = 1 - \alpha$$

defines the same t_* as the equation

$$P\big(T \le t_*\big) = 1 - \tfrac{1}{2}\,\alpha,$$

so t_* can also be determined by using the table of quantiles of the Student distribution. For purposes of this Problem ($1 - \alpha = 0.90$) we need the 95. quantile, located in the tenth row of Table D.4 in the $p = 0.95$ column, whence we again read off $t_* = t_{0.95} = 1.81$. ◁

Example The closing time x of safety valves is measured by an imprecise device reporting values with a standard deviation s_x which we know is near the value of $\sigma = 40\,\mathrm{ms}$. How many valves should we test, at confidence level $1 - \alpha = 99\%$, in order to determine the mean closing time μ to a precision of $\Delta x = 10\,\mathrm{ms}$?

Let us assume that a large sample ($n \gg 1$) will be required, so that the Student distribution can be replaced by the normal (i.e. Student in the $n \to \infty$ limit). From the last row of Table D.4, in the $1 - \alpha/2 = 0.995$ column, we get $t_* = t_{0.995} = 2.58$, corresponding to the confidence interval $\big[\overline{x} - t_* s_x/\sqrt{n}, \overline{x} + t_* s_x/\sqrt{n}\big]$, hence the population mean is determined to a precision given by

$$\mu = \overline{x} \pm \Delta x = \overline{x} \pm t_* \frac{s_x}{\sqrt{n}}.$$

The problem is asking for $t_* s_x/\sqrt{n} \le \Delta x$, whence

$$n \ge \left(\frac{t_* s_x}{\Delta x}\right)^2 \approx 106.$$

We see that the normal approximation is justified. ◁

7.3.2 Confidence Interval for Sample Variance

From Sect. 7.2.2 (see (7.16)) we know that the variable ns_X^2/σ^2 is χ^2-distributed, with $\nu = n - 1$ degrees of freedom, so we can immediately write down the confidence interval for the sample variance. Take a confidence level of $1 - \alpha = 0.90$, for example, so that the critical values of χ^2 are $\chi^2_{0.05}$ and $\chi^2_{0.95}$. So the variable ns_X^2/σ^2 is bounded as $\chi^2_{0.05} \le ns_X^2/\sigma^2 \le \chi^2_{0.95}$. The population effective deviation σ can therefore be bounded by the sample effective deviation s_X as

$$\frac{s_X\sqrt{n}}{\sqrt{\chi^2_{0.95}}} \le \sigma \le \frac{s_X\sqrt{n}}{\sqrt{\chi^2_{0.05}}}. \tag{7.20}$$

Figure 7.2 (right) shows the $\nu = 8$ case, with critical values $\chi^2_{0.05} \approx 2.73$ and $\chi^2_{0.95} \approx 15.5$ (eighth row of Table D.3). Note that the bounds are not symmetric with respect to the distribution average! See also Problem 7.6.4.

7.3.3 Confidence Region for Sample Mean and Variance

Suppose we were to use the sample $\{x_i\}_{i=1}^n$ to *simultaneously* locate, with chosen confidence level (probability $1-\alpha$) both the true mean μ *and* the true variance σ^2. By doing this, we would identify something we call a *confidence region*. If the population is normally distributed like $N(\mu, \sigma^2)$, the variables $\overline{X}$ and s_X^2 are independent, which can be proven by using characteristic functions. A confidence region at $\mathrm{CL} = 1 - \alpha$ is then obtained by simultaneous requirements

$$P_1 = P\left(-t_* \le \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \le t_*\right) = \sqrt{1 - \alpha}$$

and

$$P_2 = P\left(\chi^2_\downarrow \le \frac{ns_X^2}{\sigma^2} \le \chi^2_\uparrow\right) = \sqrt{1 - \alpha},$$

where $\pm t_*$ are the symmetric bounds in the density of the t distribution, while $\chi^2_\downarrow$ and $\chi^2_\uparrow$ are the lower and upper bounds in the density of the χ^2 distribution. The confidence region is then defined by the equation $P_1 P_2 = 1 - \alpha$, that is,

$$P\left(-t_* \le \frac{\overline{X} - \mu}{\sigma}\sqrt{n} \le t_*,\ \chi^2_\downarrow \le \frac{ns_X^2}{\sigma^2} \le \chi^2_\uparrow\right) = 1 - \alpha. \tag{7.21}$$

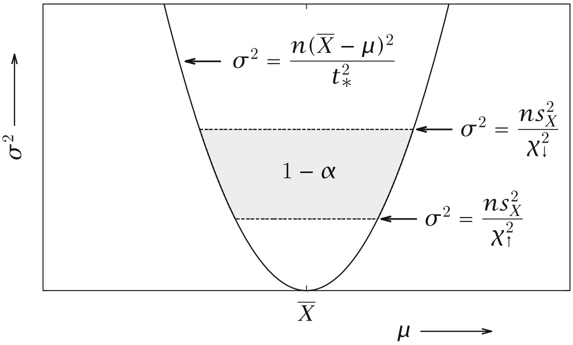

$\sigma^2 = \dfrac{n(\overline{X}-\mu)^2}{t_*^2}$
$\sigma^2 = \dfrac{ns_X^2}{\chi_\downarrow^2}$
$1-\alpha$
$\sigma^2 = \dfrac{ns_X^2}{\chi_\uparrow^2}$
σ^2
$\overline{X}$
μ

Fig. 7.5 Joint confidence region for sample mean and variance of a normally distributed population, defined by parameters t_*, $\chi_\downarrow^2$ and $\chi_\uparrow^2$ in (7.21)

An example of such a region is shown in Fig. 7.5. (We could have chosen a different sharing of $(1-\alpha)$ between the mean and the variance; the shaded area would be correspondingly narrower and taller or wider and shorter.)

7.4 Outliers and Robust Measures of Mean and Variance

Occasionally a sample contains values which obviously differ from the bulk of the sample. They are called *outliers*. Outliers may hint at an error in the experiment or may represent genuine measurements that happen to strongly deviate from the majority of observations. The ozone hole over Antarctica, for example, has been indicated by peculiar recordings by the Nimbus 7 satellite, but they were wrongly attributed to instrumental errors [2].

To determine the parameters that characterize the samples with relatively small shares of outliers, we use so-called *robust measures* and robust statistics [3]. Among other things, "robustness" implies a small sensitivity of estimates of mean and variance to the inclusion or exclusion of individual or all outliers from the sample.

Example The classical motivational case for the application of robust methods is the set of 24 measurements of copper content in bread flour [4]:

$$\begin{array}{cccccccccccc} 2.20 & 2.20 & 2.40 & 2.40 & 2.50 & 2.70 & 2.80 & 2.90 & 3.03 & 3.03 & 3.10 & 3.37 \\ 3.40 & 3.40 & 3.40 & 3.50 & 3.60 & 3.70 & 3.70 & 3.70 & 3.70 & 3.77 & 5.28 & 28.95, \end{array}$$

shown in Fig. 7.6. The arithmetic average of the whole sample is $\overline{x} = 4.28$ and the standard deviation is $s_x = 5.30$. If the value $x_{24} = 28.95$ is excluded from the sample, we get $\overline{x} = 3.21$ and $s_x = 0.69$. Clearly a single outlier may strongly modify both $\overline{x}$ and s_x, so neither $\overline{X}$ nor s_X are suitable as robust estimators of population properties. ◁

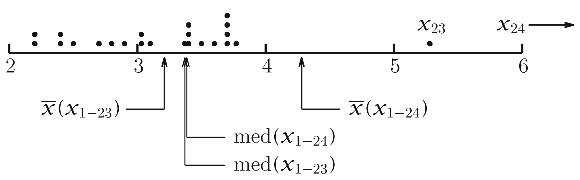


Fig. 7.6 A sample (24 values) of copper content in flour, in units of μg/g. The value $x_{24} = 28.95$ (and potentially also $x_{23} = 5.28$) is an outlier. The median is much less sensitive to the exclusion of the rightmost outlier x_{24} than the arithmetic average

A much more robust measure of the "center" of the sample $\boldsymbol{x} = \{x_i\}_{i=1}^n$ is the median. It is defined as the value (see (4.4)) which splits an ordered sample ($x_i \le x_{i+1}$) such that half of the observations are to its left and half are to its right,

$$\mathrm{med}[X] = \begin{cases} x_{(n+1)/2} & ;\ n\ \text{ odd}, \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & ;\ n\ \text{ even}. \end{cases} \tag{7.22}$$

The scattering of the data about the median can be quantified by the *median absolute deviation* (MAD)

$$\mathrm{MAD}[X] = \mathrm{med}\big[|X - \mathbf{1}_n \mathrm{med}[X]|\big], \tag{7.23}$$

where $\mathbf{1}_n = \{1, 1, \ldots, 1\}$ is a sequence of ones with length n. To be able to compare this to the standard deviation, one frequently uses the quantity

$$\mathrm{MADN}[X] \approx 1.4826\,\mathrm{MAD}[X],$$

where the factor 1.4826 is chosen such that for a normal $N(\mu, \sigma^2)$ distribution one has MADN $= \sigma$. The values for the whole flour sample are med$[X] = 3.385$ and MADN $= 0.526$, while if x_{24} is excluded, one gets 3.37 and 0.504. Both values change only insignificantly when the outlier is excluded (see Fig. 7.6).

7.4.1 Chasing Outliers

A straightforward way to exclude outliers is the "3σ"-rule. By following it we may decide to eliminate all observations deviating from the sample mean $\overline{x}$ by more than $\pm 3s_x$, or simply assign them the values $\overline{x} \pm 3s_x$. The method has several flaws. Among others, it forces us to needlessly remove, on average, three observations from an immaculate, normally distributed sample of size $n = 1000$, since the interval $[\overline{x} - 3s_x, \overline{x} + 3s_x]$ for large n contains 99.7% of the data. Besides, the calculation of the mean and variance itself is highly sensitive to outliers. It is therefore preferable to use the criterion

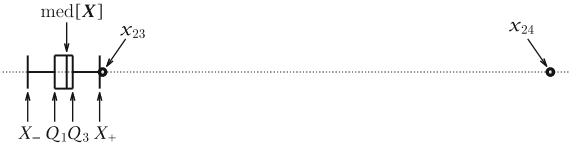


Fig. 7.7 Box diagram used to identify outlier candidates. Outliers may be expected outside of the interval $[X_-, X_+] = [Q_1 - \frac{3}{2}\,\mathrm{IQR}, Q_3 + \frac{3}{2}\,\mathrm{IQR}]$

$$x_i \text{ outlier} \iff \left|\frac{x_i - \mathrm{med}[X]}{\mathrm{MADN}[X]}\right| > 3.5.$$

A simple method to visually identify candidates for outliers is to draw a *box diagram*. One first calculates the sample median and, by a generalization of (7.22), its first and third quartile, Q_1 and Q_3: the sample is then divided into four compartments with a quarter of observations in each. One then calculates the *inter-quartile range* (IQR) and the bounds X_- and X_+, beyond which outliers are expected to appear,

$$\mathrm{IQR} = Q_3 - Q_1, \qquad X_- = Q_1 - \tfrac{3}{2}\,\mathrm{IQR}, \qquad X_+ = Q_3 + \tfrac{3}{2}\,\mathrm{IQR}.$$

This method identifies the values x_{23} and x_{24} in the flour sample as outliers (see Fig. 7.7). The interval $[X_-, X_+]$ for large n and normal distribution contains 99.3% of observations, making the method roughly equivalent to the "3σ"-rule.

7.4.2 Distribution of Sample Median (and Sample Quantiles)

There is a theorem on the distribution of sample quantiles, whose special case is the median. Let X be a continuous random variable with the probability density f_X and distribution function F_X. Let $x_{(p)}$ denote the pth quantile of X, so that $F_X\big(x_{(p)}\big) = p$, and $\widetilde{x}_{(p)}$ the sample quantile determined from the sample $\{x_1, x_2, \ldots, x_n\}$. In the limit of large samples ($n \gg 1$) it holds that [5]

$$\sqrt{n}\,\big(\widetilde{x}_{(p)} - x_{(p)}\big) \sim N\left(0\,, \frac{p(1-p)}{f_X^2\big(x_{(p)}\big)}\right).$$

Hence the sample median ($p = 0.5$) is asymptotically normally distributed with mean $x_{(0.5)}$ and variance $1/[4nf_X^2(x_{(0.5)})]$. The variance depends on the density f_X! If the population is normally distributed according to $N(\mu, \sigma^2)$, we have $x_{(p)} = \mu$ and $1/[4nf_X^2(x_{(p)})] = \pi\sigma^2/(2n)$. Therefore

$$\mathrm{med}[X] \sim N\left(\mu\,, \frac{\pi\sigma^2}{2n}\right), \qquad n \gg 1. \tag{7.24}$$

7.5 Sample Correlation

In this section we introduce measures of correlation between data sets. The correlation strength is measured by correlation coefficients. Suitable statistics are used to determine whether observed correlations are statistically significant.

7.5.1 Linear (Pearson) Correlation

The basic measure for the strength of correlation between two data sets is the linear correlation coefficient $\widehat{\rho}$. Correlations in two-dimensional data sets can often be simply "seen": characteristic patterns in the (x_i, y_i) plane for correlation coefficients $\widehat{\rho} \approx 1$ (almost complete positive correlation), $\widehat{\rho} \approx -1$ (nearly total negative correlation or anti-correlation) and $\widehat{\rho} \approx 0$ (roughly uncorrelated observations) are shown in Fig. 7.8.

The linear correlation coefficient between the data sets $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ (the estimate of the true ρ) is

$$\widehat{\rho} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}, \qquad -1 \le \widehat{\rho} \le 1. \tag{7.25}$$

The coefficient (7.25) is suitable to estimate the true correlation strength once we have figured out that a correlation exists and has a certain statistical significance.

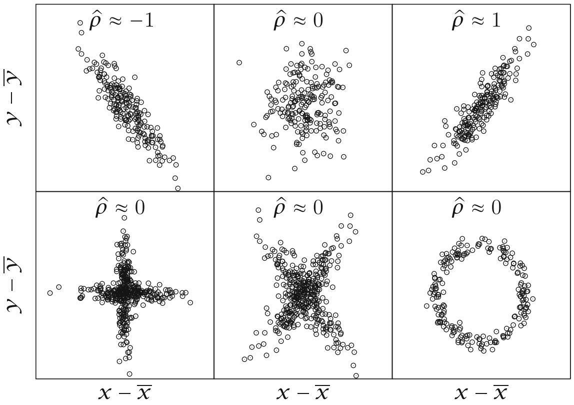


Fig. 7.8 Sets of observations (x_i, y_i) that one can describe by a two-dimensional distribution of X and Y, and corresponding estimates of the sample correlation coefficient $\widehat{\rho}$. [Top, right to left] Almost completely correlated ($\widehat{\rho} \approx 1$), uncorrelated ($\widehat{\rho} \approx 0$) and nearly anti-correlated data ($\widehat{\rho} \approx -1$). [Bottom] Three cases of realizations of uncorrelated variables which are *not* statistically independent. The requirement $\widehat{\rho} = 0$ is just a *necessary* condition for statistical independence

As in the case of the sample mean and average, the sample correlation coefficient $\widehat{\rho}$ can also be endowed with confidence intervals. To do this, one uses the Fisher transformation of $\widehat{\rho}$, namely, the statistic

$$Z = \frac{1}{2} \log \frac{1+\widehat{\rho}}{1-\widehat{\rho}} = \operatorname{Atanh} \widehat{\rho},$$

and assume that the observations x_i and y_i obey the joint binormal (two-dimensional normal) distribution. For sample sizes n of at least a few times 10 the Z statistic is approximately normally distributed, i.e.

$$Z \sim N\left(\overline{Z}, \sigma_Z^2\right) = N\left(\frac{1}{2}\left[\log \frac{1+\rho}{1-\rho} + \frac{\rho}{n-1}\right], \frac{1}{n-3}\right),$$

where ρ is the true correlation coefficient. (It turns out that correlation coefficients $\widehat{\rho}$ of samples drawn from normally distributed populations are smaller than their population counterparts ρ, hence biased. The $\rho/(2(n-1))$ helps to approximately cancel that bias.) The best estimate for ρ is then $\rho = \widehat{\rho}$, while the risk level (significance) at which one may claim that the measured $\widehat{\rho}$ differs from ρ, is given by

$$\alpha = 1 - \operatorname{erf}\left(\frac{|z-\overline{z}|\sqrt{n-3}}{\sqrt{2}}\right).$$

To determine whether the observations of x_i and y_i under conditions "1" and "2" exhibit different correlations, one compares the correlation coefficients $\widehat{\rho}_1$ and $\widehat{\rho}_2$. The statistical significance of the observed difference between $\widehat{\rho}_1$ and $\widehat{\rho}_2$ is

$$\alpha = 1 - \operatorname{erf}\left(\frac{|z_1 - z_2|}{\sqrt{2}} \sqrt{\frac{(n_1-3)(n_2-3)}{n_1+n_2-6}}\right).$$

The reverse question is: to what confidence interval $[\rho_-, \rho_+]$ can the correlation coefficient be restricted, given a confidence level of $1-\alpha$? For the commonly used $1-\alpha \approx 96\%$ the values ρ_- and ρ_+ are given by

$$\rho_- = \tanh\left(\operatorname{Atanh}\widehat{\rho} - \frac{2}{\sqrt{n}}\right), \qquad \rho_+ = \tanh\left(\operatorname{Atanh}\widehat{\rho} + \frac{2}{\sqrt{n}}\right).$$

7.5.2 *Non-parametric (Spearman) Correlation*

The linear correlation coefficient formula (7.25) contains the sample means $\overline{x}$ and $\overline{y}$, which are strongly sensitive to outliers (see Sect. 7.4). A more robust tool is called for, and one option is to define the correlation by referring to the positions (ranks)

r_i and s_i that individual x_i and y_i occupy in the ordered samples $\boldsymbol{x}$ and $\boldsymbol{y}$. The ranks are counted from 1 upwards. When several (e.g. m) equal values share m positions, they are all assigned the average rank which they would have if they differed by an infinitesimal amount. One also computes the average ranks $\overline{r} = (\sum_{i=1}^{n} r_i)/n$ and $\overline{s} = (\sum_{i=1}^{n} s_i)/n$.

Example Let us determine the rank of the sample $\{x_i\}_{i=1}^{8} = \{2, 3, 9, 3, 4, 9, 7, 3\}$. We first order the sample, obtaining $\{x_1, x_2, x_4, x_8, x_5, x_7, x_3, x_6\}$. The values $x_2 = x_4 = x_8 = 3$ share ranks 2 to 4, so their average rank is $(2 + 3 + 4)/3 = 3$. The values $x_3 = x_6 = 9$ share ranks 7 and 8, so their rank is 7.5. Therefore $\{r_i\}_{i=1}^{8} = \{1, 3, 3, 3, 5, 6, 7.5, 7.5\}$ and the average rank is $\overline{r} = 4.5$. ◁

By using the ranks r_i and s_i as well as the average ranks $\overline{r}$ and $\overline{s}$ we define the *rank correlation coefficient*

$$\widehat{\rho}_{\mathrm{r}} = \frac{\sum_{i=1}^{n}(r_i - \overline{r})(s_i - \overline{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \overline{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \overline{s})^2}} = 1 - \frac{6}{n(n^2 - 1)}\sum_{i=1}^{n}(r_i - s_i)^2. \quad (7.26)$$

To calculate $\widehat{\rho}_{\mathrm{r}}$ one refers only to the mutual position of the observations, so this kind of correlation estimation is called non-parametric. The distribution of ranked observations is uniform, and if there are just a few duplicates, the estimate (7.26) is much more robust than (7.25). The statistical significance of the measured coefficient $\widehat{\rho}_{\mathrm{r}}$ is determined by the t-test (details in Chap. 10). We form the statistic

$$t_{\mathrm{r}} = \widehat{\rho}_{\mathrm{r}}\sqrt{\frac{n-2}{1-\widehat{\rho}_{\mathrm{r}}^2}},$$

which is distributed approximately according to the Student distribution with $n - 2$ degrees of freedom. The confidence level $1 - \alpha$ (statistical significance α), at which one can reject the hypothesis that the measured correlation coefficient $\widehat{\rho}_{\mathrm{r}}$ equals the true coefficient ρ_{r}, is calculated from

$$1 - \alpha = \int_{-|t_{\mathrm{r}}|}^{|t_{\mathrm{r}}|} f_T(x;\nu)\,\mathrm{d}x = 1 - \frac{B_x(\nu/2, 1/2)}{B(\nu/2, 1/2)}, \qquad x = \frac{\nu}{\nu + t_{\mathrm{r}}^2}, \qquad \nu = n - 2,$$

where f_T is the probability density of the t distribution (see (3.22)), while $B(a, b)$ and $B_x(a, b)$ are the complete and incomplete beta functions.

7.6 Problems

7.6.1 Estimator of Third Moment

Find an unbiased estimator of the third distribution moment, $M_3 = E\big[\big(X - \overline{X}\big)^3\big]$, based on the sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$! This Problem offers a parallel to the biased and unbiased estimators of the population variance (formulas (7.8) and (7.10)).

✎ We split

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\overline{X})^3 &= \sum_i\big((X_i-\mu)-(\overline{X}-\mu)\big)^3\\ &= \underbrace{\sum_i(X_i-\mu)^3}_{T_1} - 3\underbrace{\sum_i(X_i-\mu)^2(\overline{X}-\mu)}_{T_2} + 3\underbrace{\sum_i(X_i-\mu)(\overline{X}-\mu)^2}_{T_3} - \underbrace{\sum_i(\overline{X}-\mu)^3}_{T_4},\end{aligned}$$

and calculate the expected values term by term:

$$\begin{aligned}E\big[T_1\big] &= \sum_i E\Big[\big(X_i-\mu\big)^3\Big] = nM_3,\\ E\big[T_2\big] &= E\left[\left(\sum_i(X_i-\mu)^2\right)\left(\frac{1}{n}\sum_i(X_i-\mu)\right)\right] = M_3,\\ E\big[T_3\big] &= E\left[\left(\sum_i(X_i-\mu)\right)\left(\frac{1}{n}\sum_i(X_i-\mu)\right)^2\right] = \frac{1}{n}M_3,\\ E\big[T_4\big] &= nE\left[\left(\frac{1}{n}\sum_i(X_i-\mu)\right)^3\right] = \frac{1}{n}M_3.\end{aligned}$$

Therefore

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\big(X_i-\overline{X}\big)^3\right] = \frac{1}{n}\left(nM_3 - 3M_3 + \frac{3}{n}M_3 - \frac{1}{n}M_3\right) = \frac{(n-1)(n-2)}{n^2}M_3.$$

The weighted sum at the left—surely the first form to cross one's mind—obviously results in a biased estimator. An unbiased estimator of the third moment is

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\big(X_i-\overline{X}\big)^3.$$

7.6.2 Unbiasedness of Poisson Variable Estimators

In this example [6] we realize that unbiasedness is not the holy grail to which one should strive at all costs. (See also [7].) Let the variable X be Poisson-distributed, with parameter λ, thus $P(X = x) = \lambda^x \mathrm{e}^{-\lambda}/x!$. Find an unbiased estimator for $(P(X = 0))^2 = \mathrm{e}^{-2\lambda}$!

✎ Any unbiased estimator $T(X)$ for the given quantity must satisfy the equation

$$E\big[T(X)\big] = \sum_{x=0}^{\infty} T(x)P(X = x) = \sum_{x=0}^{\infty} T(x)\frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} = \mathrm{e}^{-2\lambda}, \qquad \lambda \geq 0.$$

But the only option is $T(x) = (-1)^x$, since then $\mathrm{e}^{-\lambda}\sum_{x=0}^{\infty}(-1)^x\lambda^x/x! = \mathrm{e}^{-\lambda}\mathrm{e}^{-\lambda} = \mathrm{e}^{-2\lambda}$. What does that mean? If, for example, a single observation yields $x = 200$, we can reasonably conclude that the true $\mathrm{e}^{-2\lambda}$ is virtually zero, while the unbiasedness requirement forces us to accept the value $(-1)^{200} = 1$ as the estimate of $\mathrm{e}^{-2\lambda}$. On the other hand, if, for instance, we observe $x = 3$, the unbiased estimate is supposed to be $(-1)^3 = -1$, which is a negative value for a random quantity whose values are always on $(0, 1]$. A better estimator for $\mathrm{e}^{-2\lambda}$ is certainly e^{-2X}, even though it is biased.

7.6.3 Concentration of Mercury in Fish

(Adapted from [8]). The following average concentrations of mercury (µg per g of body weight) in fish from $n = 48$ Florida lakes are available:

$$\begin{aligned}\boldsymbol{x} = \{&1.23,\ 1.33,\ 0.04,\ 0.04,\ 1.20,\ 0.27,\ 0.49,\ 0.19,\ 0.83,\ 0.81,\ 0.71,\ 0.50,\\ &0.49,\ 1.16,\ 0.05,\ 0.15,\ 0.19,\ 0.77,\ 1.08,\ 0.98,\ 0.63,\ 0.56,\ 0.41,\ 0.73,\\ &0.59,\ 0.34,\ 0.84,\ 0.50,\ 0.34,\ 0.28,\ 0.34,\ 0.87,\ 0.56,\ 0.17,\ 0.18,\ 0.19,\\ &0.04,\ 0.49,\ 1.10,\ 0.16,\ 0.10,\ 0.21,\ 0.86,\ 0.52,\ 0.65,\ 0.27,\ 0.94,\ 0.37\}.\end{aligned}$$

The histogram of the observations is shown in Fig. 7.9 (left). Based on this sample, find the 95% confidence interval for the average concentration μ in the whole fish population!

✎ Does perhaps the sample itself indicate "normality"? A good tool to answer this question is a graph containing the ordered observations $x_1 \leq x_2 \leq \ldots \leq x_n$ on the abscissa and the variable z_i corresponding to the $(i - 0.5)/n$th quantile ξ_i of the standardized normal distribution on the ordinate axis. How is the graph constructed? When observations are sorted, the smallest observation is $x_1 = 0.04$, hence $\xi_1 = (1 - 0.5)/48 \approx 0.01042$ and $\Phi(z_1) = \xi_1$, where $\Phi(z) = \frac{1}{2}\big[1 + \mathrm{erf}(z/\sqrt{2})\big]$

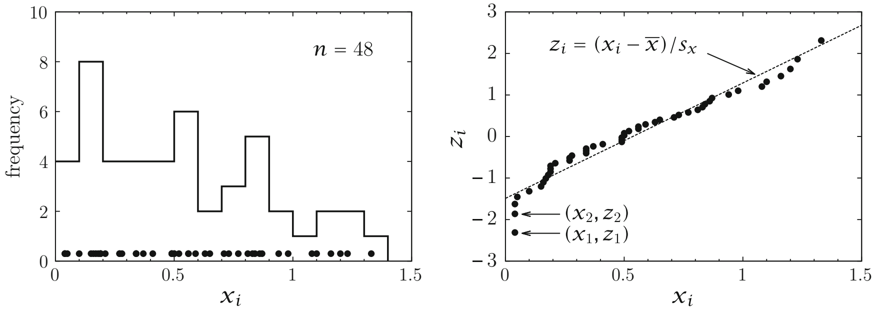


Fig. 7.9 [Left] Average mercury concentrations in fish from $n = 48$ Florida lakes. [Right] Graph of standardized variables z_i as functions of sorted observations x_i, the so called "Q–Q plot" showing the quantiles of two distributions on the respective axes

is the distribution function of the standardized normal distribution. Then $z_1 = \sqrt{2}\,\mathrm{erf}^{-1}\big[2\xi_1 - 1\big] \approx -2.311$. The pair (x_1, z_1) is the point at the extreme bottom left in Fig. 7.9 (right). The remaining points (x_2, z_2), (x_3, z_3), ... are calculated in the same manner: we plot

$$z_i = \sqrt{2}\,\mathrm{erf}^{-1}\left[2\left(\frac{i-0.5}{n}\right) - 1\right] \text{ as function of } x_i, \qquad x_1 \le x_2 \le \ldots \le x_n.$$

If the sample is normally distributed, the z_i vs. x_i graph is a straight line.

In our case the sample exhibits a distribution which does not appear to be normal: its distribution function increases faster than the normal distribution function at small x_i, and slower at large x_i. This indicates that the underlying distribution is positively skewed (see Fig. 4.4), which one can also infer from the histogram in Fig. 7.9 (left).

The confidence interval for the population mean could be calculated by (7.19), but it only applies to normally distributed X_i. However, due to the central limit theorem the mean $\overline{X}$ is approximately normally distributed at large n *regardless of the distribution of* X_i, with mean μ and variance $\sigma^2/n \approx s_X^2/n$. In our case $n = 48 \gg 1$, so (7.19) may be used nonetheless. From the sample we compute $\overline{x} \approx 0.536$ and $s_x \approx 0.360$, then use the last row of Table 7.1 at $\mathrm{CL} = 1 - \alpha = 95\%$ to obtain the critical $z_* = 1.960$. Therefore, μ can be bounded as

$$\overline{x} - z_* \frac{s_x}{\sqrt{n}} \le \mu \le \overline{x} + z_* \frac{s_x}{\sqrt{n}},$$

which amounts to $0.435 \le \mu \le 0.638$.

7.6.4 Dosage of Active Ingredient

The mass of the active pharmaceutical ingredient in pills is distributed about the known average value: in a sample of $n = 20$ pills taken for analysis we find a variance of $s_x^2 = 0.12\,\mathrm{mg}^2$ ($s_x = 0.346\,\mathrm{mg}$). Find the 80% confidence interval for the true (population) standard deviation of the active ingredient mass (a 10% chance of it being too small and a 10% chance of it being too large)!

✎ The confidence interval for the population variance is given by formula (7.20). From Table D.3 for $\nu = n - 1 = 19$ we read off the critical values of the χ^2 distribution, $\chi^2_{0.10} = 11.7$ and $\chi^2_{0.90} = 27.2$, so σ can be bounded as

$$\sqrt{ns_x^2/\chi^2_{0.90}} \le \sigma \le \sqrt{ns_x^2/\chi^2_{0.10}}$$

or $0.297\,\mathrm{mg} \le \sigma \le 0.453\,\mathrm{mg}$. Note that s_x does not lie in the middle of this interval, as we already know from Sect. 7.3.2 (Fig. 7.2).

References

1. M.R. Spiegel, J. Schiller, R.A. Srinivasan, *Theory and Problems of Probability and Statistics*, 4th edn. (McGraw-Hill, New York, 2012)
2. R. Kandel, *Our Changing Climate* (McGraw-Hill, New York, 1991)
3. L. Davies, U. Gather, Robust statistics, Chap. III.9, in *Handbook of computational statistics. Concepts and methods*, ed. by J.E. Gentle, W. Härdle, Y. Mori (Springer, Berlin, 2004), pp. 655–695
4. Analytical Methods Committee. Robust statistics – how not to reject outliers, Part 1: basic concepts. Analyst **114**, 1693 (1989), Part 2: Inter-laboratory trials. Analyst **114**, 1699 (1989)
5. A.M. Walker, A note on the asymptotic distribution of sample quantiles. J. R. Stat. Soc. B **30**, 570 (1968)
6. J.P. Romano, A.F. Siegel, *Counterexamples in Probability and Statistics* (Wadsworth & Brooks/Cole, Monterey, 1986)
7. M. Hardy, An illuminating counterexample. Am. Math. Mon. **110**, 234 (2003)
8. T.R. Lange, H.E. Royals, L.L. Connor, Influence of water chemistry on mercury concentration in largemouth bass from Florida lakes. Trans. Am. Fish. Soc. **122**, 74 (1993)

Chapter 8
Maximum-Likelihood Method

Abstract The maximum-likelihood method offers a possibility to devise estimators of unknown population parameters by circumventing the calculation of expected values like average, variance and higher moments. The likelihood function is defined and its role in formulating the principle of maximum likelihood is elucidated. The variance and efficiency of maximum-likelihood estimators is discussed, in particular in the light of its information content and possible minimum variance bound. Likelihood intervals are introduced by analogy to the confidence intervals used in standard sample-based inference. The method is extended to the case when several parameters are determined simultaneously, and to likelihood regions as generalizations of likelihood intervals.

In this chapter we discuss the possibility to devise an estimator for the unknown population parameter θ without resorting to the calculation of expected values like average, variance and higher moments (Chap. 7).

8.1 Likelihood Function

When a continuous or discrete random variable X is measured n-times (or a sample of size n is drawn from an infinite population) one obtains a set of values $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$. Assume that X is distributed according to the probability density or probability function $f_X(x; \boldsymbol{\theta})$, where

$$\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_p\} \tag{8.1}$$

are unknown parameters. Let us consider continuous variables only; the discussion of discrete variables follows the same pattern. The probability that, at given parameters $\boldsymbol{\theta}$, *just* the values x_i on intervals $[x_i, x_i + \mathrm{d}x]$ have been observed, is $\mathrm{d}P = \prod_{i=1}^{n} f_X\big(x_i; \boldsymbol{\theta}\big)\, \mathrm{d}\boldsymbol{x}$. The product of probability densities in this expression is called the *likelihood function*:

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_8

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = L(x_1, x_2, \ldots, x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(x_i; \boldsymbol{\theta}),$$

where the | sign is a warning that a joint conditional density (or probability function) is implied, since the values $\boldsymbol{x}$ have been observed given the "condition" $\boldsymbol{\theta}$. Note that the likelihood function depends on the sample and is therefore itself a random variable. We also define its logarithm, the *log-likelihood function*

$$\ell = \log L = \sum_{i=1}^{n} \log f_X(x_i; \boldsymbol{\theta}).$$

There are two good reasons for taking the log. Multiplying many $f_X(x_i; \boldsymbol{\theta})$ may result in floating-point underflow, which is avoided by turning the product into a sum. In addition, having a sum is convenient as we shall be taking the derivative of the likelihood function with respect to the parameters θ_i. Note also that the log is a monotonously increasing function with a singularity at the origin: this may be a source of numerical problems in seeking the global maximum of ℓ.

8.2 Principle of Maximum Likelihood

The *principle of maximum likelihood* states that the optimal value of the parameter θ is found by maximizing the likelihood function (or its logarithm) with respect to θ: such a measurement of $\boldsymbol{x}$ is then seen to be "most likely". We therefore wish to find such $\widehat{\theta}$ that for all possible θ it holds that $\ell\big(\boldsymbol{x}|\widehat{\theta}\big) \geq \ell\big(\boldsymbol{x}|\theta\big)$. Assume that the function ℓ is twice differentiable with respect to θ. The value $\widehat{\theta}$ is obtained by setting its first derivative to zero,

$$\ell' = \frac{\mathrm{d}\ell}{\mathrm{d}\theta} = \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_X(x_i; \theta) = \sum_{i=1}^{n} \frac{f_X'(x_i; \theta)}{f_X(x_i; \theta)} = 0, \qquad (8.2)$$

where $'$ denotes the derivative with respect to θ. This formula is known as the *likelihood equation*. The condition that we have indeed found the maximum, is

$$\ell'' = \frac{\mathrm{d}^2\ell}{\mathrm{d}\theta^2} = \sum_{i=1}^{n} \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f_X(x_i; \theta) < 0. \qquad (8.3)$$

Should we wish to determine several parameters (8.1) simultaneously, the likelihood equation needs to be solved for each parameter separately,

$$\frac{\partial}{\partial\theta_j}\ell(\boldsymbol{x}|\boldsymbol{\theta}) = 0, \qquad j = 1, 2, \ldots, p. \qquad (8.4)$$

By analogy to (8.3) we also identify the sufficient condition that the absolute maximum of ℓ has been found: the square matrix A with the elements

$$A_{ij}\big(\widehat{\boldsymbol{\theta}}\big) = -\left(\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \tag{8.5}$$

must be positive definite.

Example We have been measuring the same quantity with several devices with different uncertainties σ_i. The observations $\{x_1, x_2, \ldots, x_n\} = \boldsymbol{x}$ are scattered about the true value μ. Suppose that the fluctuations about the mean are normally distributed. The probability that, given the value of parameter μ, the observation x_i is on the interval $[x_i, x_i + \mathrm{d}x]$, is then

$$f_X(x_i;\, \mu)\, \mathrm{d}x = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right) \mathrm{d}x.$$

The corresponding likelihood function is

$$L(\boldsymbol{x}|\mu) = \prod_{i=1}^{n} f_X(x_i;\, \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right),$$

and its logarithm is

$$\ell = \log L(\boldsymbol{x}|\mu) = -\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma_i^2} - \sum_{i=1}^{n} \log\sigma_i + \text{const.} \tag{8.6}$$

By solving the likelihood equation

$$\frac{\mathrm{d}\ell}{\mathrm{d}\mu} = \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma_i^2} = 0$$

we obtain the estimate $\widehat{\mu}$ for the parameter μ:

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \qquad w_i = \frac{1}{\sigma_i^2}, \tag{8.7}$$

which is the familiar formula for the weighted average of x_i with normally distributed errors σ_i. The second derivative of ℓ with respect to μ reveals that $\widehat{\mu}$ indeed corresponds to the maximum ℓ, since $\mathrm{d}^2\ell/\mathrm{d}\mu^2 = -\sum_{i=1}^{n} w_i < 0$. ◁

Example Six vertical lines in Fig. 8.1 represent the observations $\{x_1, x_2, \ldots, x_6\}$, assumed to originate in a Cauchy-distributed population with the width parameter $s = 0.0001$ and unknown mean μ (see (3.20)). What is the maximum-likelihood estimate for the mean, $\widehat{\mu}$, based on this sample?

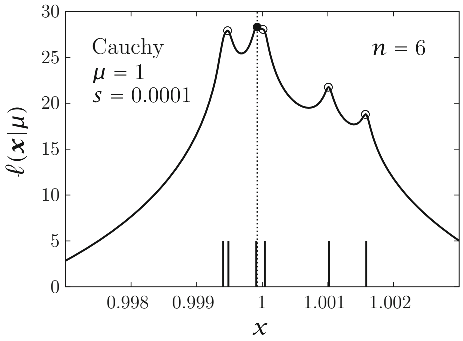


Fig. 8.1 Sample of six values from a Cauchy-distributed population with the width parameter $s = 0.0001$ and unknown mean μ. (The sample was in fact generated by a Cauchy random generator with $\mu = 1$ and $s = 0.0001$.) Also shown is the log-likelihood function with four local maxima (○) and the global maximum (•) at $\widehat{\mu} \approx 0.999922$

The log-likelihood function for a sample of size n is

$$\ell(\boldsymbol{x}|\mu) = \log\left(\prod_{i=1}^{n} \frac{1}{\pi} \frac{s}{s^2 + (x_i - \mu)^2}\right) = \sum_{i=1}^{n} \log\left(\frac{1}{\pi} \frac{s}{s^2 + (x_i - \mu)^2}\right),$$

and the likelihood equation is

$$\frac{\partial \ell}{\partial \mu} = 2 \sum_{i=1}^{n} \frac{x_i - \mu}{s^2 + (x_i - \mu)^2} = 0.$$

This can be written as $p(\mu) = 0$, where p is a polynomial of degree $2n - 1$. Thus, in general, the likelihood equation has $2n - 1$ solutions, some of which correspond to local maxima of ℓ. The optimal $\widehat{\mu}$ corresponds to the global maximum, which one usually finds numerically and tends to be near the sample median. In our case there are four local maxima and a global maximum at $\widehat{\mu} \approx 0.999922$. ◁

8.3 Variance of Estimator

Consistency and unbiasedness (see (7.2)–(7.4)) are not the only desirable properties of a statistical estimator. One would also like it to have as small a variance as possible. In general, different estimators of the same quantity have different variances: for example, both the sample mean and the sample median are consistent and unbiased estimators of the "center" of a population with known variance. Yet—as we shall see—the variance of the mean is smaller than the variance of the median, so the sample median is a "better" estimator.

8.3.1 Limit of Large Samples

For large samples ($n \gg 1$) one expects the estimate $\widehat{\theta}$ to be not very different from the true value of the parameter θ_0. We may therefore divide the likelihood equation (8.2) by n, expand it in a Taylor series and keep just the first two terms:

$$\frac{1}{n}\frac{\mathrm{d}\ell}{\mathrm{d}\theta} \approx \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\,\log f_X(x_i;\theta)\right]_{\theta_0}}_{\alpha(\boldsymbol{x};\,\theta_0)} + (\theta-\theta_0)\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2}{\partial\theta^2}\,\log f_X(x_i;\theta)\right]_{\theta_0}}_{\beta(\boldsymbol{x};\,\theta_0)} = 0$$

or, briefly,

$$\alpha + (\theta-\theta_0)\beta = 0. \tag{8.8}$$

The quantities α and β have a random nature, as they depend on the current sample $\boldsymbol{x}$. What are their expected values? For large n the sums can be replaced by integrals: the role of the sum weights $1/n$ are taken over by the probability densities. Then one can write

$$E[\alpha] = E\left[\frac{\partial\log f_X}{\partial\theta}\right]_{\theta_0} = \left(\int\frac{1}{f_X}\frac{\partial f_X}{\partial\theta}f_X\,\mathrm{d}x\right)_{\theta_0} = \left(\frac{\partial}{\partial\theta}\int f_X(x;\theta)\,\mathrm{d}x\right)_{\theta_0} = 0,$$

where we have considered the fact that the probability density f_X is normalized regardless of the value of its parameter, $\int f_X(x;\theta)\,\mathrm{d}x = 1$, and the derivative of a constant is zero. We play the same game with β:

$$\begin{aligned} E[\beta] = E\left[\frac{\partial^2\log f_X}{\partial\theta^2}\right]_{\theta_0} &= \left(\int\left[\frac{1}{f_X}\frac{\partial^2 f_X}{\partial\theta^2} - \frac{1}{f_X^2}\left(\frac{\partial f_X}{\partial\theta}\right)^2\right]f_X\,\mathrm{d}x\right)_{\theta_0} \\ &= \left(\frac{\partial^2}{\partial\theta^2}\int f_X(x;\theta)\,\mathrm{d}x\right)_{\theta_0} - \left(\int\left(\frac{\partial\log f_X}{\partial\theta}\right)^2 f_X(x;\theta)\,\mathrm{d}x\right)_{\theta_0}. \end{aligned}$$

The first term vanishes for the same reason as in $E[\alpha]$, while the second term is the (negative) expected value of the quantity $(\partial\log f_X/\partial\theta)^2$, so

$$E[\beta] = E\left[\frac{\partial^2\log f_X}{\partial\theta^2}\right]_{\theta_0} = -E\left[\left(\frac{\partial\log f_X}{\partial\theta}\right)^2\right]_{\theta_0} \neq 0 \tag{8.9}$$

for all non-degenerate cases. Then we see from (8.8) that $\widehat{\theta}$ at large n approaches the true value θ_0, since $\lim_{n\to\infty}\left(\widehat{\theta}-\theta_0\right) = -\lim_{n\to\infty}\alpha/\beta = -E[\alpha]/E[\beta] = 0$. Therefore, the estimate for the parameter θ_0 is

$$\widehat{\theta} \approx \theta_0 - \frac{\alpha}{E[\beta]}.$$

What is the variance of this estimate? We compute it as the expected value

$$\mathrm{var}\big[\widehat{\theta}\big] = E\left[\big(\widehat{\theta}-\theta_0\big)^2\right] = E\left[\left(\frac{\alpha}{E[\beta]}\right)^2\right] = \frac{E\big[\alpha^2\big]}{E[\beta]^2}.$$

The denominator has already been calculated in (8.9), and the numerator is

$$E\big[\alpha^2\big] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left[\frac{\partial \log f_X(X_i;\theta)}{\partial\theta}\frac{\partial \log f_X(X_j;\theta)}{\partial\theta}\right]_{\theta_0}.$$

Individual random variables X_i are mutually independent, so all mixed terms may be discarded, only the n quadratic terms survive:

$$E\big[\alpha^2\big] = \frac{1}{n^2}\sum_{i=1}^{n} E\left[\left(\frac{\partial \log f_X(X_i;\theta)}{\partial\theta}\right)^2\right]_{\theta_0} = -\frac{E[\beta]}{n}.$$

So the variance we are looking for is

$$\mathrm{var}\big[\widehat{\theta}\big] = -\frac{1}{nE[\beta]} = \frac{1}{n}\left\{E\left[\left(\frac{\partial \log f_X}{\partial\theta}\right)^2\right]_{\theta_0}\right\}^{-1} = -\frac{1}{n}\left\{E\left[\frac{\partial^2 \log f_X}{\partial\theta^2}\right]_{\theta_0}\right\}^{-1}.$$

Denoting $f_{X,i} = f_X(X_i;\theta)$ and taking into account the mutual independence of individual X_i, the following relations also hold true:

$$E\big[(\ell')^2\big] = E\Big[\Big(\sum_i \log f_{X,i}\Big)'^2\Big] = E\left[\left(\sum_i \frac{f'_{X,i}}{f_{X,i}}\right)'^2\right] = E\left[\sum_i\left(\frac{f'_{X,i}}{f_{X,i}}\right)^2\right] = n\,E\left[\left(\frac{f'_X}{f_X}\right)^2\right],$$

$$E\big[\ell''\big] = E\Big[\Big(\sum_i \log f_{X,i}\Big)''\Big] = n\,E\Big[\big(\log f_X\big)''\Big] = n\,E\left[\left(\frac{f'_X}{f_X}\right)'\right] = -n\,E\Big[\big(\log f_X\big)'^2\Big].$$

The variance can thus be computed in at least four equivalent ways; two require the log-likelihood function, and two require the probability density:

$$\Big(\mathrm{var}\big[\widehat{\theta}\big]\Big)^{-1} = E\big[(\ell')^2\big] = -E\big[\ell''\big] = n\,E\left[\left(\frac{f'_X}{f_X}\right)^2\right] = -n\,E\Big[\big(\log f_X\big)''\Big]. \quad (8.10)$$

Example Let us determine the parameter a in the Pareto distribution (3.16) based on the measured sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ and the assumption that the other parameter, b, is known. The likelihood function is

$$L(\boldsymbol{x}|a) = \prod_{i=1}^{n} f_X(x_i;a) = \prod_{i=1}^{n}\frac{a}{b}\left(\frac{b}{x_i}\right)^{a+1},$$

and its logarithm is

$$\ell = \sum_{i=1}^{n} \left[\log \frac{a}{b} + (a+1) \log \frac{b}{x_i} \right] = n \log \frac{a}{b} + (a+1) \sum_{i=1}^{n} \log \frac{b}{x_i}.$$

The likelihood equation $\partial\ell/\partial a = 0$ yields $n/a + \sum_{i=1}^{n} \log(b/x_i) = 0$, whence the estimate

$$\widehat{a} = n \left[\sum_{i=1}^{n} \log \frac{x_i}{b} \right]^{-1} . \tag{8.11}$$

Formula (8.10) then gives its variance:

$$\begin{aligned} \left(\mathrm{var}\big[\widehat{a}\big]\right)^{-1} &= -n\,E\Big[\Big(\log f_X\Big)''\Big] = -n\,E\bigg[\bigg(\log \frac{a}{b} + (a+1) \log \frac{b}{X}\bigg)''\bigg] \\ &= -n\,E\bigg[\bigg(\frac{1}{a} + \log \frac{b}{X}\bigg)'\bigg] = -n\,E\bigg[-\frac{1}{a^2}\bigg] = \frac{n}{a^2}, \end{aligned}$$

hence $\mathrm{var}\big[\widehat{a}\big] = a^2/n$. ◁

8.4 Efficiency of Estimator

There is a relation between the bias and the variance of an estimator, based on the information contained in the sample. (The concept of information will be discussed in detail in Chap. 11.) The equivalent quantities in (8.10) can be interpreted as the *information* of the sample *with respect to parameter* θ,

$$I(\theta) = E\big[(\ell')^2\big] = -E\big[\ell''\big] \propto n.$$

One can prove that the variance of estimates, obtained with a specific estimator, is bounded from below [1, 2]. The lower bound of the variance of estimator T with bias b is given by the Cramér-Rao or *information inequality*

$$\mathrm{var}[T] \geq \frac{[1 + b'(\theta)]^2}{I(\theta)}.$$

If b does not depend on θ—or if the estimator is unbiased ($b = 0$)—the *minimum variance bound* is given by the inequality

$$\mathrm{var}[T] \geq \frac{1}{I(\theta)} \propto \frac{1}{n} \tag{8.12}$$

with a very clear interpretation: by increasing the information (sample size) it is possible to reduce the variance of the estimate obtained from the sample with estimator T. It can be shown [3] that the lower bound is attained precisely when

$$\ell' = A(\theta)\big(T - E[T]\big) = A(\theta)\big(T - \theta - b(\theta)\big). \tag{8.13}$$

Here A is an arbitrary quantity independent of $\boldsymbol{x}$, but it may depend on the parameter θ. Integrating the above relation we get $\ell = \int \ell' \,\mathrm{d}\theta = B(\theta)T + C(\theta) + D$, whence, by inverting the log,

$$L = \mathrm{e}^{\ell} = d \exp\big[B(\theta)T + C(\theta)\big], \tag{8.14}$$

where D and d do not depend on θ. Hence T is a *minimum variance estimator* if the likelihood function L has the particular form (8.14). If, in addition, such an estimator is unbiased, it also follows from (8.12) that

$$\mathrm{var}[T] = \frac{1}{I(\theta)} = \frac{1}{E[(\ell')^2]} = \frac{1}{\big(A(\theta)\big)^2 E\big[(T - E[T])^2\big]} = \frac{1}{\big(A(\theta)\big)^2 \mathrm{var}[T]}$$

or

$$\mathrm{var}[T] = \frac{1}{|A(\theta)|}. \tag{8.15}$$

The quality of an estimator is expressed by its *efficiency*, defined as the ratio of the minimal and actual variance of the estimator,

$$\mathrm{eff}[T] = \big(\mathrm{var}[T]\big)_{\min} \big/ \mathrm{var}[T] \ , \qquad 0 \le \mathrm{eff}[T] \le 1.$$

In principle high efficiency is desirable, although it does not say much about other qualities of an estimator, for example, its robustness.

Example We have acquired an integer sample $\{x_1, x_2, \ldots, x_n\}$, assumed to stem from a Poisson-distributed population, corresponding to the probability function $f_X(x; \lambda) = \lambda^x \mathrm{e}^{-\lambda}/x!$ and unknown parameter λ. We wish to determine this parameter. The log-likelihood function is

$$\ell = \log\left(\prod_{i=1}^{n} f_X(x_i; \lambda)\right) = \sum_{i=1}^{n} \big[x_i \log\lambda - \log(x_i!) - \lambda\big].$$

By comparing its derivative

$$\ell' = \frac{\mathrm{d}\ell}{\mathrm{d}\lambda} = \sum_{i=1}^{n} \left(\frac{x_i}{\lambda} - 1\right) = \frac{1}{\lambda}\sum_{i=1}^{n} (x_i - \lambda) = \frac{n}{\lambda}\big(\overline{x} - \lambda\big)$$

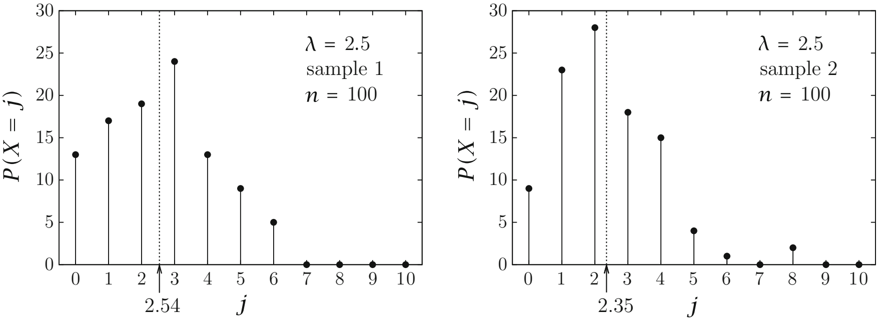


Fig. 8.2 Samples of size $n = 100$ taken from a Poisson-distributed population with parameter $\lambda = 2.5$. [Left] Sample with arithmetic mean $\overline{x} = \widehat{\lambda} = 2.54$. [Right] Sample with $\overline{x} = \widehat{\lambda} = 2.35$. The expected variance of the sample mean is $\lambda/n = 0.025$ and the effective deviation is $\sqrt{0.025} \approx 0.158$, consistent with the values shown

and formula (8.13) we see that the arithmetic mean $T = \overline{X}$ is an unbiased estimator for the parameter $\theta = \lambda$ with variance $|A(\lambda)|^{-1} = \lambda/n$. How this works in practice is shown in Fig. 8.2. ◁

Example Let us examine the variances of the sample mean $\overline{x} = \frac{1}{n}\sum_i x_i$ and the sample median (definition (7.22)) as two possible estimators for the true population mean μ, assuming that the population is normally distributed, with known variance σ^2. The derivative of the log-likelihood function with n observations is

$$\ell' = \frac{\partial \ell}{\partial \mu} = \frac{\partial}{\partial \mu} \log \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right] = \frac{n}{\sigma^2}\big(\overline{x} - \mu\big).$$

This has the form (8.13), where $T = \overline{X}, \theta = \mu, b(\theta) = 0$ and $A(\theta) = n/\sigma^2$. By (8.15) therefore

$$\mathrm{var}\big[\overline{X}\big] = \frac{1}{|A(\theta)|} = \frac{\sigma^2}{n}.$$

We know this already, for example, from (7.7). What about the median? The sample median of large samples ($n \gg 1$) is normally distributed according to (7.24), so

$$\mathrm{var}\big[\mathrm{med}[\boldsymbol{X}]\big] = \frac{\pi\sigma^2}{2n}.$$

Therefore the sample median is a less efficient estimator for the population mean than the sample mean, since its efficiency is only

$$\mathrm{eff}\big[\mathrm{med}[\boldsymbol{X}]\big] = \frac{\mathrm{var}[\overline{X}]}{\mathrm{var}\big[\mathrm{med}[\boldsymbol{X}]\big]} = \frac{\sigma^2}{n}\,\frac{2n}{\pi\sigma^2} = \frac{2}{\pi} \approx 0.637.$$

In plain English: after many samplings, an equally good estimate for the true mean μ of a normally distributed population can be obtained from a sample mean of 637 observations as from the median of 1000 observations. ◁

8.5 Likelihood Intervals

Likelihood intervals are analogs of confidence intervals discussed in Sect. 7.3 with a slightly different interpretation. A confidence interval $[\overline{X} - s_X/\sqrt{n}, \overline{X} + s_X/\sqrt{n}]$ expresses the probability that the true mean μ will be on this interval, namely $P\big(\overline{X} - s_X/\sqrt{n} \le \mu \le \overline{X} + s_X/\sqrt{n}\big) \approx 0.683$. On the other hand, a likelihood interval

$$\big[\widehat{\theta} - \sigma \le \theta \le \widehat{\theta} + \sigma\big], \tag{8.16}$$

where it is assumed that the true σ is known, and the corresponding probability

$$p = P\big(\widehat{\theta} - \sigma \le \theta \le \widehat{\theta} + \sigma\big) \approx 0.683, \tag{8.17}$$

measure our "belief" that the observations $\boldsymbol{x}$ were generated by a random process with parameter θ from the interval (8.16).

How can a likelihood interval be determined? The distribution of $\widehat{\theta}$ is generally unknown, so we also do not know how to compute the probability (8.17). However, if we resort to the large-sample limit ($n \gg 1$), all maximum-likelihood estimates attain the minimum variance bound [4]. In addition, in the asymptotic regime $n \to \infty$ the likelihood function becomes independent of the sample $\boldsymbol{x}$ and tends to the normal distribution in θ, with mean $\widehat{\theta}$ and variance σ^2:

$$L(\boldsymbol{x}|\theta) \to L(\theta) = L(\widehat{\theta})\, \mathrm{e}^{-\frac{1}{2}R} = L_{\max}\, \mathrm{e}^{-\frac{1}{2}R}, \qquad R = (\theta - \widehat{\theta})^2/\sigma^2. \tag{8.18}$$

Here the *true* θ is to be seen as "dancing" about the parameter $\widehat{\theta}$. In its vicinity, the log-likelihood function has a parabolic shape

$$\ell(\theta) - \ell(\widehat{\theta}) = \ell(\theta) - \ell_{\max} = -\frac{1}{2}\frac{(\theta - \widehat{\theta})^2}{\sigma^2}, \tag{8.19}$$

shown in Fig. 8.3.

An arbitrary likelihood interval $[\theta^-, \theta^+]$ is then defined by the formula

$$p = P\big(\theta^- \le \theta \le \theta^+\big) = \Phi\left(\frac{\theta^+ - \widehat{\theta}}{\sigma}\right) - \Phi\left(\frac{\theta^- - \widehat{\theta}}{\sigma}\right), \tag{8.20}$$

where Φ is the distribution function of the standard normal distribution (3.11). We are mostly interested in symmetric intervals $[\theta^-, \theta^+] = [\widehat{\theta} - m\sigma, \widehat{\theta} + m\sigma]$ with probability content p and probabilities $\frac{1}{2}(1 - p)$ to the left and right of them:

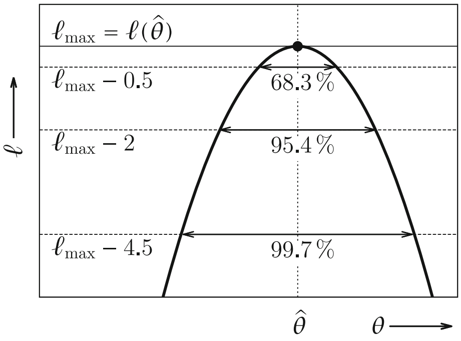


Fig. 8.3 The parabolic shape of the log-likelihood function near $\widehat{\theta}$ with the corresponding likelihood intervals for $p = 68.3\%$, $p = 95.4\%$ and 99.7%

$$p = P\big(\widehat{\theta} - m\sigma \le \theta \le \widehat{\theta} + m\sigma\big) = 2\Phi(m) - 1. \tag{8.21}$$

The intervals correspond to line segments between the intersections of parabolas with horizontal lines at a below $\ell_{\max}$, where $a = R/2 = m^2/2$, as shown in Fig. 8.3. Likelihood intervals with $p = 68.3$, 95.4 and 99.7% ($m = 1$, 2 and 3) correspond to $a = 0.5$, 2.0 and 4.5, respectively. The method approximately works even in the asymmetric case, as shown in the following Example.

Example Based on measured decay instances $\boldsymbol{t} = \{t_1, t_2, \ldots, t_n\}$ we wish to determine the decay time of radioactive nuclei. The sample $\boldsymbol{t}$ may not be large: Fig. 8.4 (left) shows $n = 5$ decays, and Fig. 8.4 (right) shows $n = 50$ decays.

The probability for the nucleus to decay in the time interval $[t, t + \mathrm{d}t]$ is given by the exponential distribution, so that $f_T(t)\,\mathrm{d}t = \tau^{-1}\,\mathrm{e}^{-t/\tau}\,\mathrm{d}t$. The likelihood function for the sample $\boldsymbol{t}$ with parameter τ is

$$L(\boldsymbol{t}|\tau) = \prod_{i=1}^{n} f_T(t_i; \tau) = \frac{1}{\tau^n} \exp\left(-\frac{1}{\tau}\sum_{i=1}^{n} t_i\right) = \frac{1}{\tau^n} \exp\left(-\frac{n}{\tau}\bar{t}\right),$$

while its log is $\ell = -n\bar{t}/\tau - n\log\tau$. Comparing the likelihood equation for τ,

$$\ell' = \frac{\mathrm{d}\ell}{\mathrm{d}\tau} = \frac{n}{\tau}\left(\frac{\bar{t}}{\tau} - 1\right) = \frac{n}{\tau^2}\left(\bar{t} - \tau\right) = 0,$$

to (8.13) we deduce that $\widehat{\tau} = \bar{t} = \left(\sum_i t_i\right)/n$ is an unbiased estimator for the mean decay time, with variance $\mathrm{var}[\widehat{\tau}] = \tau^2/n$, so that the uncertainty of the parameter $\widehat{\tau}$ is $\tau/\sqrt{n} \approx \bar{t}/\sqrt{n}$. At $\widehat{\tau} = \bar{t}$ one has $\ell(\widehat{\tau}) = \ell_{\max} = -n\big(1 + \log\widehat{\tau}\big)$ or

$$-\big(\ell(\tau) - \ell(\widehat{\tau})\big) = n\left(\frac{\widehat{\tau}}{\tau} + \log\frac{\tau}{\widehat{\tau}} - 1\right).$$

For small n this does not have the parabolic shape (8.19) in τ, so one can not determine a symmetric likelihood interval by using (8.21). But one can still define an asymmetric

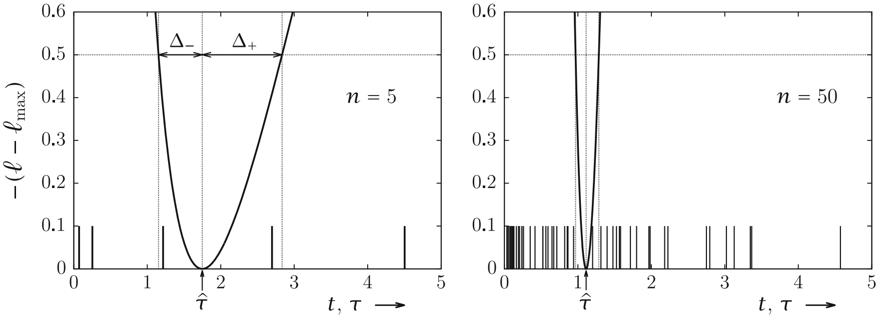


Fig. 8.4 Determination of the mean decay time of nuclei from the sample of [Left] $n = 5$ and [Right] $n = 50$ measured decay instances. Also shown are the graphs of log-likelihood functions. The true decay time, used to generate the events, is $\tau = 1$

interval $[\tau_-, \tau_+]$, where $\tau_- = \widehat{\tau} - \Delta_-$ and $\tau_+ = \widehat{\tau} + \Delta_+$. For, say, $p = 68.3\%$, this interval is defined by

$$-\big(\ell(\tau_\pm) - \ell(\widehat{\tau})\big) = \tfrac{1}{2}$$

and is shown in the upper part of Fig. 8.4 (left). When $n \gg 1$, $\ell - \ell_{\max}$ becomes more and more parabolic and the corresponding likelihood interval more and more symmetric (see Fig. 8.4 (right)). Ultimately, in the limit, $n \to \infty$ we finally obtain a symmetric interval $\Delta_\pm = \sqrt{\mathrm{var}[\widehat{\tau}]}$. ◁

8.6 Simultaneous Determination of Multiple Parameters

Let us revisit the issue of determining multiple parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_p\}$, whose values we wish to infer from the sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$. The likelihood equations with corresponding log-likelihood functions used to obtain the estimates for individual θ_j have already been written down: see (8.4). What complicates the matter is the determination of uncertainties (variances) of these parameters and their correlations.

8.6.1 General Method for Arbitrary (Small or Large) Samples

If estimates can be written as explicit functions of the variables $\boldsymbol{X}$, that is, in the form $\widehat{\theta}_j = \widehat{\theta}_j(X_1, X_2, \ldots, X_n)$, the covariances of $\widehat{\theta}_i$ and $\widehat{\theta}_j$ can be defined as

$$\mathrm{cov}\big[\widehat{\theta}_i, \widehat{\theta}_j\big] = \int \big(\widehat{\theta}_i - \theta_i\big)\big(\widehat{\theta}_j - \theta_j\big) L\big(\boldsymbol{x}|\boldsymbol{\theta}\big)\,\mathrm{d}\boldsymbol{x}. \tag{8.22}$$

The variances of individual θ_i are obtained when this formula is applied at $i = j$,

$$\mathrm{var}\big[\widehat{\theta_i}\big] = \int \big(\widehat{\theta_i} - \theta_i\big)^2 L\big(\boldsymbol{x}|\boldsymbol{\theta}\big)\,\mathrm{d}\boldsymbol{x}.$$

The multiple integral should be calculated on the whole definition domain of the random variables X_i, corresponding to sample values x_i. This method is applicable to samples of any size, small or large.

Example By using (8.22) let us show that the variance of the estimator $\widehat{\tau} = \overline{t}$ for the mean decay time is indeed τ^2/n, as shown in Example on p. 213:

$$\begin{aligned}\mathrm{var}\big[\widehat{\tau}\big] &= \underbrace{\int_0^\infty \int_0^\infty \cdots \int_0^\infty}_{n} \big(\widehat{\tau} - \tau\big)^2 \prod_{i=1}^n \left(\frac{1}{\tau}\,\mathrm{e}^{-t_i/\tau}\,\mathrm{d}t_i\right) \\ &= \int \cdots \int \left(\frac{1}{n}\sum_{k=1}^n t_k\right)\left(\frac{1}{n}\sum_{j=1}^n t_j\right)\prod_{i=1}^n \left(\frac{1}{\tau}\,\mathrm{e}^{-t_i/\tau}\,\mathrm{d}t_i\right) \\ &\quad -2\tau \int \cdots \int \left(\frac{1}{n}\sum_{j=1}^n t_j\right)\prod_{i=1}^n \left(\frac{1}{\tau}\,\mathrm{e}^{-t_i/\tau}\,\mathrm{d}t_i\right) + \tau^2 \int \cdots \int \prod_{i=1}^n \left(\frac{1}{\tau}\,\mathrm{e}^{-t_i/\tau}\,\mathrm{d}t_i\right) \\ &= \left(\frac{2}{n} + \frac{n-1}{n}\right)\tau^2 - 2\tau^2 + \tau^2 = \frac{\tau^2}{n}. \end{aligned} \tag{8.23}$$

Of course, one can also be very brief: $\mathrm{var}\big[\widehat{\tau}\big] = \mathrm{var}\big[\overline{T}\big] = n\,\mathrm{var}[T]/n^2 = \tau^2/n$. We shall revisit the decay time determination in Problem 8.8.1. ◁

8.6.2 Asymptotic Method (Large Samples)

For large samples ($n \gg 1$) the dependence (8.19) can be generalized to multiple parameters as

$$\ell(\boldsymbol{\theta}) - \ell(\widehat{\boldsymbol{\theta}}) \approx -\tfrac{1}{2}\big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\big)^{\mathrm{T}} A \big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\big)^{\mathrm{T}},$$

where A is the matrix of negative second derivatives of ℓ with respect to θ_j as in (8.5). Its expected value $B = E[A]$ is a symmetric matrix with the elements

$$B_{ij} = E\big[A_{ij}\big] = -E\left[\left(\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\right)\right]_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \tag{8.24}$$

and the likelihood function has the form of a p-dimensional normal density (4.23),

$$L(\boldsymbol{x}|\boldsymbol{\theta}) \propto \exp\left(-\tfrac{1}{2}\big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\big)^{\mathrm{T}} C^{-1} \big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\big)\right),$$

where $C = B^{-1}$ is the covariance matrix of the parameters $\boldsymbol{\theta}$. Its elements are

$$C_{ii} = \mathrm{var}\big[\widehat{\theta}_i\big], \qquad C_{ij} = \mathrm{cov}\big[\widehat{\theta}_i, \widehat{\theta}_j\big], \tag{8.25}$$

while the correlation coefficient of an arbitrary parameter pair is

$$\rho\big(\widehat{\theta}_i, \widehat{\theta}_j\big) = \frac{\mathrm{cov}\big[\widehat{\theta}_i, \widehat{\theta}_j\big]}{\sqrt{\mathrm{var}\big[\widehat{\theta}_i\big]}\sqrt{\mathrm{var}\big[\widehat{\theta}_j\big]}} = \frac{C_{ij}}{\sqrt{C_{ii}}\sqrt{C_{jj}}}.$$

Example A sample $\{x_1, x_2, \ldots, x_n\}$ presumably stems from a normally distributed population. We are interested in the estimates for its mean μ and effective deviation σ as well as their uncertainties. We already know the log-likelihood function (8.6), except that now all effective deviations are the same. Since two parameters are involved, $\theta_1 = \mu$ and $\theta_2 = \sigma$, there are two likelihood equations:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0, \qquad \frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0.$$

The usual formulas for sample mean and variance follow:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\big(x_i - \widehat{\mu}\big)^2.$$

To calculate their uncertainties, we need the second derivatives:

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\frac{n}{\sigma^2}, \qquad \frac{\partial^2 \ell}{\partial \mu \partial \sigma} = -\frac{2\sum_{i=1}^{n}\big(x_i - \mu\big)}{\sigma^3}, \qquad \frac{\partial^2 \ell}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{3\sum_i\big(x_i - \mu\big)^2}{\sigma^4},$$

whose expected values are

$$B_{11} = E\big[A_{11}\big] = -E\left[\left(\frac{\partial^2 \ell}{\partial \mu^2}\right)\right]_{\widehat{\mu},\widehat{\sigma}} = \frac{n}{\widehat{\sigma}^2},$$

$$B_{12} = E\big[A_{12}\big] = -E\left[\left(\frac{\partial^2 \ell}{\partial \mu \partial \sigma}\right)\right]_{\widehat{\mu},\widehat{\sigma}} = \frac{2\sum_{i=1}^{n}\big(\overbrace{E[X_i - \mu]}^{0}\big)_{\widehat{\mu}}}{\widehat{\sigma}^3} = 0,$$

$$B_{22} = E\big[A_{22}\big] = -E\left[\left(\frac{\partial^2 \ell}{\partial \sigma^2}\right)\right]_{\widehat{\mu},\widehat{\sigma}} = -\frac{n}{\widehat{\sigma}^2} + \frac{3}{\widehat{\sigma}^4}\sum_{i=1}^{n}\Big(\underbrace{E\big[(X_i - \mu)^2\big]}_{\sigma^2}\Big)_{\widehat{\sigma}} = \frac{2n}{\widehat{\sigma}^2}.$$

Inverting the matrix B yields the covariance matrix:

$$B = \begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix} \quad \Longrightarrow \quad C = B^{-1} = \begin{pmatrix} \widehat{\sigma}^2/n & 0 \\ 0 & \widehat{\sigma}^2/2n \end{pmatrix},$$

so

$$\mathrm{var}\big[\widehat{\mu}\big] = \frac{\widehat{\sigma}^2}{n}, \quad \mathrm{var}\big[\widehat{\sigma}\big] = \frac{\widehat{\sigma}^2}{2n}, \quad \mathrm{cov}\big[\widehat{\mu}, \widehat{\sigma}\big] = 0.$$

The estimates $\widehat{\mu}$ and $\widehat{\sigma}$—in this particular case—are uncorrelated. Repeat the calculation for $\theta_2 = \sigma^2$ instead of $\theta_2 = \sigma$! What is the difference? ◁

8.7 Likelihood Regions

When the maximum-likelihood method is used to determine multiple parameters $\boldsymbol{\theta}$ and one would like to specify their uncertainties, likelihood intervals are replaced by *likelihood regions*. Similar to (8.20) one is usually interested in the probabilities that parameters θ_1 and θ_2 simultaneously lie on their corresponding intervals,

$$p = P\big(\theta_1^- \le \theta_1 \le \theta_1^+,\ \theta_2^- \le \theta_2 \le \theta_2^+\big).$$

Let us restrict the discussion to two, generally correlated parameters by using large samples ($n \gg 1$). By analogy to the one-dimensional case (8.18) the likelihood function near the optimal values $\widehat{\theta_1}$ and $\widehat{\theta_2}$ has the form of a binormal density in θ_1 and θ_2 (see Example on p. 108), with possible correlations:

$$L(\boldsymbol{x}|\theta_1, \theta_2) \to L(\theta_1, \theta_2) = L(\widehat{\theta_1}, \widehat{\theta_2})\, \mathrm{e}^{-\frac{1}{2}R} = L_{\max}\, \mathrm{e}^{-\frac{1}{2}R}$$

or

$$\ell\big(\theta_1, \theta_2\big) = \ell_{\max} - \tfrac{1}{2}R,$$

where R is a random variable

$$R = \frac{1}{1-\rho^2}\left[\frac{\big(\theta_1 - \widehat{\theta_1}\big)^2}{\sigma_1^2} - 2\rho\frac{\big(\theta_1 - \widehat{\theta_1}\big)\big(\theta_2 - \widehat{\theta_2}\big)}{\sigma_1\sigma_2} + \frac{\big(\theta_2 - \widehat{\theta_2}\big)^2}{\sigma_2^2}\right]. \tag{8.26}$$

The curves of constant likelihood are ellipses centered at $(\widehat{\theta_1}, \widehat{\theta_2})$, defining the corresponding likelihood region. The limiting value $R = 1$ defines the *covariance ellipse*, an example of which is shown in Fig. 8.5 (left).

It turns out [4] that R—regardless of $\widehat{\theta_1}, \widehat{\theta_2}, \sigma_1, \sigma_2$ and ρ—is distributed according to the χ^2 distribution with $\nu = 2$, so that with a chosen probability p one has

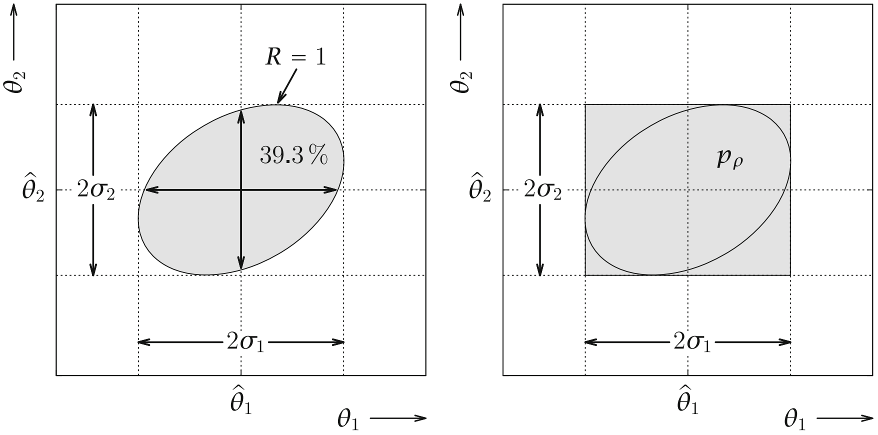


Fig. 8.5 [Left] Covariance ellipse as the boundary of the likelihood region ($p = 39.3\%$) for parameters θ_1 and θ_2. The arrows with lengths $2\sigma_1$ and $2\sigma_2$ denote the usual likelihood intervals ($p = 68.3\%$) for an individual parameter *regardless of the other*. [Right] A rectangle circumscribing the ellipse as an alternative likelihood region

$$P\big(R \le R_p\big) = \int_0^{R_p} f_{\chi^2}(x;2)\,\mathrm{d}x = \frac{1}{2}\int_0^{R_p} \mathrm{e}^{-x/2}\,\mathrm{d}x = 1 - \mathrm{e}^{-R_p/2} = p, \qquad (8.27)$$

where f_{χ^2} is given by (3.21). The probability that $R \le R_p$ is equal to the probability that θ_1 and θ_2 are simultaneously within the ellipse defined by the equation $R(\theta_1,\theta_2) = R_p$. One first chooses a probability p with which one would like to jointly "capture" θ_1 and θ_2 in the elliptic region: the corresponding ellipses are then the intersections of the surface $\big(\theta_1,\theta_2,\ell(\theta_1,\theta_2)\big)$ with parallel planes $\ell = \ell_{\max} - a$, where $a = R_p/2$—just like in the one-dimensional case in Fig. 8.3. Solving (8.27) for R_p,

$$R_p = -2\log(1-p),$$

then yields the equation of the ellipse (8.26) with $R(\theta_1,\theta_2) = R_p$. Some typical pairs of p and R_p are listed in the table below.

p	0.393	0.500	0.683	0.865	0.954	0.989	0.997
R_p	1	1.39	2.30	4	6.16	9	11.62

8.7.1 Alternative Likelihood Regions

Likelihood regions may be defined by any prescription that algebraically or geometrically maps the parameter uncertainties to the chosen probability p in a unique way.

Instead of an elliptic region, for example, one can define a rectangular one, such that its sides correspond to the parameter ranges

$$\theta_1 = \widehat{\theta_1} \pm m\sigma_1, \qquad \theta_2 = \widehat{\theta_2} \pm m\sigma_2, \qquad m \in \mathbb{R}.$$

The probability p that θ_1 and θ_2 are simultaneously within the rectangle depends on the correlation parameter ρ. It is given by the formula [4]

$$p_\rho(m) = \frac{1}{\sqrt{2\pi}} \int_{-m}^{m} \left[\Phi\left(\frac{m - \rho t}{\sqrt{1-\rho^2}} \right) - \Phi\left(\frac{-m - \rho t}{\sqrt{1-\rho^2}} \right) \right] \mathrm{e}^{-t^2/2}\, \mathrm{d}t,$$

where Φ is the distribution function of the standard normal distribution (3.11). The integral is calculated numerically. If a rectangle circumscribes the covariance ellipse as shown in Fig. 8.5 (right), one obtains a likelihood region with $m = 1$, on which θ_1 and θ_2 are found with probability

$$P\left(\widehat{\theta_1} - \sigma_1 \le \theta_1 \le \widehat{\theta_1} + \sigma_1, \widehat{\theta_2} - \sigma_2 \le \theta_2 \le \widehat{\theta_2} + \sigma_2\right) = p_\rho(1).$$

Some typical pairs of ρ and $p_\rho(1)$ are shown in the table below.

ρ	0.0	0.2	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$p_\rho(1)$	0.466	0.471	0.486	0.498	0.514	0.534	0.561	0.596	0.683

The column with $\rho = 0.0$ corresponds to uncorrelated parameters, for which

$$\begin{aligned} p_\rho(1) &= P\left(\widehat{\theta_1} - \sigma_1 \le \theta_1 \le \widehat{\theta_1} + \sigma_1, \widehat{\theta_2} - \sigma_2 \le \theta_2 \le \widehat{\theta_2} + \sigma_2\right) \\ &= P\left(\widehat{\theta_1} - \sigma_1 \le \theta_1 \le \widehat{\theta_1} + \sigma_1\right) P\left(\widehat{\theta_2} - \sigma_2 \le \theta_2 \le \widehat{\theta_2} + \sigma_2\right) \\ &\approx 0.683^2 \approx 0.466. \end{aligned}$$

8.8 Problems

8.8.1 *Lifetime of Particles in Finite Detector*

A detector is used to measure the lifetime of unstable particles, yielding the sample $\boldsymbol{t} = \{t_1, t_2, \ldots, t_n\}$. The times $t_i = l_i/(\gamma_i v_i)$, where $\gamma_i = (1 - v_i^2/c^2)^{-1/2}$ is the Lorentz factor, are calculated for each particle from the measured length l_i of its trajectory and its velocity v_i. Use the sample $\boldsymbol{t}$ to obtain the maximum-likelihood estimate of the mean decay time and its variance! Discuss the cases that the detector is ① infinitely large or ② finite.

✎ ① Decay times are exponentially distributed with density $f_T(t;\tau) = \tau^{-1}\mathrm{e}^{-t/\tau}$, where τ is the true decay time. In an infinite detector one may have $0 \le t < \infty$ and the density f_T is correctly normalized. The likelihood function for n observations is

$$L(\boldsymbol{t}|\tau) = \prod_{i=1}^{n} f_T(t_i;\tau) = \frac{1}{\tau^n}\exp\left(-\frac{1}{\tau}\sum_{i=1}^{n} t_i\right) = \frac{1}{\tau^n}\exp\left(-\frac{n}{\tau}\bar{t}\right),$$

where $\bar{t} = (1/n)\sum_{i=1}^{n} t_i$ is the usual sample mean. The log-likelihood function is $\ell = -n\log\tau - n\bar{t}/\tau$, and the corresponding likelihood equation is

$$\frac{\mathrm{d}\ell}{\mathrm{d}\tau} = \frac{n}{\tau}\left(\frac{\bar{t}}{\tau} - 1\right) = \frac{n}{\tau^2}\left(\bar{t} - \tau\right) = 0.$$

The estimate $\widehat{\tau}$ for the true decay time τ is therefore the sample mean, $\widehat{\tau} = \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$. Its variance is given by formula (8.23).

② The finite-detector case is more interesting. Here the decays can be described by the probability density

$$f_T(t;\tau) = \frac{\tau^{-1}\mathrm{e}^{-t/\tau}}{\int_0^T \tau^{-1}\mathrm{e}^{-t/\tau}\,\mathrm{d}t} = \frac{1}{\tau}\frac{\mathrm{e}^{-t/\tau}}{1-\mathrm{e}^{-T/\tau}}, \qquad 0 \le t \le T,$$

where T is the *potential decay time*. Namely, for the ith particle the time can only be measured on a finite interval $0 \le t_i \le T_i = l_{\mathrm{max},i}/(\gamma_i v_i)$. The log-likelihood function is now

$$\ell = -n\log\tau - \frac{n\bar{t}}{\tau} - \sum_{i=1}^{n}\log\left[1 - \mathrm{e}^{-T_i/\tau}\right],$$

while the likelihood equation is

$$\frac{\mathrm{d}\ell}{\mathrm{d}\tau} = \frac{n}{\tau}\left(\frac{\bar{t}}{\tau} - 1\right) + \frac{1}{\tau^2}\sum_{i=1}^{n}\frac{T_i\,\mathrm{e}^{-T_i/\tau}}{1-\mathrm{e}^{-T_i/\tau}} = 0.$$

Multiplying by τ^2 yields an implicit equation for τ,

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\left[t_i + \frac{T_i\,\mathrm{e}^{-T_i/\tau}}{1-\mathrm{e}^{-T_i/\tau}}\right] = \bar{t} + \frac{1}{n}\sum_{i=1}^{n}\frac{T_i\,\mathrm{e}^{-T_i/\tau}}{1-\mathrm{e}^{-T_i/\tau}},$$

which can be solved iteratively with the initial condition obtained in Problem ① . With $\widehat{\tau}$ we also calculate its variance by using the formula

$$\mathrm{var}\left[\widehat{\tau}\right] = \left[\left(\frac{\mathrm{d}^2\ell}{\mathrm{d}\tau^2}\right)\right]^{-1}_{\tau=\widehat{\tau}}.$$

8.8.2 Device Failure Due to Corrosion

Figure 8.6 (left) shows $n = 32$ device lifetimes t_i (times until failure) in dependence of the corrosion level of its components [5]. Assume that the lifetime is an exponentially distributed random variable T with the probability density $f_T(t; a, b) = \lambda\,\mathrm{e}^{-\lambda t} = ax^b\,\mathrm{e}^{-ax^b t}$, where x is the corrosion level, while a and b are unknown parameters. Determine a and b and their variances by using the maximum-likelihood method!

✎ The sample consists of $n = 32$ pairs (x_i, t_i). The likelihood function and its logarithm are

$$L(\boldsymbol{t}, \boldsymbol{x}|a, b) = \prod_{i=1}^{n} ax_i^b\,\mathrm{e}^{-ax_i^b t_i}, \qquad \ell = \log L = n\log a + b\sum_{i=1}^{n}\log x_i - a\sum_{i=1}^{n} x_i^b t_i.$$

The likelihood equations are

$$\frac{\partial \ell}{\partial a} = \frac{n}{a} - \sum_{i=1}^{n} x_i^b t_i = 0,$$
$$\frac{\partial \ell}{\partial b} = \sum_{i=1}^{n}\log x_i - a\sum_{i=1}^{n} x_i^b t_i \log x_i = 0.$$

These equations can not be solved analytically for a and b, so one can either seek a numerical solution or directly maximize ℓ. Either way,

$$\widehat{a} \approx 1.10, \quad \widehat{b} \approx 0.49, \tag{8.28}$$

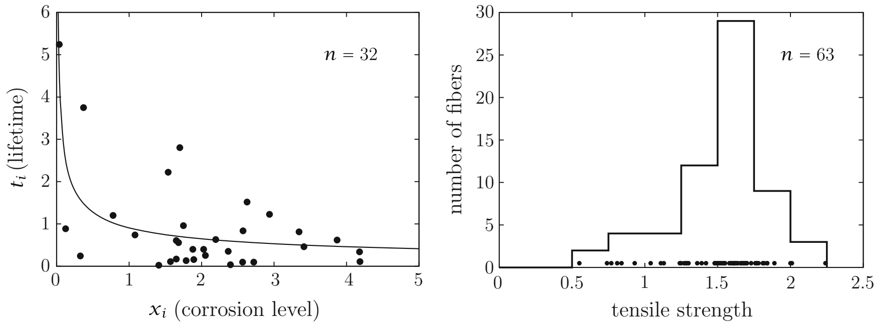


Fig. 8.6 [Left] Device lifetimes as functions of corrosion levels present in its components. The *curve* represents the model with parameters determined by the maximum-likelihood method. [Right] Measured distribution of glass fibers with respect to their tensile strength. Both data sets can be found on the book's website

so that $\ell\big(\widehat{a},\widehat{b}\big) \approx -22.3$. Since the variable T is exponentially distributed, its expected value is $E[T] = 1/\lambda = a^{-1}x^{-b}$. This curve for optimal parameters (8.28) is shown in Fig. 8.6 (left). To compute the variances we also need the second derivatives

$$\ell_{aa} = \frac{\partial^2\ell}{\partial a^2} = -\frac{n}{a^2}, \quad \ell_{ab} = \frac{\partial^2\ell}{\partial a \partial b} = -\sum_{i=1}^{n} x_i^b t_i \log x_i, \quad \ell_{bb} = \frac{\partial^2\ell}{\partial b^2} = -a\sum_{i=1}^{n} x_i^b t_i \big(\log x_i\big)^2.$$

We arrange these expressions in matrix B by formula (8.24). Its inverse is the covariance matrix of the optimal parameters:

$$B = \begin{pmatrix} -\ell_{aa} & -\ell_{ab} \\ -\ell_{ab} & -\ell_{bb} \end{pmatrix} \approx \begin{pmatrix} 26.47 & 12.33 \\ 12.33 & 36.45 \end{pmatrix}, \qquad C = B^{-1} \approx \begin{pmatrix} 0.0448 & -0.0152 \\ -0.0152 & 0.0326 \end{pmatrix}.$$

By using (8.25) we finally obtain $\sqrt{\mathrm{var}\big[\widehat{a}\big]} = \sqrt{C_{11}} \approx 0.21$ and $\sqrt{\mathrm{var}\big[\widehat{b}\big]} = \sqrt{C_{22}} \approx 0.18$.

8.8.3 Distribution of Extreme Rainfall

Let us revisit the Example from p. 159. By fitting the probability density (6.13) to the histogram in Fig. 6.7 (right) we obtained the parameter values listed in that Figure ($\widehat{\mu} = 53.9\,\mathrm{mm}$, $\widehat{\sigma} = 14.8\,\mathrm{mm}$, $\widehat{\xi} = 0.077$). Use the maximum-likelihood method to determine μ, σ and ξ as well as their uncertainties!

✎ The distribution of the measured $n = 151$ extreme values $\boldsymbol{x} = \{x_i\}_{i=1}^{n}$ is modeled by the probability density of the form (6.13). The appropriate log-likelihood function is

$$\ell(\boldsymbol{x}|\mu,\sigma,\xi) = -n\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left[1+\xi\,\frac{x_i-\mu}{\sigma}\right] - \sum_{i=1}^{n}\left[1+\xi\,\frac{x_i-\mu}{\sigma}\right]^{-1/\xi}. \quad (8.29)$$

To compute the estimates $\widehat{\mu}$, $\widehat{\sigma}$ and $\widehat{\xi}$ we need to solve the likelihood equations $\partial\ell/\partial\mu = 0$, $\partial\ell/\partial\sigma = 0$ and $\partial\ell/\partial\xi = 0$, but clearly this can be rather annoying. Such cases call for a numerical tool like MATHEMATICA to directly maximize $\ell = \texttt{logL}[\mu,\sigma,\xi]$. We already know the approximate parameter values: if, on the other hand, we have not the slightest idea of what they should be, we can simply plot ℓ: see Fig. 8.7.

Inspecting the plot allows us to narrow down the search region:

$$\texttt{NMaximize}\big[\big\{\texttt{logL}[\mu,\sigma,\xi], 50 \le \mu \le 60\,\&\&10 \le \sigma \le 20\,\&\&0.03 \le \xi \le 0.2\big\}, \big\{\mu,\sigma,\xi\big\}\big].$$

We get

$$\widehat{\mu} \approx 54.0\,\mathrm{mm}, \quad \widehat{\sigma} \approx 13.8\,\mathrm{mm}, \quad \widehat{\xi} \approx 0.11, \qquad (8.30)$$

where $\ell\big(\widehat{\mu},\widehat{\sigma},\widehat{\xi}\big) \approx -645.0$. These values are denoted by black dots in Fig. 8.7.

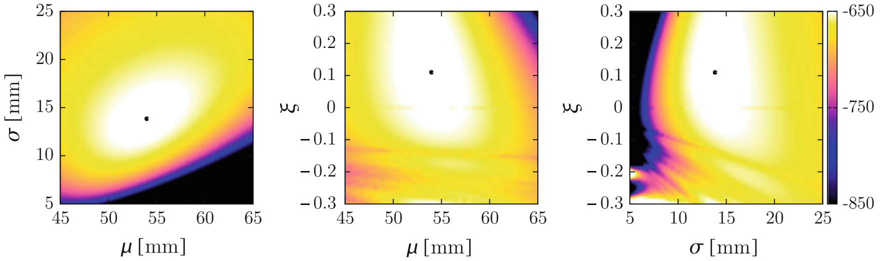


Fig. 8.7 The values of the log-likelihood function as a function of parameters μ, σ and ξ pertaining to the distribution of extreme rainfall in Engelberg. The *black symbols* denote the maximal value $\ell \approx -645.0$ attained by the parameter set (8.30)

Without the use of modern computational tools the calculation of variances of optimal parameters is even more strenuous. One needs second derivatives

$$\frac{\partial^2\ell}{\partial\mu^2} = \cdots, \quad \frac{\partial^2\ell}{\partial\mu\partial\sigma} = \cdots, \quad \frac{\partial^2\ell}{\partial\mu\partial\xi} = \cdots, \quad \frac{\partial^2\ell}{\partial\sigma^2} = \cdots, \quad \frac{\partial^2\ell}{\partial\sigma\partial\xi} = \cdots, \quad \frac{\partial^2\ell}{\partial\xi^2} = \cdots,$$

to construct the matrix B by formula (8.24) and the covariance matrix $C = B^{-1}$,

$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\mu}, \widehat{\sigma}, \widehat{\xi}\right), \qquad B_{jk} = -\left(\frac{\partial^2\ell}{\partial\theta_j\partial\theta_k}\right)_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \qquad C = B^{-1}.$$

With more and more parameters, this is no longer manageable by hand. In MATHEMATICA, on the contrary, one simply defines the parameter array $\boldsymbol{\theta}$ = `Table[q[i], {i, 3}]` and the 3×3 matrix B, which one fills with negative second derivatives:

```
logL[μ_, σ_, ξ_] := −n Log[σ]− . . . ; (* formula (8.29) *)
For[j = 1, j ≤ 3, j++, {
    For[k = 1, k ≤ 3, k++, {
        B[[j, k]] = −D[ logL[ θ[[1]], θ[[2]], θ[[3]] ], q[j], q[k] ];
  }];
}];
```

The matrix elements, calculated symbolically, need to be evaluated with the optimal parameters (8.30). The only remaining task is to compute the covariance matrix (use `Inverse[B]`):

$$C = B^{-1} \approx \begin{pmatrix} 1.6916 & 0.6702 & -0.0355 \\ 0.6702 & 0.9986 & -0.0182 \\ -0.0355 & -0.0182 & 0.0052 \end{pmatrix},$$

so the parameter uncertainties are

$$\sqrt{\mathrm{var}[\widehat{\mu}]} = \sqrt{C_{11}} \approx 1.30, \qquad \sqrt{\mathrm{var}[\widehat{\sigma}]} = \sqrt{C_{22}} \approx 1.00, \qquad \sqrt{\mathrm{var}[\widehat{\xi}]} = \sqrt{C_{33}} \approx 0.072.$$

8.8.4 Tensile Strength of Glass Fibers

In modeling the tension of glass fibers one can imagine that each fiber consists of many smaller fibers, so that the whole breaks when the weakest link in the chain breaks. Figure 8.6 (right) shows the measured distribution of fibers with respect to their tensile strength [6]. The measured strengths are therefore a kind of *minimal* extreme values; describe them with an appropriate extreme distribution of the type (6.15), and determine its parameters and their variances by using the maximum-likelihood method!

✎ As explained in Sect. 6.6.3, the same problem can be solved if one negates the data ($x_i \mapsto -x_i$) and finds the distribution of *maximal* values with the sign of the mean parameter reversed, $\widetilde{\mu} = -\widehat{\mu}$. Then the optimal parameters and their variances can be calculated precisely by the procedure outlined in Problem 8.8.3. The log-likelihood function is given by formula (8.29), and maximizing it gives the estimates

$$\widehat{\mu} = -\widetilde{\mu} \approx -1.64, \quad \widehat{\sigma} \approx 0.27, \quad \widehat{\xi} \approx -0.084,$$

at which the value of the log-likelihood function is $\ell\big(\widehat{\mu}, \widehat{\sigma}, \widehat{\xi}\big) \approx -14.3$. The parameter covariance matrix is

$$C \approx 10^{-3} \begin{pmatrix} 1.4082 & 0.2142 & -0.7947 \\ 0.2142 & 0.6516 & -0.4412 \\ -0.7947 & -0.4412 & 4.8930 \end{pmatrix},$$

and the parameter uncertainties are

$$\sqrt{\mathrm{var}[\widehat{\mu}]} = \sqrt{C_{11}} \approx 0.038, \qquad \sqrt{\mathrm{var}[\widehat{\sigma}]} = \sqrt{C_{22}} \approx 0.026, \qquad \sqrt{\mathrm{var}[\widehat{\xi}]} = \sqrt{C_{33}} \approx 0.070.$$

References

1. H. Cramér, *Mathematical Methods of Statistics* (Princeton University Press, Princeton, 1946)
2. C.R. Rao, Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 81 (1945)
3. S. Brandt, *Data Analysis*, 4th edn. (Springer, Berlin, 2014)

4. A.G. Frodesen, O. Skjeggestad, H. Tøfte, *Probability and Statistics in Particle Physics* (Universitetsforlaget, Bergen, 1979)
5. S. Coles, *An Introduction to Statistical Modeling of Extreme Values* (Springer, Berlin, 2001)
6. R.L. Smith, J.C. Naylor, A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. Appl. Stat. **36**, 358 (1987)

Chapter 9
Method of Least Squares

Abstract The method of least squares is the basic tool of developing and verifying models by fitting theoretical curves to data. Fitting functions that linearly depend on model parameters (linear regression) is treated first, discussing the distinct cases of known and unknown experimental uncertainties, finding confidence intervals for the optimal parameters, and estimating the quality of the fit. Regression with standard and orthogonal polynomials, straight-line fitting and fitting a constant are analyzed separately. Linear regression for binned data, linear regression with constraints, general linear regression by using singular-value decomposition, and robust linear regression are presented, followed by a discussion of non-linear regression.

Almost on a daily basis one encounters the problem of fitting a chosen function to the pairs $\{(x_i, y_i)\}_{i=1}^{n}$, i.e. the values y_i measured at points x_i, arranged in vectors

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}, \quad \boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}. \tag{9.1}$$

The observations y_i may be correlated. These correlations—to be precise, the estimates of correlations—can of course only be determined by multiple (M-fold) measurement of the whole set $\boldsymbol{y}$ at the same $\boldsymbol{x}$. The obtained variances and covariances can be stored in the *sample covariance matrix*

$$\Sigma_{\boldsymbol{y}} = \frac{1}{M}\sum_{m=1}^{M}\big(\boldsymbol{y}_m - \overline{\boldsymbol{y}}\big)\big(\boldsymbol{y}_m - \overline{\boldsymbol{y}}\big)^{\mathrm{T}}, \qquad \overline{\boldsymbol{y}} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{y}_m,$$

where $(\Sigma_{\boldsymbol{y}})_{ij} \approx \mathrm{cov}[Y_i, Y_j]$. If the measurements are independent, the covariance matrix is diagonal,

$$\Sigma_{\boldsymbol{y}} = \mathrm{diag}\big(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\big), \tag{9.2}$$

where σ_i is the uncertainty of the individual y_i. The function f being used to fit the data contains a certain set of model parameters. The final estimates for their values should be sensitive to the precision or uncertainty of the data. Searching for the appropriate model function f is called *regression*.

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_9

9.1 Linear Regression

Linear regression means that the model function linearly depends on the parameters $\boldsymbol{\theta}$. This kind of regression is used when we seek a polynomial

$$f(x) = \theta_1 + \theta_2 x + \cdots + \theta_p x^{p-1} \tag{9.3}$$

that best fits the data (9.1) or, say, a function of the form

$$f(x) = \theta_1 + \theta_2\,\mathrm{e}^x + \theta_3 \sin x. \tag{9.4}$$

We assume that each value y_i is a realization of a random variable, distributed about the unknown true value with the uncertainty σ_i. For the parameter set $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_p\}$ one therefore generally writes

$$f_i = f\big(x_i; \boldsymbol{\theta}\big) = \sum_{j=1}^{p} a_{ij}(x_i)\theta_j, \qquad i = 1, 2, \ldots, n, \qquad p \le n,$$

or, in matrix form,

$$\boldsymbol{f} = A(\boldsymbol{x})\boldsymbol{\theta}.$$

Here $\boldsymbol{f}$ is a vector of dimension n and A is a $n \times p$ matrix with elements A_{ij} that in general are functions of x. In (9.4), for example, we have $A_{i1} = 1$, $A_{i2} = \mathrm{e}^{x_i}$ and $A_{i3} = \sin x_i$ for each i.

The main idea of fitting is to minimize the sum of squares $(y_i - f_i)^2$ with respect to the uncertainties σ_i. This is the core of the *method of least squares*. (Regression, however, is not a uniquely solvable problem, as many other measures of deviation of y_i from f_i exist; least squares just happen to be by far the most popular.) One therefore tries to find the parameters $\boldsymbol{\theta}$ minimizing the quadratic form

$$X^2 = \big(\boldsymbol{y} - A\boldsymbol{\theta}\big)^{\mathrm{T}} \Sigma_{\boldsymbol{y}}^{-1} \big(\boldsymbol{y} - A\boldsymbol{\theta}\big) \tag{9.5}$$

or, in the case of uncorrelated uncertainties (9.2),

$$X^2 = \sum_{i=1}^{n} \frac{\big(y_i - f(x_i; \boldsymbol{\theta})\big)^2}{\sigma_i^2}. \tag{9.6}$$

The deviation of y_i from the model value $f_i = f(x_i)$ is "punished" inversely proportional to the absolute error of y_i. The measure of deviation X^2 is minimized when its minimum is found by requiring

$$\frac{\partial X^2}{\partial \theta_j} = 0, \qquad j = 1, 2, \ldots, p,$$

or, in vector form, $\partial X^2/\partial\boldsymbol{\theta} = -2\big(A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}\boldsymbol{y} - A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A\boldsymbol{\theta}\big) = \mathbf{0}$. This yields the so-called *normal system* of linear equations for the parameters θ_j,

$$\big(A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A\big)\boldsymbol{\theta} = A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}\boldsymbol{y}. \tag{9.7}$$

Its solution is the vector of optimal parameters

$$\widehat{\boldsymbol{\theta}} = \Xi\,\boldsymbol{y}, \qquad \Xi = \big(A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A\big)^{-1}A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}.$$

What are the uncertainties (variances) and covariances of components of $\widehat{\boldsymbol{\theta}}$? In other words, what is the connection between the $p \times p$ covariance matrix of parameters $\boldsymbol{\theta}$ and the $n \times n$ covariance matrix of the values $\boldsymbol{y}$? We use $D = \Xi$ in the error-propagation formula (4.2.7) to derive

$$\begin{aligned}\Sigma_{\widehat{\boldsymbol{\theta}}} = \Xi\Sigma_{\boldsymbol{y}}\Xi^{\mathrm{T}} &= \big(A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A\big)^{-1}A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}\Sigma_{\boldsymbol{y}}\Sigma_{\boldsymbol{y}}^{-1}A\big(A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A\big)^{-1} \\ &= \big(A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A\big)^{-1}\big(A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A\big)\big(A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A\big)^{-1} = \big(A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A\big)^{-1}.\end{aligned} \tag{9.8}$$

Therefore

$$\operatorname{var}\big[\widehat{\theta}_j\big] = \big(\Sigma_{\widehat{\boldsymbol{\theta}}}\big)_{jj}, \quad \operatorname{cov}\big[\widehat{\theta}_j, \widehat{\theta}_k\big] = \big(\Sigma_{\widehat{\boldsymbol{\theta}}}\big)_{jk}, \quad \operatorname{corr}\big[\widehat{\theta}_j, \widehat{\theta}_k\big] = \frac{\big(\Sigma_{\widehat{\boldsymbol{\theta}}}\big)_{jk}}{\sqrt{\big(\Sigma_{\widehat{\boldsymbol{\theta}}}\big)_{jj}}\sqrt{\big(\Sigma_{\widehat{\boldsymbol{\theta}}}\big)_{kk}}}, \tag{9.9}$$

where $j, k = 1, 2, \ldots, p$. The estimate $\widehat{\boldsymbol{\theta}}$ is unbiased, $E\big[\widehat{\boldsymbol{\theta}}\big] = \boldsymbol{\theta}$. Because the relation between $\boldsymbol{\theta}$ and $\boldsymbol{y}$ is linear, the Gauss–Markov theorem [1] also tells us that it has the smallest possible variance.

Note that the dependence of the measure of deviation on parameters $\boldsymbol{\theta}$ has the general form

$$X^2(\boldsymbol{\theta}) = X^2\big(\widehat{\boldsymbol{\theta}}\big) + \big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\big)^{\mathrm{T}}\Sigma_{\widehat{\boldsymbol{\theta}}}^{-1}\big(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\big).$$

In the case of a single parameter θ and a *constant* covariance matrix of observations $\Sigma_{\widehat{\boldsymbol{\theta}}}^{-1}$ one sees that X^2 has a parabolic shape:

$$X^2(\theta) = X^2\big(\widehat{\theta}\big) + \big(\operatorname{var}\big[\widehat{\theta}\big]\big)^{-1}\big(\theta - \widehat{\theta}\big)^2, \qquad \operatorname{var}\big[\widehat{\theta}\big] = 2\left(\frac{\partial^2 X^2}{\partial\theta^2}\right)^{-1}. \tag{9.10}$$

In the following we shall describe the fitting of functions to data in a pedagogically rather non-orthodox sequence: polynomials, orthogonal polynomials, straight line, constant. We assume throughout that the observations $\boldsymbol{y}$ are independent, so that their covariance matrix is given by (9.2).

9.1.1 Fitting a Polynomial, Known Uncertainties

When (9.6) is minimized with respect to θ_j ($j = 1, 2, \ldots, p$) with model (9.3), the system (9.7) becomes

$$\left(V^{\mathrm{T}} W V\right)\boldsymbol{\theta} = V^{\mathrm{T}} W \boldsymbol{y}.$$

Here V is an $n \times p$ Vandermonde matrix with the elements

$$V_{ij} = x_i^{j-1}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p, \tag{9.11}$$

while $W = \mathrm{diag}\left(1/\sigma_1^2, 1/\sigma_2^2, \ldots, 1/\sigma_n^2\right)$ is the weight matrix. Denoting $B = V^{\mathrm{T}} W V$ ($p \times p$ matrix, $B = B^{\mathrm{T}}$) and $\boldsymbol{b} = V^{\mathrm{T}} W \boldsymbol{y}$ (p-dimensional vector) the system can be rewritten as

$$B\boldsymbol{\theta} = \boldsymbol{b} \tag{9.12}$$

or

$$\sum_{j=1}^{p} B_{kj}\theta_j = b_k, \qquad B_{kj} = \sum_{i=1}^{n} \frac{x_i^{k+j-2}}{\sigma_i^2}, \qquad b_k = \sum_{i=1}^{n} \frac{x_i^{k-1} y_i}{\sigma_i^2}, \tag{9.13}$$

where $j, k = 1, 2, \ldots, p$. The solution of (9.12) is the vector of optimal parameters,

$$\widehat{\boldsymbol{\theta}} = B^{-1}\boldsymbol{b}, \tag{9.14}$$

while their variances and covariances are

$$\mathrm{var}\left[\widehat{\theta}_j\right] = \left(B^{-1}\right)_{jj}, \qquad \mathrm{cov}\left[\widehat{\theta}_j, \widehat{\theta}_k\right] = \left(B^{-1}\right)_{jk}. \tag{9.15}$$

Example An experiment results in the angular distribution of scattered particles, shown in the Table below and in Fig. 9.1. The independent variables x_i are the cosines of the scattering angles and are "almost exactly known": a measurement at x_i involves angles on the interval around $\cos\phi_i$ which is much smaller than the distance between the neighboring points, $|x_i - x_{i-1}|$. The dependent variables y_i are the numbers of detected particles at given angle. The uncertainties σ_i change with the angle and increase in the backward direction.

$x_i = \cos\phi_i$	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
$y_i = N_i$	1412	908	881	534	501	352	218	355	278	482
σ_i	320	205	180	160	120	103	90	88	80	76

Let us fit the data by a parabola ($p = 3$). Formulas (9.13) yield

$$\boldsymbol{b} = \begin{pmatrix} 0.351 \\ 0.0890 \\ 0.134 \end{pmatrix}, \qquad B = 10^{-3} \begin{pmatrix} 8.492 & 3.167 & 2.926 \\ 3.167 & 2.926 & 1.791 \\ 2.926 & 1.791 & 1.746 \end{pmatrix},$$

and solving the system (9.14) gives the optimal parameters

$$\widehat{\boldsymbol{\theta}} = \bigl(376.11,\ -501.45,\ 650.94\bigr)^{\mathrm{T}}.$$

Their covariance matrix (see (9.15)) is

$$B^{-1} = \begin{pmatrix} 2808.1 & -423.55 & -4272.8 \\ -423.55 & 9251.0 & -8782.2 \\ -4272.8 & -8782.2 & 21902.9 \end{pmatrix}.$$

The uncertainties of parameters $\widehat{\boldsymbol{\theta}}$, i.e. the effective deviations $\Delta\theta_i = \sqrt{\mathrm{var}\bigl[\widehat{\theta}_i\bigr]}$, can be read off from its diagonal elements:

$$\Delta\widehat{\theta}_1 = \sqrt{\bigl(B^{-1}\bigr)_{11}} = 52.99, \quad \Delta\widehat{\theta}_2 = \sqrt{\bigl(B^{-1}\bigr)_{22}} = 96.18, \quad \Delta\widehat{\theta}_3 = \sqrt{\bigl(B^{-1}\bigr)_{33}} = 148.0.$$

The parabola with these parameters that fits the data optimally is shown in Fig. 9.1. The correlations between the calculated parameters are most clearly identified in the correlation matrix ρ with the elements

$$\rho_{jk} = \frac{\mathrm{cov}\bigl[\widehat{\theta}_j, \widehat{\theta}_k\bigr]}{\sqrt{\mathrm{var}\bigl[\widehat{\theta}_j\bigr]}\sqrt{\mathrm{var}\bigl[\widehat{\theta}_k\bigr]}} = \frac{\bigl(B^{-1}\bigr)_{jk}}{\sqrt{\bigl(B^{-1}\bigr)_{jj}}\sqrt{\bigl(B^{-1}\bigr)_{kk}}},$$

hence

$$\rho = \begin{pmatrix} 1 & -0.0831 & -0.5448 \\ -0.0831 & 1 & -0.6170 \\ -0.5448 & -0.6170 & 1 \end{pmatrix}.$$

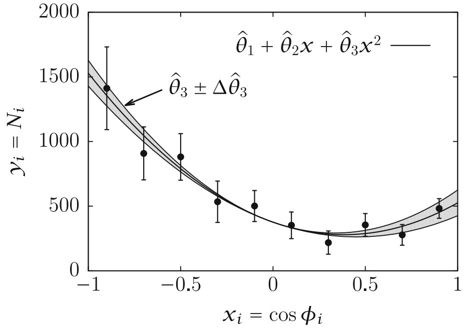


Fig. 9.1 Distribution of particles with respect to scattering angles (angular distribution), together with the best-fit parabola. The *shading* indicates the area corresponding to the variation of the parabola's leading coefficient by one effective deviation

The diagonal elements (auto-correlations of parameters $\widehat{\theta}_i$) are equal to 1. The parameter cross-correlations are given by the off-diagonal terms: a smaller absolute value implies a smaller correlation. A positive sign means proportionality (correlation), a negative sign implies anti-correlation. ◁

9.1.2 Fitting Observations with Unknown Uncertainties

So far we have assumed that the uncertainties of $\boldsymbol{y}$ were known, providing each observation y_i with an error σ_i having zero mean, that is,

$$\boldsymbol{y} = A\boldsymbol{\theta} + \boldsymbol{\sigma}, \qquad E\big[\boldsymbol{\sigma}\big] = \boldsymbol{0}, \qquad E\big[\boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathrm{T}}\big] = \Sigma_{\boldsymbol{y}},$$

where $\Sigma_{\boldsymbol{y}}$ is the covariance matrix. If the uncertainties are unknown, one usually assumes that the variance is constant for all observations, σ^2, and that the observations are independent. Then the covariance matrix is simply $\Sigma_{\boldsymbol{y}} = \sigma^2 I_n$, where I_n is an $n \times n$ unit matrix. In this case the optimal parameters $\boldsymbol{\theta}$ can be calculated without even knowing σ^2 since it follows from (9.7) that

$$\widehat{\boldsymbol{\theta}} = \big(A^{\mathrm{T}}A\big)^{-1}A^{\mathrm{T}}\boldsymbol{y}. \tag{9.16}$$

How can one nevertheless still estimate σ^2 and the uncertainties of $\widehat{\boldsymbol{\theta}}$? We compute the *sum of squared residuals* (SSR)

$$\mathrm{SSR} = X^2_{\mathrm{min}} = \big(\boldsymbol{y} - A\widehat{\boldsymbol{\theta}}\big)^{\mathrm{T}}\big(\boldsymbol{y} - A\widehat{\boldsymbol{\theta}}\big),$$

measuring the deviation of the model $A\boldsymbol{\theta}$ (evaluated with optimal parameters) from the observations $\boldsymbol{y}$. The unbiased estimator for the unknown variance σ^2 is then

$$\widehat{\sigma^2} = \frac{\mathrm{SSR}}{n-p}, \tag{9.17}$$

while the covariance matrix of the optimal parameters is

$$\Sigma_{\widehat{\boldsymbol{\theta}}} = \frac{\mathrm{SSR}}{n-p}\big(A^{\mathrm{T}}A\big)^{-1}, \tag{9.18}$$

which can be compared to (9.8) when the uncertainties were known. Variances, covariances and correlation of optimal parameters are then obtained from (9.9).

Long example Fig. 9.2 (top left) shows the deviation of the global average temperature of the Earth surface from the average value obtained between 1951 and 1980, the so-called temperature anomaly [2]. The circles denote the average annual anomalies

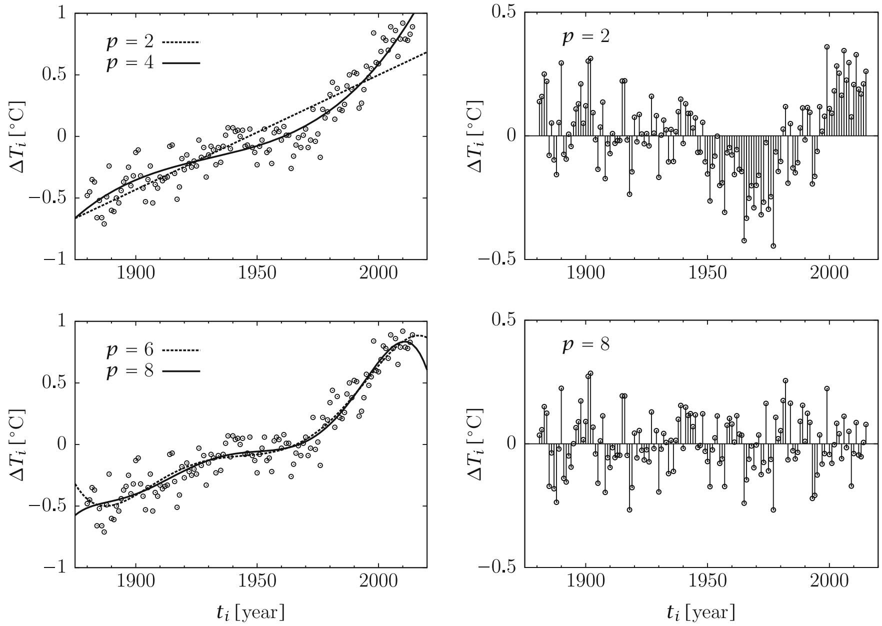


Fig. 9.2 [Left top and bottom] Time dependence of the temperature anomaly (Earth surface) and best-fit polynomials of odd degrees. [Right] Deviations (residuals) $y_i - f(x_i)$, calculated with the optimal parameter values $\boldsymbol{\theta}$ in the case [Top] $p = 2$ and [Bottom] $p = 8$. Notice how the residuals have shrunk and become more random

as a function of time (135 observations between 1880 and 2014). The uncertainties of $\Delta T_i = y_i$ are unknown.

It is useful to shift the origin, $t_i \to x_i = t_i - 1880$, so that one can work with smaller numbers on the interval [0, 134]. The data are fitted by polynomials (9.3) of various odd degrees ($p = 2$, 4, 6 and 8). The matrix A in formula (9.16) is of the Vandermonde form (9.11). The calculated optimal parameters minimizing the measure of deviation X^2 are listed in Table 9.1. The corresponding minimal values X^2_{min} are given in the second column of Table 9.2.

How can we judge whether the chosen model function is “good”? The basic diagnostic tool are the differences between y_i and $f(x_i)$ once the minimization of X^2 has been done, that is, the *residuals*

$$y_i - f\big(x_i; \widehat{\boldsymbol{\theta}}\big).$$

The distribution of residuals should be as random as possible. The residuals of the linear fit ($p = 2$) are shown in Fig. 9.2 (top right). This is clearly unsatisfactory as the residuals are large and even exhibit some sort of oscillations. When the fit degree is increased, the residuals shrink and tend to become more and more random, while X^2 keeps on dropping. However, one should not perpetuate this as it is wise to describe

Table 9.1 Values of optimal parameters $\widehat{\boldsymbol{\theta}}$ (units of $\widehat{\theta}_i$ are °C/yr^{i-1}) in regression analysis of the temperature anomaly with polynomials of various degrees

p	$\widehat{\theta}_1$	$10^3\widehat{\theta}_2$	$10^3\widehat{\theta}_3$	$10^5\widehat{\theta}_4$	$10^7\widehat{\theta}_5$	$10^9\widehat{\theta}_6$	$10^{11}\widehat{\theta}_7$	$10^{13}\widehat{\theta}_8$
2	−0.618	9.311						
4	−0.583	15.02	−0.206	0.135				
6	−0.442	−16.71	1.424	−3.048	2.626	−0.772		
8	−0.515	8.354	−0.536	3.072	−6.475	5.875	−2.147	0.195

See also Table 9.2

Table 9.2 Minimal values X^2_{min}, the coefficient of determination R and the estimate of variance $\widehat{\sigma}^2$ (9.17) in polynomial regression of the temperature anomaly

p	X^2_{min}	R	$\widehat{\sigma}^2$
2	3.688	0.828	0.0278
4	2.095	0.902	0.0160
6	1.859	0.913	0.0144
8	1.792	0.917	0.0141

the observations with as few parameters as possible: any data $\{(x_i, y_i)\}_{i=1}^{p}$ can be even exactly interpolated by a polynomial of degree $p-1$, but such fits invariably become too “wild”: signs of such behaviour can be glimpsed already at $p=8$ near the edges of Fig. 9.2 (bottom left).

Even if X^2_{min} drops when p is increased that does not necessarily mean that a high-degree polynomial is better than a lower-degree polynomial. Indeed such a function does a better job in describing the variation in the data, but it is unclear how much of it can be assigned to the uncertainties of observations and how much to the model having too many parameters. A good measure of the variance that can be assigned solely to the model is the *coefficient of determination*

$$R = 1 - \frac{X^2_{\mathrm{min}}/n}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2/n},$$

where the denominator is the total variance of the observations about their mean. A better regression results in the values of R being closer to 1. The values of R for our Example are given in the third column of Table 9.2. A seventh-degree polynomial ($p=8$), for instance, describes 91.7% of the data variation.

Two more tools are at hand to evaluate the relevance of individual parameters θ_j for the description of data. The first tool are their correlations, obtained by the usual formula (9.9). Fitting with $p=4$, say, gives the correlation matrix

$$\rho = \begin{pmatrix} 1 & -0.860 & 0.735 & -0.650 \\ -0.860 & 1 & -0.967 & 0.915 \\ 0.735 & -0.967 & 1 & -0.986 \\ -0.650 & 0.915 & -0.986 & 1 \end{pmatrix},$$

indicating strong correlations, in particular between θ_2 and θ_3 as well as θ_3 and θ_4. The other tool are the ratios of absolute parameter values and their variances—a kind of "signal-to-noise" ratios—

$$\xi_j = |\widehat{\theta}_j| \Big/ \sqrt{\mathrm{var}[\widehat{\theta}_j]}\,,$$

where the variances are obtained from the diagonal elements of the covariance matrix (9.18). If a parameter θ_j is thought to be statistically relevant, the value ξ_j should be much larger than 1. In fitting with $p = 4$ we get $\xi_1 = 13.8$, $\xi_2 = 5.47$, $\xi_3 = 4.31$ and $\xi_4 = 5.75$, while for $p = 6$ we get $\xi_1 = 7.60$, $\xi_2 = 1.88$, $\xi_3 = 3.43$, $\xi_4 = 3.87$, $\xi_5 = 4.04$ and $\xi_6 = 4.00$, indicating that parameter θ_2 might be superfluous. ◁

9.1.3 Confidence Intervals for Optimal Parameters

Confidence intervals for parameters $\boldsymbol{\theta}$ are calculated by analogy to the confidence intervals for the sample mean (Sect. 7.3.1). If the uncertainties σ_i of y_i are unknown—yet assumed to be independent and identically distributed—we first compute the covariance matrix (9.18) and extract the variances

$$\mathrm{var}[\widehat{\theta}_j] = \left(\Sigma_{\widehat{\theta}}\right)_{jj}, \qquad j = 1, 2, \ldots, p.$$

At chosen confidence level $1 - \alpha$, parameter θ_j then has the confidence interval

$$\left[\widehat{\theta}_j - t_*\sqrt{\mathrm{var}[\widehat{\theta}_j]},\, \widehat{\theta}_j + t_*\sqrt{\mathrm{var}[\widehat{\theta}_j]}\right] \tag{9.19}$$

on which the unknown true value θ_j lies. The critical value t_* can be read off from Table 7.1, taking into account $\nu = n - p$ degrees of freedom. In other words,

$$P\left(\widehat{\theta}_j - t_*\sqrt{\mathrm{var}[\widehat{\theta}_j]} \le \theta_j \le \widehat{\theta}_j + t_*\sqrt{\mathrm{var}[\widehat{\theta}_j]}\right) = 1 - \alpha.$$

If $n - p \gg 1$, the t distribution turns into the standardized normal distribution and t_* can be replaced by z_* from the last row of Table 7.1.

Example In the previous Example we have seen that in fitting a fifth-degree polynomial ($p = 6$) the least reliable parameter was θ_2, for which we got

$$\widehat{\theta}_2 = -0.01671, \qquad \xi_2 = |\widehat{\theta}_2| \Big/ \sqrt{\mathrm{var}\big[\widehat{\theta}_2\big]} = 0.01671/0.00889 \approx 1.88.$$

Let us choose a confidence level of $1 - \alpha = 0.99$. The sample is large, $n - p = 135 - 6 = 129$, so the limit the of normal distribution is justified and we can simply take $t_* = z_* = 2.576$. The confidence interval (9.19) for θ_2 is then

$$\big[-0.01671 - 2.576 \cdot 0.00889 \le \theta_2 \le -0.01671 + 2.576 \cdot 0.00889\big],$$

thus $P\big(-0.0396 \le \theta_2 \le 0.0062\big) = 0.99$. ◁

9.1.4 How "Good" Is the Fit?

If the uncertainties of the n observations are mutually independent and normally distributed, and if the p regression parameters are also independent, the minimized sum of squared residuals is χ^2-distributed (3.21) with $\nu = n - p$ degrees of freedom,

$$X^2_{\mathrm{min}} = \sum_{i=1}^{n} \frac{\big(y_i - [A\widehat{\boldsymbol{\theta}}]_i\big)^2}{\sigma_i^2} \sim \chi^2(n-p).$$

Therefore

$$G = \int_{X^2_{\mathrm{min}}}^{\infty} f_{\chi^2}(x;\nu)\,\mathrm{d}x = 1 - F_{\chi^2}\big(X^2_{\mathrm{min}};\nu\big),$$

where F_{χ^2} is the distribution function. This formula is used to quantify the *goodness of fit*. A small X^2_{min} or large G indicate a "good fit", large X^2_{min} and small G imply that a fit is "bad". See also Sect. 10.3.

9.1.5 Regression with Orthogonal Polynomials ⋆

The matrix B in (9.12) tends to be ill-conditioned: its condition number exponentially grows with dimension p, so $\kappa(B) \approx C\exp(\alpha p)$, see Sect. 3.2.5 in [3]. At desired precision ε of parameters θ_j the procedure described above can accommodate polynomials of degree $p \lesssim (1/\alpha)\log(\varepsilon_{\mathrm{M}}/C\varepsilon)$, where ε_{M} is the machine precision. In `double` precision this usually means $p \lesssim 10$.

One can at least partly avoid these stability problems if the points $\{x_i\}_{i=1}^{n}$ coincide with the definition domain of some system of orthogonal polynomials. Most suitable for regression purposes are the *discrete-variable orthogonal polynomials* $\{p_j(x) : \deg(p_j) = j\}_{j=0}^{p}$ which are mutually independent and orthogonal on a discrete set of points $\{x_i\}$ in the sense

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2} p_j(x_i) p_k(x_i) = A_j \delta_{j,k}$$

with some weight $1/\sigma_i^2$. A model function may be devised as a linear combination

$$f(x; \boldsymbol{\theta}) = \sum_{j=0}^{p} \theta_j p_j(x),$$

where the expansion coefficients θ_j are unknown parameters. They are again determined such that the measure of deviation (9.6) is minimal. We get

$$X^2 = \sum_{j=0}^{p} \theta_j^2 A_j - 2 \sum_{j=0}^{p} \theta_j B_j + C, \qquad B_j = \sum_{i=1}^{n} \frac{p_j(x_i) y_i}{\sigma_i^2}, \qquad C = \sum_{i=1}^{n} \frac{y_i^2}{\sigma_i^2}.$$

The condition for a minimum, $\partial X^2 / \partial \theta_j = 0$, gives $2\theta_j A_j - 2B_j = 0$ or

$$\widehat{\theta}_j = \frac{B_j}{A_j} \qquad \text{and} \qquad X^2 = C - \sum_{j=0}^{p} \frac{B_j^2}{A_j} = \min.$$

Example The most popular system of discrete-variable orthogonal polynomials are the Chebyshev polynomials of the first kind, $T_j(x)$: see Sect. 4.3 in [3]. A linear combination $f(x) = \sum_{j=0}^{p} \theta_j T_j(x)$ of these polynomials minimizes the measure $X^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$ with the nodes $x_i = \cos\bigl(\pi(2i-1)/(2n)\bigr)$ and coefficients

$$\theta_0 = \frac{1}{n} \sum_{i=1}^{n} y(x_i), \qquad \theta_j = \frac{2}{n} \sum_{i=1}^{n} y(x_i) T_j(x_i), \qquad j = 1, 2, \ldots, p,$$

which is known as the Chebyshev approximation. The problem is well defined for $p + 1 \leq n$. In the limiting case $p + 1 = n$ the function f interpolates the data y_i and the measure of deviation is $X^2 = 0$. ◁

9.1.6 Fitting a Straight Line

Seeking a straight line $f(x) = \theta_1 + \theta_2 x$ fitting the data $\{(x_i, y_i)\}_{i=1}^{n}$ with known uncertainties σ_i is the most common two-parameter linear regression. When the measure

X^2 (9.6) is minimized with respect to θ_1 and θ_2, we obtain an analytically solvable system

$$\theta_1 S + \theta_2 S_x = S_y,$$
$$\theta_1 S_x + \theta_2 S_{xx} = S_{xy},$$

where we have denoted

$$S = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}, \quad S_x = \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}, \quad S_{xx} = \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2}, \quad S_{xy} = \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2}, \quad S_y = \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}.$$

The matrix B from (9.12) is immediately recognizable, as well as its inverse,

$$B = \begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix}, \qquad B^{-1} = \frac{1}{\det B} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}, \tag{9.20}$$

where we assume that $\det B = S_{xx}S - S_x^2 \neq 0$. The coefficients $\widehat{\theta}_1$ and $\widehat{\theta}_2$ that minimize X^2 are

$$\widehat{\theta}_1 = \frac{S_{xx}S_y - S_x S_{xy}}{S_{xx}S - S_x^2}, \qquad \widehat{\theta}_2 = \frac{S S_{xy} - S_x S_y}{S_{xx}S - S_x^2}.$$

A sample fit is shown in Fig. 9.3 (left).

For constant uncertainties ($\sigma_i = \sigma$) the parameters of the straight line are

$$\widehat{\theta}_1 = \overline{y} - \widehat{\theta}_2 \overline{x}, \qquad \widehat{\theta}_2 = \frac{s_{xy}}{s_x^2},$$

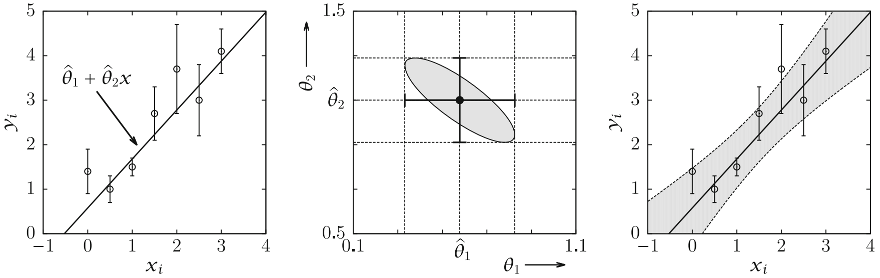


Fig. 9.3 Fitting a straight line to the data $y_i = 1.4, 1.0, 1.5, 2.7, 3.7, 3.0, 4.1$ with errors $\sigma_i = 0.5, 0.3, 0.2, 0.6, 1.0, 0.8, 0.5$ at $x_i = 0, 0.5, 1, 1.5, 2, 2.5, 3$ by using the method of least squares. [Left] The straight line that minimizes the measure X^2 ($\widehat{\theta}_1 = 0.578$, $\widehat{\theta}_2 = 1.100$). [Center] Covariance ellipse with the center at $(\widehat{\theta}_1, \widehat{\theta}_2)$ and uncertainties $\pm\sigma(\widehat{\theta}_1) = \pm 0.247$ and $\pm\sigma(\widehat{\theta}_2) = \pm 0.190$. Parameters within the ellipse correspond to straight lines contained in the shaded area of the [Right] panel. The straight line corresponding to the true θ_1 and θ_2 has a probability $1 - \mathrm{e}^{-1/2}$ of lying within this area

where $\overline{x} = \left(\sum_{i=1}^{n} x_i\right)/n$ and $\overline{y} = \left(\sum_{i=1}^{n} y_i\right)/n$ are the arithmetic means of the observations, while

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \qquad \text{and} \qquad s_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

are their sample variance and covariance. We see that the straight line goes through the "center of gravity" $(\overline{x}, \overline{y})$ of the data. From (9.15) and (9.20) we also obtain the variances of the parameters $\widehat{\theta}_1$ and $\widehat{\theta}_2$, which do not depend on the position of the points along the y-axis:

$$\operatorname{var}\big[\widehat{\theta}_1\big] = \left(B^{-1}\right)_{11} = \frac{S_{xx}}{S_{xx}S - S_x^2}, \qquad \operatorname{var}\big[\widehat{\theta}_2\big] = \left(B^{-1}\right)_{22} = \frac{S}{S_{xx}S - S_x^2}. \qquad (9.21)$$

The off-diagonal elements of the covariance matrix B^{-1} for estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are

$$\operatorname{cov}\big[\widehat{\theta}_1, \widehat{\theta}_2\big] = \left(B^{-1}\right)_{12} = \left(B^{-1}\right)_{21} = \frac{-S_x}{S_{xx}S - S_x^2},$$

so that the linear correlation coefficient between $\widehat{\theta}_1$ and $\widehat{\theta}_2$ is equal to

$$\widehat{\rho} = \frac{\operatorname{cov}\big[\widehat{\theta}_1, \widehat{\theta}_2\big]}{\sqrt{\operatorname{var}\big[\widehat{\theta}_1\big]}\sqrt{\operatorname{var}\big[\widehat{\theta}_2\big]}}.$$

Let us denote

$$\widehat{\sigma}_1 = \sqrt{\operatorname{var}\big[\widehat{\theta}_1\big]}, \quad \widehat{\sigma}_2 = \sqrt{\operatorname{var}\big[\widehat{\theta}_2\big]}.$$

(Do not confuse these with the uncertainties σ_i of the observations y_i.) The estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$, their uncertainties $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$, together with the correlation coefficient $\widehat{\rho}$, define a covariance ellipse centered at $(\widehat{\theta}_1, \widehat{\theta}_2)$ with semi-axes r_1 and r_2, rotated by angle α in the (θ_1, θ_2)-plane. The ellipse parameters are given by the formulas [4]

$$\begin{aligned} \tan 2\alpha &= 2\widehat{\rho}\widehat{\sigma}_1\widehat{\sigma}_2 \,/\, \left(\widehat{\sigma}_1^2 - \widehat{\sigma}_2^2\right), \\ r_1^2 &= \widehat{\sigma}_1^2\widehat{\sigma}_2^2\left(1 - \widehat{\rho}^2\right) \Big/ \left[\widehat{\sigma}_1^2 \sin^2\alpha - \widehat{\rho}\widehat{\sigma}_1\widehat{\sigma}_2 \sin 2\alpha + \widehat{\sigma}_2^2 \cos^2\alpha\right], \\ r_2^2 &= \widehat{\sigma}_1^2\widehat{\sigma}_2^2\left(1 - \widehat{\rho}^2\right) \Big/ \left[\sigma_1^2 \cos^2\alpha + \widehat{\rho}\widehat{\sigma}_1\widehat{\sigma}_2 \sin 2\alpha + \widehat{\sigma}_2^2 \sin^2\alpha\right]. \end{aligned}$$

An example of a covariance ellipse is shown in Fig. 9.3 (center). The points within the covariance ellipse define a bundle of straight lines, indicated by the shaded area in Fig. 9.3 (right).

Let us show how the uncertainties $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$ change with increasing number of points n. Suppose that the points $\{x_i\}_{i=1}^{n}$ are uniformly distributed on $[\alpha, \beta]$, so that $x_i = \alpha + (i-1)\Delta x$ with $\Delta x = (\beta - \alpha)/(n-1)$, and that each observation has an error of $\sigma_i = \sigma$. We calculate $S = n/\sigma^2$, $S_x = n(\alpha+\beta)/(2\sigma^2)$ and $S_{xx} = n[(\alpha^2+\alpha\beta+\beta^2)(2n-1) - 3\alpha\beta]/[6(n-1)\sigma^2]$. From (9.21) we then extract the asymptotic behaviour in the $n \to \infty$ limit,

$$\widehat{\sigma}_1^2 \sim \frac{4(\alpha^2+\alpha\beta+\beta^2)\sigma^2}{n(\alpha-\beta)^2} + \mathcal{O}\left(\frac{1}{n^2}\right), \qquad \widehat{\sigma}_2^2 \sim \frac{12\sigma^2}{n(\alpha-\beta)^2} + \mathcal{O}\left(\frac{1}{n^2}\right).$$

Therefore, with increasing n, the straight-line parameters $\widehat{\theta}_1$ and $\widehat{\theta}_2$ gain in precision just as any other statistical average, i.e. $\widehat{\sigma}_1, \widehat{\sigma}_2 \sim \mathcal{O}(n^{-1/2})$.

9.1.7 Fitting a Straight Line with Uncertainties in both Coordinates

In linear regression with a straight line $f(x) = \theta_1 + \theta_2 x$, where the values in *both* y_i and x_i possess uncertainties, we strive to minimize the measure

$$X^2 = \sum_{i=1}^{n} \frac{\left(y_i - \theta_1 - \theta_2 x_i\right)^2}{\sigma_{xi}^2\theta_2^2 + \sigma_{yi}^2}.$$

The determination of optimal $\widehat{\theta}_1$ and $\widehat{\theta}_2$ requires us to simultaneously fulfill the conditions $\partial X^2/\partial\theta_1 = 0$ and $\partial X^2/\partial\theta_2 = 0$. Because θ_2 enters non-linearly, the problem is non-trivial; a reliable algorithm is given in [3], p. 233.

9.1.8 Fitting a Constant

In zero-degree polynomial regression the observations y_i are fitted by a constant: we minimize (9.6) with $f(x_i) = c$. The condition $\partial X^2/\partial c = 0$ yields

$$\widehat{c} = \frac{1}{S}\sum_{i=1}^{n}\frac{y_i}{\sigma_i^2}, \qquad S = \sum_{i=1}^{n}\frac{1}{\sigma_i^2}, \qquad \mathrm{var}\big[\widehat{c}\big] = \frac{1}{S}, \tag{9.22}$$

which is precisely the weighted average (8.7). It can be used in place of the arithmetic average whenever measurements of the same quantity have different errors. The asymptotic behaviour is $(\mathrm{var}[\widehat{c}])^{1/2} \sim \mathcal{O}(n^{-1/2})$ when $n \to \infty$.

We shall see in Chap. 10 that the measure $X^2 = \sum_{i=1}^{n}(y_i - \widehat{c})^2/\sigma_i^2$ helps us quantify the assumption of the normal distribution of uncertainties σ_i. At chosen significance (risk level) α, say, $\alpha = 5\%$, we determine χ_+^2 from the equation

$$\int_{\chi_+^2}^{\infty} f_{\chi^2}(x; n-1)\,\mathrm{d}x = \alpha,$$

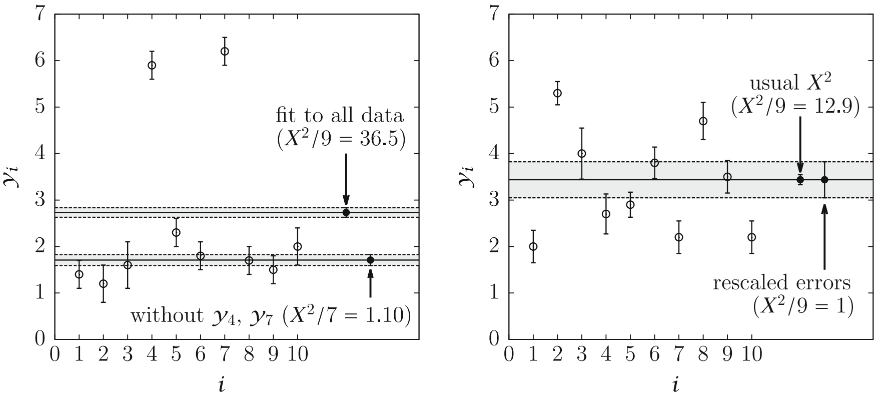


Fig. 9.4 Fitting a constant to the data $\{y_i \pm \sigma_i\}_{i=1}^n$, $n = 10$. [Left] Weighted average in the presence of two outliers (y_4 and y_7) and without them. [Right] Weighted average of data with an unexplained systematic error, resulting in an unreasonably small uncertainty of the parameter $\widehat{c}$, as well as the average involving rescaled errors (9.23)

where f_{χ^2} is the density (3.21). We compare the obtained χ^2_+ with the value of X^2 calculated from the data. If $X^2 < \chi^2_+$, the optimal value $\widehat{c}$ and its uncertainty may be considered to be consistent with the data.

Example This apparent simplicity conceals many pitfalls. Figure 9.4 (left) shows $n = 10$ observations, which we fit by a constant c. The procedure outlined above yields $\widehat{c} = 2.73 \pm 0.10$ and $X^2/(n-1) = 36.5$. With a chosen significance $\alpha = 5\%$ and $\nu = n - 1 = 9$, we use Fig. 10.6 to read off $\chi^2_+(\alpha = 5\%)/9 \approx 1.88$. Since $X^2/(n - 1) > \chi^2_+/(n-1)$, the basic premise of normal errors may be rejected. (In other words, a constant probability density is inconsistent with the measured sample with a probability much higher than α.) But if the outliers y_4 and y_7 are omitted, we get $\widehat{c} = 1.71 \pm 0.12$ and $X^2/7 = 1.10$, while $\chi^2_+(\alpha = 5\%)/7 \approx 2.01$. Now the fit appears to be consistent with the data.

A different problem is revealed in Fig. 9.4 (right). By using the method of Sect. 7.3.1 we can show that individual observations are outside of the confidence interval for the sample mean. In this case we obtain $X^2/9 = 12.9$ and again $\chi^2_+(\alpha = 5\%)/9 \approx 1.88$. But now outliers can not be blamed for a large value of X^2, as the measurements obviously include an unknown, underestimated systematic error. In such cases the measurement uncertainties may be rescaled:

$$\sigma'_i = \sigma_i \sqrt{\frac{X^2}{n-1}}, \qquad i = 1, 2, \ldots, n. \tag{9.23}$$

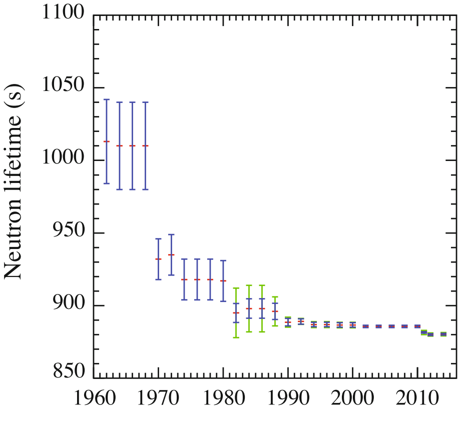
The weighted average to compute the desired parameter $\widehat{c}$ can still be formed by (9.22), but its uncertainty now becomes

$$\sqrt{\mathrm{var}[\widehat{c}]} = \sqrt{\frac{X^2}{n-1}}\left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\right)^{-1/2} .$$

This yields a more sensible result and, by construction, $X^2/(n-1) = 1$. ◁

9.1.9 *Are We Allowed to Simply Discard Some Data?*

We certainly are! Among various reasons a specific measurement or an individual observation may be simply removed, the Particle Data Group [5] lists the following: • the measurement is superseded by or included in later results; • its error is not given; • it involves questionable assumptions; • it has a poor signal-to-noise ratio, low statistical significance, or is otherwise of poorer quality than other data available; • it is clearly inconsistent with other results that appear to be more reliable; • it is not independent of other results. The figure shows the mean values of the neutron decay time, as known over the years 1960–2015. (There were several independent experiments each year; shown are the annual averages.) Think about which of the values shown in the plot should be trusted today, based on the above criteria!



9.2 Linear Regression for Binned Data

Often the data are *histogrammed* or *binned* in the x variable. This means that n observations $\{x_1, x_2, \ldots, x_n\}$ are classified into N mutually exclusive classes or *bins*. The ith bin then contains $y_i = n_i$ observations. Figure 9.5 (left) shows an example:

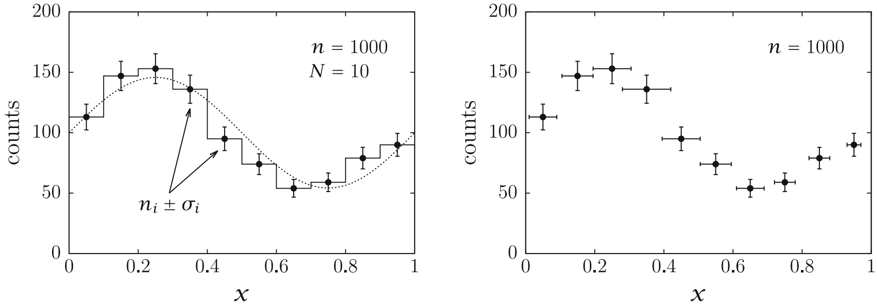


Fig. 9.5 [Left] Histogram of $n = 1000$ counts arranged in $N = 10$ bins. [Right] The same data set, but with specified uncertainties in the x variable which partially overlap (the classes are not mutually exclusive). These two arrangements are not equivalent!

the first bin $x \in [0.0, 0.1]$ has $n_1 = 113$ counts, the second bin $x \in [0.1, 0.2]$ has $n_2 = 147$, and so on.

Let the probability of a certain value landing in the ith bin be $p_i(\boldsymbol{\theta})$. Here $\boldsymbol{\theta}$ is the parameter set that determines the model distribution of the observations. The expected number of counts in the ith bin is therefore $f_i(\boldsymbol{\theta}) = np_i(\boldsymbol{\theta})$. The histogram contains all observations, hence

$$\sum\nolimits_{i=1}^{N} n_i = \sum\nolimits_{i=1}^{N} np_i\big(\boldsymbol{\theta}\big) = n. \tag{9.24}$$

In Sect. 5.2 we have demonstrated that the distribution of n_i in N bins is multinomial, with the covariance matrix

$$\Sigma = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_N \\ -np_2p_1 & np_2(1-p_2) & \cdots & -np_2p_N \\ \vdots & \vdots & \ddots & \vdots \\ -np_Np_1 & -np_Np_2 & \cdots & np_N(1-p_N) \end{pmatrix}.$$

Due to normalization (9.24) the matrix Σ is singular ($|\Sigma| = 0$), but the least-squares problem can still be formulated. Namely, one of the bins—say, the Nth—can be eliminated, since one always has $n_N = n - n_1 - n_2 - \cdots - n_{N-1}$. This results in an $(N-1) \times (N-1)$ non-singular matrix Σ' which we use to minimize the measure of deviation

$$X^2 = \big(\boldsymbol{y} - n\boldsymbol{p}\big)^{\mathrm{T}}\big(\Sigma'\big)^{-1}\big(\boldsymbol{y} - n\boldsymbol{p}\big) = \sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\big(n_i - np_i\big)\left(\big(\Sigma'\big)^{-1}\right)_{ij}\big(n_j - np_j\big),$$

where $\boldsymbol{y} = (n_1, n_2, \ldots, n_{N-1})^{\mathrm{T}}$ and $\boldsymbol{p} = (p_1, p_2, \ldots, p_{N-1})^{\mathrm{T}}$, and the matrix elements are $\Sigma'_{ij} = np_i(\delta_{i,j} - p_j)$, $i, j = 1, 2, \ldots, N-1$, so that $\Sigma' = n(D - \boldsymbol{pp}^{\mathrm{T}})$, where

$D^{-1}\boldsymbol{p} = \mathbf{1}$. By invoking the Sherman-Morrison formula we get

$$\left(D - \boldsymbol{p}\boldsymbol{p}^{\mathrm{T}}\right)^{-1} = D^{-1} + \frac{D^{-1}\boldsymbol{p}\boldsymbol{p}^{\mathrm{T}}D^{-1}}{1 - \boldsymbol{p}^{\mathrm{T}}D^{-1}\boldsymbol{p}} = D^{-1} + \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{p_N},$$

where we have used $1 - \boldsymbol{p}^{\mathrm{T}}D^{-1}\boldsymbol{p} = 1 - p_1 - p_2 - \cdots - p_{N-1} = p_N$. The expression for X^2 can therefore be rewritten as

$$\begin{aligned} X^2 &= \left(\boldsymbol{y} - n\boldsymbol{p}\right)^{\mathrm{T}} \frac{1}{n}\left(D^{-1} + \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{p_N}\right)\left(\boldsymbol{y} - n\boldsymbol{p}\right) \\ &= \sum_{i=1}^{N-1} \frac{(n_i - np_i)^2}{np_i} + \frac{(n_N - np_N)^2}{np_N} = \sum_{i=1}^{N}\left(\frac{n_i - np_i(\boldsymbol{\theta})}{\sqrt{np_i(\boldsymbol{\theta})}}\right)^2. \end{aligned} \tag{9.25}$$

Minimizing (9.25) is not a trivial exercise, as the parameters $\{\theta_1, \theta_2, \ldots, \theta_p\}$ enter non-linearly. However, if the number of bins N is large enough, individual p_i are so small the all off-diagonal elements of the matrix Σ may be neglected. Consequently, $(\Sigma)_{ii} = \sigma_i^2 = np_i(1 - p_i) \approx np_i = f_i$ and the parameters $\boldsymbol{\theta}$ can be obtained by solving the system of equations

$$\frac{\partial X^2}{\partial \theta_j} = -2\sum_{i=1}^{N}\left(\frac{n_i - f_i}{f_i} + \frac{1}{2}\left(\frac{n_i - f_i}{f_i}\right)^2\right)\frac{\partial f_i}{\partial \theta_j} = 0, \qquad j = 1, 2, \ldots, p. \tag{9.26}$$

Two simplifications are possible—none of which eliminates non-linearity. The first option is at hand: imagine that the denominators in (9.25) are independent of $\boldsymbol{\theta}$. This is equivalent to the system (9.26) without the quadratic term. Besides, we can replace f_i by n_i in the denominators, as these values can not be *that* different! In this simplified approach, one only needs to solve the system

$$\frac{\partial X^2}{\partial \theta_j} = -2\sum_{i=1}^{N}\left(\frac{n_i - f_i}{f_i}\right)\frac{\partial f_i}{\partial \theta_j} \approx -2\sum_{i=1}^{N}\left(1 - \frac{f_i}{n_i}\right)\frac{\partial f_i}{\partial \theta_j} = 0. \tag{9.27}$$

Example Let us revisit Fig. 9.5 (left). A total of $n = 1000$ counts have been classified into $N = 10$ equidistant bins $[x_i, x_i + \Delta x]$ of width $\Delta x = 0.1$, where $x_i = (i - 1)\Delta x$, $i = 1, 2, \ldots, N$. There are $n_1, n_2, \ldots, n_N = 113$, 147, 153, 136, 95, 74, 54, 59, 79, 90 counts in individual bins. Assume that the observations are described by the probability density

$$f(x; \theta) = 1 + \theta\,\sin(2\pi x), \qquad 0 \le x \le 1, \tag{9.28}$$

where θ is an unknown parameter, and that the uncertainties of n_i are Poissonian. The corresponding f_i can be calculated by integrating f over each bin:

$$f_i = n \int_{x_i}^{x_i+\Delta x} f(x;\theta)\,\mathrm{d}x = n\Big[\underbrace{\Delta x}_{a_i} + \underbrace{(2\pi)^{-1}[\cos(2\pi x_i) - \cos(2\pi(x_i+\Delta x))]}_{b_i}\,\theta\Big].$$

The estimate for θ is obtained by (9.27), where we use $f_i = n(a_i + b_i\theta)$:

$$\frac{\partial X^2}{\partial\theta} = -2\sum_{i=1}^{N} \frac{\big(n_i - n(a_i + b_i\theta)\big)}{n_i}\, n b_i = 0,$$

We obtain

$$\widehat{\theta} = \frac{1}{n}\left[\sum_{i=1}^{N}\frac{b_i^2}{n_i}\right]^{-1} \sum_{i=1}^{N} b_i\left(1 - \frac{na_i}{n_i}\right) \approx 0.457.$$

We are dealing with a linear problem, so X^2 near $\widehat{\theta}$ has a parabolic shape (9.10), which we write as $X^2\big(\theta\big) = X^2\big(\widehat{\theta}\big) + \big(\sum_{i=1}^{N} n^2 b_i^2/n_i\big)\big(\theta - \widehat{\theta}\big)^2$, whence

$$\mathrm{var}\big[\widehat{\theta}\big] = 2\left(\frac{\partial^2 X^2}{\partial\theta^2}\right)^{-1} = \frac{1}{n^2}\left(\sum_{i=1}^{N}\frac{b_i^2}{n_i}\right)^{-1} \approx 0.0018 \quad\Longrightarrow\quad \sqrt{\mathrm{var}\big[\widehat{\theta}\big]} \approx 0.042.$$

Being a density, the best-fit function (9.28) now only needs to be normalized. As we have $n = 1000$ counts in $N = 10$ bins, it obviously has to be multiplied by $n/N = 100$, so the final form is $f(x) = 100\,\big(1 + 0.457\sin(2\pi x)\big)$. It is shown by the dashed line in Fig. 9.5 (left). ◁

9.3 Linear Regression with Constraints

The *true* quantities being measured—call them η_i—are often algebraically related through some kind of *constraints*. The actual observations y_i have uncertainties σ_i, so they do not necessarily satisfy the constraints which, however, should be satisfied by the *estimates* $\widehat{\eta}_i$. The following classic example outlines two general approaches to including constraints in the method of least squares.

Example A measurement of the interior angles of a triangle yields $y_1 = 51°, y_2 = 42°$ and $y_3 = 85°$ with errors $\sigma_1 = \sigma_2 = \sigma_3 \equiv \sigma = 1°$. The measured values add up to $y_1 + y_2 + y_3 = 178°$, not the required 180°. If the true angles are considered as unknown parameters ($\eta_i = \theta_i$), they can be estimated by the method of least squares. We minimize

$$X^2(\boldsymbol{\theta}) = \sum_{i=1}^{3}\left(\frac{y_i - \theta_i}{\sigma_i}\right)^2,$$

besides, we require that

$$\theta_1 + \theta_2 + \theta_3 - 180^\circ = 0. \tag{9.29}$$

This constraint can be used to eliminate one variable from X^2, say, θ_3:

$$X^2(\boldsymbol{\theta}) = \left(\frac{y_1 - \theta_1}{\sigma_1}\right)^2 + \left(\frac{y_2 - \theta_2}{\sigma_2}\right)^2 + \left(\frac{y_3 - (180^\circ - \theta_1 - \theta_2)}{\sigma_3}\right)^2.$$

We calculate the derivative of this measure with respect to θ_1 and θ_2 and set it to zero, obtaining two equations for two unknowns. With their solution we exploit (9.29) to obtain the remaining third angle, thus finally

$$\widehat{\theta_1} = 51\tfrac{2}{3}^\circ, \quad \widehat{\theta_2} = 42\tfrac{2}{3}^\circ, \quad \widehat{\theta_3} = 180^\circ - \widehat{\theta_1} - \widehat{\theta_2} = 85\tfrac{2}{3}^\circ,$$

so that the requirement $\widehat{\theta_1} + \widehat{\theta_2} + \widehat{\theta_3} = 180^\circ$ is fulfilled by construction. It can be seen that the method has uniformly distributed the missing 2° from 178° to the correct value 180° among the three observations.

The problem can also be solved by using Lagrange multipliers. The constraint equation (9.29) is taken into account by endowing it with a weight representing a new unknown parameter with respect to which the measure of deviation needs to be minimized:

$$X^2(\theta_1, \theta_2, \theta_3, \lambda) = \sum_{i=1}^{3} \left(\frac{y_i - \theta_i}{\sigma_i}\right)^2 + 2\lambda\left(\sum_{i=1}^{3} \theta_i - 180^\circ\right).$$

By solving the equations $\partial X^2/\partial\theta_1 = \partial X^2/\partial\theta_2 = \partial X^2/\partial\theta_3 = \partial X^2/\partial\lambda = 0$ we get

$$\widehat{\lambda} = \frac{1}{3\sigma^2}\left(\sum_{i=1}^{3} y_i - 180^\circ\right), \qquad \widehat{\theta_i} = y_i - \sigma^2\widehat{\lambda} = y_i - \frac{1}{3}\left(\sum_{i=1}^{3} y_i - 180^\circ\right), \tag{9.30}$$

whence $\widehat{\lambda} = -\frac{2}{3}$, yielding the same $\widehat{\theta_1}$, $\widehat{\theta_2}$ and $\widehat{\theta_3}$ as before. ◁

It is worthwhile to outline the method of least squares with linear constraints in a more general way. If we must determine p parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_p\}$ satisfying q constraints with the appropriate multipliers $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$, we must minimize

$$X^2(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \left(\boldsymbol{y} - A\boldsymbol{\theta}\right)^{\mathrm{T}} \Sigma_{\boldsymbol{y}}^{-1} \left(\boldsymbol{y} - A\boldsymbol{\theta}\right) + 2\boldsymbol{\lambda}^{\mathrm{T}}\left(B\boldsymbol{\theta} - \boldsymbol{b}\right).$$

Here B is a $q \times p$ matrix and $\boldsymbol{b}$ is a q-dimensional vector, while the observations vector $\boldsymbol{y}$, their covariance matrix $\Sigma_{\boldsymbol{y}}$ and the regression matrix A have their usual, well-known roles. The measure X^2 is minimized by setting its derivatives with respect to each of the $p + q$ parameters to zero:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} X^2 &= -2\bigl(A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}\boldsymbol{y} - A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A\boldsymbol{\theta}\bigr) + 2B^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{0},\\ \nabla_{\boldsymbol{\lambda}} X^2 &= \quad 2\bigl(B\boldsymbol{\theta} - \boldsymbol{b}\bigr) = \mathbf{0}.\end{aligned}$$

By denoting

$$\Sigma_{\boldsymbol{c}}^{-1} = A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}A, \quad \boldsymbol{c} = A^{\mathrm{T}}\Sigma_{\boldsymbol{y}}^{-1}\boldsymbol{y}, \quad \Sigma_{\boldsymbol{b}} = B\Sigma_{\boldsymbol{c}}B^{\mathrm{T}}, \tag{9.31}$$

this can be written as

$$\begin{aligned}\Sigma_{\boldsymbol{c}}^{-1}\boldsymbol{\theta} + B^{\mathrm{T}}\boldsymbol{\lambda} &= \boldsymbol{c},\\ B\boldsymbol{\theta} &= \boldsymbol{b}.\end{aligned}$$

The solution of this system [1] are $\boldsymbol{\lambda}$, $\boldsymbol{\theta}$, and their variances and covariances:

$$\widehat{\boldsymbol{\lambda}} = \Sigma_{\boldsymbol{b}}^{-1}\bigl(B\Sigma_{\boldsymbol{c}}\boldsymbol{c} - \boldsymbol{b}\bigr), \tag{9.32}$$

$$\widehat{\boldsymbol{\theta}} = \Sigma_{\boldsymbol{c}}\boldsymbol{c} - \Sigma_{\boldsymbol{c}}B^{\mathrm{T}}\Sigma_{\boldsymbol{b}}^{-1}\bigl(B\Sigma_{\boldsymbol{c}}\boldsymbol{c} - \boldsymbol{b}\bigr), \tag{9.33}$$

$$\Sigma_{\widehat{\boldsymbol{\theta}}} = \Sigma_{\boldsymbol{c}} - \bigl(B\Sigma_{\boldsymbol{c}}\bigr)^{\mathrm{T}}\Sigma_{\boldsymbol{b}}^{-1}\bigl(B\Sigma_{\boldsymbol{c}}\bigr). \tag{9.34}$$

Exercise Let us revisit the triangle angles problem. The observations vector is $\boldsymbol{y} = (y_1, y_2, y_3)^{\mathrm{T}}$, their covariance matrix is $\Sigma_{\boldsymbol{y}} = \mathrm{diag}\bigl(\sigma^2, \sigma^2, \sigma^2\bigr)$, and the regression matrix is $A = \mathrm{diag}(1, 1, 1)$. The constraint equation $B\boldsymbol{\theta} = \boldsymbol{b}$ is embodied by the 1×3 "matrix" B and the 1-dimensional "vector" $\boldsymbol{b}$:

$$B = \bigl(1\;1\;1\bigr), \qquad \boldsymbol{b} = \bigl(180^\circ\bigr).$$

The quantities from the definition (9.31) are

$$\Sigma_{\boldsymbol{c}}^{-1} = \frac{1}{\sigma^2}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} = \Sigma_{\boldsymbol{y}}^{-1}, \qquad \boldsymbol{c} = \frac{1}{\sigma^2}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix} = \frac{1}{\sigma^2}\boldsymbol{y}, \qquad \Sigma_{\boldsymbol{b}} = \bigl(3\sigma^2\bigr).$$

Using (9.32) and (9.33) immediately leads to (9.30), and formula (9.34) gives the variances and covariances of the parameter estimates:

$$\Sigma_{\widehat{\boldsymbol{\theta}}} = \begin{pmatrix}2/3 & -1/3 & -1/3\\ -1/3 & 2/3 & -1/3\\ -1/3 & -1/3 & 2/3\end{pmatrix}(^\circ)^2.$$

We see that the estimated "optimal" values of the angles have a smaller effective deviation than the measured ones—namely$(\sqrt{2/3})^\circ$ instead of 1°—but they have become correlated (non-zero off-diagonal elements). ◁

9.4 General Linear Regression by Singular-Value Decomposition ⋆

A more general form of linear regression,

$$f(x) = \sum_{j=1}^{p} \theta_j \phi_j(x),$$

where ϕ_j are basis functions, can be used to fit the model function f to a large data set with fewer parameters. Such problems are over-determined, but the data are often not rich enough to nail down a unique linear combination of the basis functions: one may obtain many functions that minimize the measure

$$X^2 = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i; \boldsymbol{\theta})}{\sigma_i} \right)^2 \tag{9.35}$$

almost equally well. (From a strict mathematical point of view, of course, the least-square solution of an over-determined problem is unique.)

A better control over the importance of the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{T}}$ can be obtained by using singular-value decomposition (see Sects. 3.3.2 and 3.3.3 in [3]), which excels also in terms of numerical stability. By denoting

$$A_{ij} = \frac{\phi_j(x_i)}{\sigma_i}, \qquad b_i = \frac{y_i}{\sigma_i},$$

where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, we see that (9.35) has the typical least-squares form $X^2 = \|A\boldsymbol{\theta} - \boldsymbol{b}\|^2$. We then perform the singular-value decomposition of A: $A = U \Lambda V^{\mathrm{T}} \in \mathbb{R}^{n \times p}$, where $n \geq p$. The matrix $U = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p)$ has n-dimensional columns $\boldsymbol{u}_i$, the matrix $V = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p)$ has p-dimensional columns $\boldsymbol{v}_i$, and the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ contains the singular values λ_i. The vector of optimal parameters is obtained by the sum

$$\widehat{\boldsymbol{\theta}} = \sum_{i=1}^{p} \frac{\boldsymbol{u}_i^{\mathrm{T}} \boldsymbol{b}}{\lambda_i} \boldsymbol{v}_i.$$

The variances and covariances of $\widehat{\boldsymbol{\theta}}$ are

$$\mathrm{var}\big[\widehat{\theta}_j\big] = \sum_{i=1}^{p} \frac{V_{ji}^2}{\lambda_i^2}, \qquad \mathrm{cov}\big[\widehat{\theta}_j, \widehat{\theta}_k\big] = \sum_{i=1}^{p} \frac{V_{ji} V_{ki}}{\lambda_i^2}. \tag{9.36}$$

We must be alert to the singular values λ_i for which $\lambda_i / \lambda_{\max} \lesssim n\varepsilon_{\mathrm{M}}$, where ε_{M} is the machine precision. Such values, appearing in the denominators of (9.36), increase the

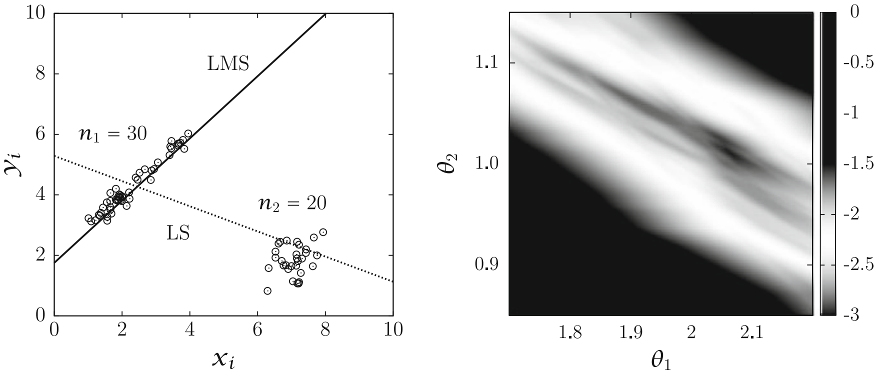


Fig. 9.6 Robust linear regression on data containing a significant fraction of outliers. [Left] The standard least-squares method (LS) yields a straight line that goes through both data clusters, but does not describe their main part correctly. The LMS method delivers a reasonable description. [Right] The function being minimized in the LMS method.

uncertainties of parameters θ_j and indicate that including further parameters would be pointless. They also contribute insignificantly to the minimization of X^2, so they can be eliminated. This is done by setting $1/\lambda_i = 0$ (for further explanations see e.g. the comment to the `Fitsvd` algorithm in [6]). One may also discard those singular values for which the ratio $\lambda_i/\lambda_{\max}$ is *larger* than $\approx n\varepsilon_{\mathrm{M}}$, at least until X^2 starts to increase significantly.

9.5 Robust Linear Regression

As all estimates of distribution location and scale, regression methods are sensitive to outliers (see Sect. 7.4). A telling example is shown in Fig. 9.6 (left). A set of $n_1 = 30$ data of the form $y_i = \theta_1 + \theta_2 x_i + z_i$, where $\theta_1 = 2$, $\theta_2 = 1$, x_i and z_i are realizations of $X \sim U(1, 4)$ and $Z \sim N(0, 0.04)$, and a relatively large set of $n_2 = 20$ presumed outliers (x_i, y_i), realizations of $X \sim N(7, 0.25)$ and $Y \sim N(2, 0.25)$, are fitted by a straight line. The standard least-squares method (LS) yields a result that does not describe the bulk of the data.

A simple method exists where we minimize the median of the squares of residuals $y_i - (\theta_1 + \theta_2 x_i)$ called *least median of squares* (LMS). We seek parameters θ_1 and θ_2 that minimize the measure of deviation

$$\mathrm{med}_i \left[\left(y_i - \theta_1 - \theta_2 x_i \right)^2 \right]. \tag{9.37}$$

The main issue with the LMS method is precisely the minimization of (9.37). The function being minimized with respect to θ_j has $\mathcal{O}(n^{p+1})$ local minima, where

n is the number of points (x_i, y_i) and p is the degree of the regression polynomial. The example in the figure has $n = n_1 + n_2 = 50$ and $p = 1$ (straight line), so there are $\approx$2500 local minima, among which the global one needs to be found, as shown in Fig. 9.6 (right). For introductory reading on robust regression see [7].

9.6 Non-linear Regression

In non-linear regression the dependence of the model function on regression parameters is non-linear, for example

$$f(x; \boldsymbol{\theta}) = \theta_1 + \theta_2\, \mathrm{e}^{\theta_3 x} + \theta_4 \sin(x + \theta_5).$$

(Compare this to (9.4)!) As usual, the observations y_i at x_i are arranged in vectors $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, and the components of the regression function in $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}) = \big(f(x_1; \boldsymbol{\theta}), f(x_2; \boldsymbol{\theta}), \ldots, f(x_n; \boldsymbol{\theta})\big)^{\mathrm{T}}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{T}}$. By analogy to (9.5) the measure of deviation is defined as

$$X^2 = \big(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})\big)^{\mathrm{T}} \Sigma_{\boldsymbol{y}}^{-1} \big(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})\big).$$

If the measurement errors are uncorrelated, the covariance matrix is diagonal, $\Sigma_{\boldsymbol{y}} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$, and the above expression can be simplified to

$$X^2 = \sum_{i=1}^{n} \frac{\big(y_i - f(x_i; \boldsymbol{\theta})\big)^2}{\sigma_i^2}.$$

This is where the problems start. Minimization of X^2 now implies solving a system of p (in general non-linear) equations $\partial X^2/\partial\theta_j = 0$ $(j = 1, 2, \ldots, p)$. Such problems are therefore solved iteratively: we ride on the sequence

$$\boldsymbol{\theta}^{(\nu+1)} = \boldsymbol{\theta}^{(\nu)} + \Delta\boldsymbol{\theta}^{(\nu)}, \qquad \nu = 0, 1, 2, \ldots, \tag{9.38}$$

where $\Delta\boldsymbol{\theta}^{(\nu)}$ is obtained by solving a *linear* problem, to approach the optimal parameter set. In the νth step the update can be calculated by minimizing

$$X^2(\Delta\boldsymbol{\theta}) = \Big[\boldsymbol{y} - \boldsymbol{f}\Big(\boldsymbol{\theta}^{(\nu)} + \Delta\boldsymbol{\theta}\Big)\Big]^{\mathrm{T}} \Sigma_{\boldsymbol{y}}^{-1} \Big[\boldsymbol{y} - \boldsymbol{f}\Big(\boldsymbol{\theta}^{(\nu)} + \Delta\boldsymbol{\theta}\Big)\Big],$$

where we have suppressed the dependence of $\boldsymbol{f}$ on $\boldsymbol{x}$. If $\Delta\boldsymbol{\theta}$ is small, $\boldsymbol{f}$ can be expanded as $\boldsymbol{f}\big(\boldsymbol{\theta}^{(\nu)} + \Delta\boldsymbol{\theta}\big) \approx \boldsymbol{f}\big(\boldsymbol{\theta}^{(\nu)}\big) + J\big(\boldsymbol{\theta}^{(\nu)}\big)\, \Delta\boldsymbol{\theta}$, where J is the Jacobi matrix with the elements

$$J_{ij} = \left(\frac{\partial f_i}{\partial \theta_j}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(\nu)}}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p.$$

Let us denote $\boldsymbol{f}_\nu = \boldsymbol{f}\big(\boldsymbol{\theta}^{(\nu)}\big)$ and $J_\nu = J\big(\boldsymbol{\theta}^{(\nu)}\big)$. We must minimize the expression

$$X^2(\Delta\boldsymbol{\theta}) = \Big[\big(\boldsymbol{y} - \boldsymbol{f}_\nu\big) - J_\nu \Delta\boldsymbol{\theta}\Big]^{\mathrm{T}} \Sigma_{\boldsymbol{y}}^{-1} \Big[\big(\boldsymbol{y} - \boldsymbol{f}_\nu\big) - J_\nu \Delta\boldsymbol{\theta}\Big].$$

From the requirement $\partial X^2/\partial(\Delta\boldsymbol{\theta}) = \mathbf{0}$ it follows that

$$\big[J_\nu^{\mathrm{T}} \Sigma_{\boldsymbol{y}}^{-1} J_\nu\big]\, \Delta\boldsymbol{\theta}^{(\nu)} = J_\nu^{\mathrm{T}} \Sigma_{\boldsymbol{y}}^{-1} \big(\boldsymbol{y} - \boldsymbol{f}_\nu\big).$$

This system of *linear* equations must be solved in order to obtain the update $\Delta\boldsymbol{\theta}$ in the νth iteration (9.38). If the data uncertainties are uncorrelated, the procedure can be summarized as

$$G\big(\boldsymbol{\theta}^{(\nu)}\big)\, \Delta\boldsymbol{\theta}^{(\nu)} = \boldsymbol{g}\big(\boldsymbol{\theta}^{(\nu)}\big), \tag{9.39}$$

$$\boldsymbol{\theta}^{(\nu+1)} = \boldsymbol{\theta}^{(\nu)} + \Delta\boldsymbol{\theta}^{(\nu)}, \qquad \nu = 0, 1, 2, \ldots, \tag{9.40}$$

where

$$G_{jk}\big(\boldsymbol{\theta}^{(\nu)}\big) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(\frac{\partial f_i}{\partial \theta_j} \frac{\partial f_i}{\partial \theta_k}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(\nu)}}, \qquad g_j\big(\boldsymbol{\theta}^{(\nu)}\big) = \sum_{i=1}^{n} \left(\frac{y_i - f_i}{\sigma_i^2} \left(\frac{\partial f_i}{\partial \theta_j}\right)\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(\nu)}}.$$

We are still missing the uncertainties of $\widehat{\boldsymbol{\theta}}$. Near the optimum X^2 is approximately parabolic, so the covariance matrix of the regression parameters can be computed as in (9.8),

$$\Sigma_{\widehat{\boldsymbol{\theta}}} = \big[J^{\mathrm{T}}\big(\widehat{\boldsymbol{\theta}}\big) \Sigma_{\boldsymbol{y}}^{-1} J\big(\widehat{\boldsymbol{\theta}}\big)\big]^{-1},$$

or simply $\Sigma_{\widehat{\boldsymbol{\theta}}} = \big[G\big(\widehat{\boldsymbol{\theta}}\big)\big]^{-1}$, if σ_i are uncorrelated. As usual the diagonal elements of this matrix are the parameter variances, and the off-diagonal are the covariances, see (9.9). If the uncertainties of $\boldsymbol{y}$ are unknown, the parameter covariance matrix can be calculated by analogy to (9.18),

$$\Sigma_{\widehat{\boldsymbol{\theta}}} = \frac{\mathrm{SSR}}{n-p} \big[J^{\mathrm{T}}\big(\widehat{\boldsymbol{\theta}}\big) J\big(\widehat{\boldsymbol{\theta}}\big)\big]^{-1}, \qquad \mathrm{SSR} = X^2\big(\widehat{\boldsymbol{\theta}}\big). \tag{9.41}$$

Example The circles in Fig. 9.7 (left) show $n = 132$ values of annual global release of CO_2 from fossil fuels over 1880–2011, measured in millions of tons [8]. We would like to model the data by a two-parameter ($p = 2$) function of the form

$$f(x) = \theta_1\, \mathrm{e}^{\theta_2 x}, \tag{9.42}$$

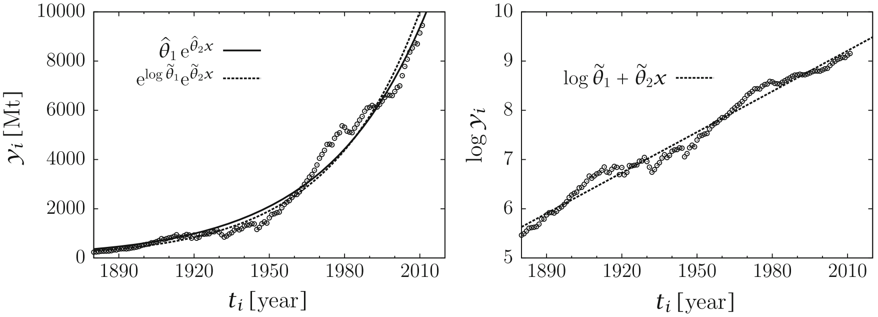


Fig. 9.7 Global release of CO$_2$. [Left] Exponential model, calculated by non-linear regression (*full curve*), and the curve obtained by transforming the straight-line regression of the logarithmic data (*dashed*). [Right] Linear regression of logarithmic data

where $x = t - 1880$ (origin shift as in Fig. 9.2). The data errors are unknown, which amounts to $\sigma_i = 1$ for all i. The iterative method requires the derivatives

$$\frac{\partial f_i}{\partial\theta_1} = \mathrm{e}^{\theta_2 x_i}, \qquad \frac{\partial f_i}{\partial\theta_2} = \theta_1 x_i\,\mathrm{e}^{\theta_2 x_i},$$

to form the 2×2 matrix G,

$$G_{11}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\mathrm{e}^{2\theta_2 x_i}, \qquad G_{12}(\boldsymbol{\theta}) = \theta_1\sum_{i=1}^{n}x_i\,\mathrm{e}^{2\theta_2 x_i}, \qquad G_{22}(\boldsymbol{\theta}) = \theta_1^2\sum_{i=1}^{n}x_i^2\,\mathrm{e}^{2\theta_2 x_i},$$

where $G_{21} = G_{12}$, as well as the right-hand side of (9.39):

$$g_1(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left(y_i - \theta_1\mathrm{e}^{\theta_2 x_i}\right)\mathrm{e}^{\theta_2 x_i}, \qquad g_2(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left(y_i - \theta_1\mathrm{e}^{\theta_2 x_i}\right)\theta_1 x_i\,\mathrm{e}^{\theta_2 x_i}.$$

Iteration (9.40) is started by the initial approximation $(\theta_1, \theta_2)^{(0)} = (200, 0.02)$. After less than ten iterations we obtain the final values

$$\widehat{\theta}_1 = (363.8 \pm 18.3)\mathrm{Mt}, \qquad \widehat{\theta}_2 = (2.499 \pm 0.045) \times 10^{-2}/\mathrm{yr},$$

where the uncertainties have been computed by formula (9.41). The result of non-linear regression is shown by the full curve in Fig. 9.7 (left). However, do think about it: a standard polynomial regression of such strongly scattered data would yield an equally likable result.

There is another way to proceed. If we take the logarithm of the values y_i, non-linear regression becomes linear, since

$$\log f(x) = \log\theta_1 + \theta_2 x. \tag{9.43}$$

Now the regression parameters are $\log\theta_1$ (not θ_1) and θ_2. By the recipe of Sect. 9.1.6 we compute their optimal values

$$\log\widetilde{\theta}_1 = (5.630 \pm 0.026)\log(\mathrm{Mt}), \qquad \widetilde{\theta}_2 = (2.755 \pm 0.034) \times 10^{-2}/\mathrm{yr},$$

corresponding to the straight line in Fig. 9.7 (right). But if the obtained $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ are reinserted in (9.42), one gets a different description of the non-logarithmic data, namely the dashed curve in Fig. 9.7 (left). The reason for this disagreement is clear. Even though the models (9.42) and (9.43) are mathematically identical, the chosen statistical models $y_i = \theta_1\,\mathrm{e}^{\theta_2 x_i} + \sigma_i$ and $\log y_i = \log\theta_1 + \theta_2 x_i + \varepsilon_i$ are *not* equivalent because the uncertainties σ_i and ε_i in each case have different distributions. Besides, the method of least squares by itself is not "invariant" with respect to such transformations. ◁

9.7 Problems

The following Problems refer to standard data sets of NIST, suitable for studying non-linear regression and testing computer programs [9]. The data files can be found on the website of the book.

9.7.1 Two Gaussians on Exponential Background

A frequent problem in all areas of physics is describing the data by a model involving several normal (Gaussian) distributions with different means and widths, superposed on exponential background. An example is shown in Fig. 9.8 (left). Fit the data

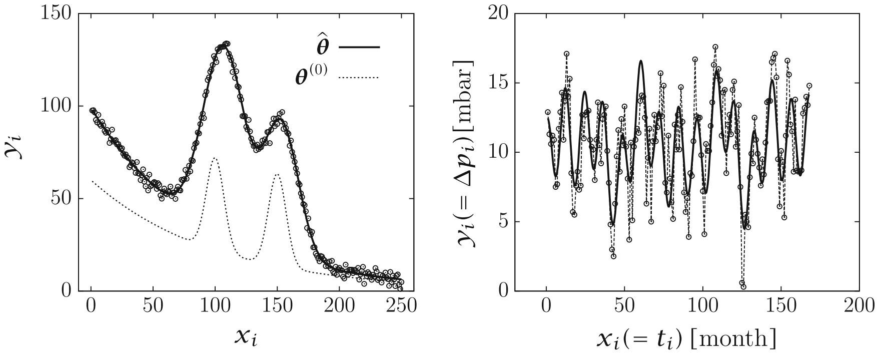


Fig. 9.8 [Left] Two Gaussians on exponential background (model function (9.44)). The *dashed line* is the initial approximation with parameters (9.45), and the *full curve* is the final result of the fit. [Right] Time dependence of the pressure gradient during the "El Niño Southern Oscillation" (ENSO) phenomenon

$\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ at $n = 250$ points $\{x_1, x_2, \ldots, x_n\} = \{1, 2, \ldots, n\}$ by a eight-parameter ($p = 8$) function

$$f(x; \boldsymbol{\theta}) = \theta_1 \mathrm{e}^{-\theta_2 x} + \theta_3 \exp\left(-\frac{(x-\theta_4)^2}{\theta_5^2}\right) + \theta_6 \exp\left(-\frac{(x-\theta_7)^2}{\theta_8^2}\right) \tag{9.44}$$

by using the method of non-linear regression described in Sect. 9.6! Use the initial approximation

$$\boldsymbol{\theta}^{(0)} = \{\theta_1, \theta_2, \ldots, \theta_8\} = \{60, 0.01, 50, 100, 10, 50, 150, 10\}, \tag{9.45}$$

corresponding to the dashed line.

✎ The final (rounded) set of parameters and their uncertainties is

$$\begin{aligned}
&\widehat{\theta}_1 = 99.02 \pm 0.54,\ \widehat{\theta}_2 = 0.011 \pm 0.0001,\ \widehat{\theta}_3 = 101.88 \pm 0.59,\ \widehat{\theta}_4 = 107.03 \pm 0.15,\\
&\widehat{\theta}_5 = 23.58 \pm 0.23,\ \widehat{\theta}_6 = 72.05 \pm 0.62,\quad \widehat{\theta}_7 = 153.27 \pm 0.19,\ \widehat{\theta}_8 = 19.53 \pm 0.26.
\end{aligned}$$

It corresponds to the full curve in Fig. 9.8 (left).

9.7.2 Time Dependence of the Pressure Gradient

Figure 9.8 (right) shows $n = 168$ monthly averages of pressure differences between Easter Island in the Pacific and the Australian city of Darwin [10]. This pressure gradient is responsible for the trade winds in the southern hemisphere. The Fourier analysis of the data (the signal is indicated by the dashed line connecting the points) reveals peaks at three frequencies. The most prominent one corresponds to the annual cycle (12-month period), but one can detect two further components with 26 and 44-month periods, characteristic for the so-called *El Niño Southern Oscillation* (ENSO) phenomenon. Fit the data by a nine-parameter ($p = 9$) function

$$\begin{aligned}
f(x; \boldsymbol{\theta}) = \theta_1 &+ \theta_2 \cos\bigl(2\pi x/12\bigr) + \theta_3 \cos\bigl(2\pi x/12\bigr)\\
&+ \theta_5 \cos\bigl(2\pi x/\theta_4\bigr) + \theta_6 \cos\bigl(2\pi x/\theta_4\bigr)\\
&+ \theta_8 \cos\bigl(2\pi x/\theta_7\bigr) + \theta_9 \cos\bigl(2\pi x/\theta_7\bigr)
\end{aligned}$$

by the method of non-linear regression (Sect. 9.6) with the initial approximation

$$\boldsymbol{\theta}^{(0)} = \{\theta_1, \theta_2, \ldots, \theta_9\} = \{11, 3, 0.5, 40, -0.7, -1.3, 25, -0.3, 1.4\}.$$

✎ The final set of parameters and their uncertainties is

$$\begin{aligned}
&\widehat{\theta}_1 = 10.51 \pm 0.17, && \widehat{\theta}_2 = 3.076 \pm 0.243, && \widehat{\theta}_3 = 0.533 \pm 0.244,\\
&\widehat{\theta}_4 = 44.31 \pm 0.94, && \widehat{\theta}_5 = -1.623 \pm 0.281, && \widehat{\theta}_6 = 0.526 \pm 0.481,\\
&\widehat{\theta}_7 = 26.89 \pm 0.42, && \widehat{\theta}_8 = 0.213 \pm 0.515, && \widehat{\theta}_9 = 1.497 \pm 0.254.
\end{aligned}$$

It corresponds to the full curve in Fig. 9.8 (right).

9.7.3 *Thermal Expansion of Copper*

Figure 9.9 (left) shows the measured thermal expansion coefficient of copper as a function of temperature [11]. Describe the data ($n = 236$) by the rational function

$$f(x;\boldsymbol{\theta}) = \frac{\theta_1 + \theta_2 x + \theta_3 x^2 + \theta_4 x^3}{1 + \theta_5 x + \theta_6 x^2 + \theta_7 x^3}, \tag{9.46}$$

depending on $p = 7$ parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_7\}$. Use the initial sets

$$\boldsymbol{\theta}^{(0)} = \{10, -1, 0.05, -10^{-5}, -0.05, 0.001, -10^{-6}\}, \tag{9.47}$$

$$\boldsymbol{\theta}^{(0)} = \{10, -0.1, 0.005, -10^{-6}, -0.005, 10^{-4}, -10^{-7}\}, \tag{9.48}$$

and make precisely 20 steps of (9.39). Does the iteration converge in both cases? Plot the solutions corresponding to both parameter sets.

✎ Choosing good initial parameters is crucial. The iteration started with (9.47) does not converge: the solution after 20 iterations is the dashed curve in the figure. The iteration initialized with (9.48) does converge (full curve); the final parameter set is

$$\begin{aligned}
&\widehat{\theta}_1 = 1.08 \pm 0.17,\\
&\widehat{\theta}_2 = -0.123 \pm 0.012, && \widehat{\theta}_3 = (4.08 \pm 0.23) \times 10^{-3}, && \widehat{\theta}_4 = (-1.43 \pm 0.28) \times 10^{-6},\\
&\widehat{\theta}_5 = (-5.76 \pm 0.25) \times 10^{-3}, && \widehat{\theta}_6 = (2.40 \pm 0.10) \times 10^{-4}, && \widehat{\theta}_7 = (-1.23 \pm 0.13) \times 10^{-7}.
\end{aligned}$$

9.7.4 *Electron Mobility in Semiconductor*

Figure 9.9 (right) shows the measured electron mobilities in silicon as a function of the (log)-concentration of donor admixtures at a certain temperature [12]. Fit the data ($n = 37$) by a model of the form (9.46)! Make 20 steps of the iteration (9.39) with two initial regression parameter sets:

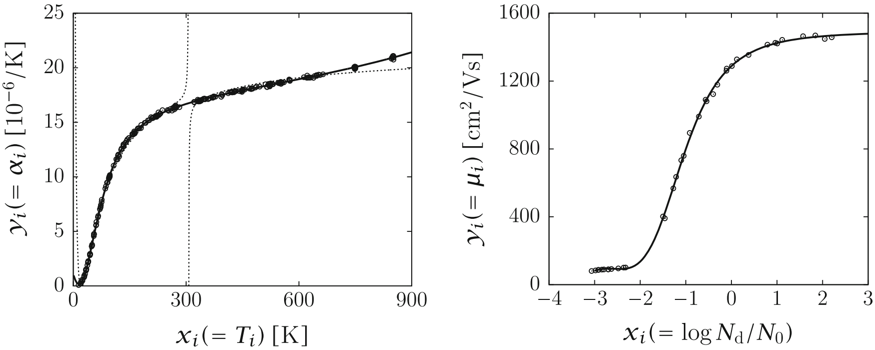


Fig. 9.9 [Left] Fitting the function (9.46) to the measured thermal expansion coefficient of copper after 20 iterations: the *dashed curve* corresponds to the initial conditions (9.47), while the *full curve* corresponds to (9.48). [Right] Regression analysis of the data on electron mobility in silicon as a function of donor concentration

$$\boldsymbol{\theta}^{(0)} = \{1000, 1000, 400, 40, 0.7, 0.3, 0.03\},$$
$$\boldsymbol{\theta}^{(0)} = \{1300, 1500, 500, 75, 1.0, 0.4, 0.05\}.$$

✎ The method started with the first initial set does not converge. Using the second set does lead to convergence and the final parameters are

$$\begin{aligned}&\widehat{\theta_1} = 1288.1 \pm 4.7,\\ &\widehat{\theta_2} = 1491.1 \pm 39.6,\ \widehat{\theta_3} = 583.2 \pm 28.7,\quad \widehat{\theta_4} = 75.4 \pm 5.6,\\ &\widehat{\theta_5} = 0.966 \pm 0.03,\quad \widehat{\theta_6} = 0.398 \pm 0.015,\ \widehat{\theta_7} = 0.0497 \pm 0.0066.\end{aligned}$$

The solution with these parameters is shown by the full curve in the figure.

9.7.5 Quantum Defects in Iodine Atoms

Figure 9.10 (left) shows the data from a study of quantum defects in iodine atoms [9], with the excited-state energies on the abscissa and the number of defects on the ordinate axis. Fit the $n = 25$ values by a four-parameter ($p = 4$) function

$$f(x; \boldsymbol{\theta}) = \theta_1 - \theta_2 x - \frac{1}{\pi}\arctan\left(\frac{\theta_3}{x - \theta_4}\right) \tag{9.49}$$

by using the method of non-linear least squares, described in Sect. 9.6! Initialize the algorithm by the regression parameters

$$\boldsymbol{\theta}^{(0)} = \{\theta_1, \theta_2, \theta_3, \theta_4\} = \{0.2, -5 \cdot 10^{-6}, 1200, -150\}.$$

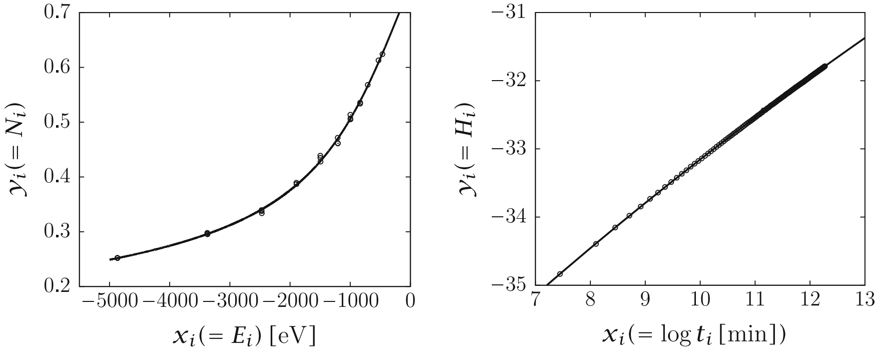


Fig. 9.10 [Left] Modeling the number of quantum defects in iodine atoms by the function (9.49). [Right] Fitting the function (9.50) to the observed values of the magnetic field strength in a superconductor as a function of time

✎ The final values of the parameters are

$$\widehat{\theta}_1 = 0.202 \pm 0.019,\ \widehat{\theta}_2 = (-6.2 \pm 3.2) \cdot 10^{-6},\ \widehat{\theta}_3 = 1204 \pm 74,\ \widehat{\theta}_4 = -181.3 \pm 49.6.$$

9.7.6 Magnetization in Superconductor

The circles in Fig. 9.10 (right) represent the $n = 154$ observed magnetic field strengths as a function of time from a study of magnetization in superconductors [9]. Model the data by the function

$$f(x; \boldsymbol{\theta}) = \theta_1 \big(\theta_2 + x\big)^{-1/\theta_3}. \tag{9.50}$$

Use the iterative method of non-linear least squares described in Sect. 9.6. The initial parameter set is $\boldsymbol{\theta}^{(0)} = \{\theta_1, \theta_2, \theta_3\} = \{-1500, 45, 0.85\}$.

✎ The final parameter set when iteration (9.39) terminates is $\widehat{\theta}_1 = -2523.5 \pm 297.2$, $\widehat{\theta}_2 = 46.74 \pm 1.24$ and $\widehat{\theta}_3 = 0.932 \pm 0.020$.

References

1. A.G. Frodesen, O. Skjeggestad, H. Tøfte, *Probability and Statistics in Particle Physics* (Universitetsforlaget, Bergen, 1979)
2. GISS/NASA Analysis, based on: J. Hansen, R. Ruedy, M. Sato, K. Lo, *Global surface temperature change*, Rev. Geophys. **48** (2010) RG4004
3. S. Širca, M. Horvat, *Computational Methods for Physicists* (Springer, Berlin, 2012)
4. S. Brandt, *Data Analysis*, 4th edn. (Springer, Berlin, 2014)

5. K.A. Olive et al. (Particle Data Group), The review of particle physics. Chin. Phys. C **38**, 090001 (2014)
6. W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, 3rd edn. (Cambridge University Press, Cambridge, 2007)
7. R.A. Maronna, R.D. Martin, V.J. Yohai, *Robust statistics* (John Wiley & Sons, Chichester, Theory and methods, 2006)
8. T.A. Boden, G. Marland, R.J. Andres, *Global, regional, and national fossil-fuel* CO_2 *emissions*, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, doi:10.3334/CDIAC/00001_V2015
9. NIST Statistical Reference Datasets (2003). The data files and explanations are accessible at http://www.itl.nist.gov/div898/strd/
10. D. Kahaner, C. Moler, S. Nash, *Numerical Methods and Software* (Prentice-Hall, Englewood Cliffs, 1989)
11. T.A. Hahn, Thermal expansion of copper from 20 to 800 K—standard reference material 736. J. Appl. Phys. **41**, 5096 (1970)
12. S.S. Li, W.R. Thurber, The dopant density and temperature dependence of electron mobility and resistivity in *n*-type silicon. Solid State Electron. **20**, 609 (1977)

Chapter 10
Statistical Tests: Verifying Hypotheses

Abstract Statistical tests based on hypotheses are used to statistically verify or disprove, at a certain level of significance, models of populations and their probability distributions. The null and alternative hypothesis are the corner-stones of each such verification, and go hand-in-hand with the possibility of inference errors; these are defined first, followed by the exposition of standard parametric tests for normally distributed variables (tests of mean, variance, comparison of means, variances). Pearson's χ^2-test is introduced as a means to ascertain the quality of regression (goodness of fit) in the case of binned data. The Kolmogorov–Smirnov test with which binned data can be compared to a continuous distribution function or two binned data sets can be compared to each other, is discussed as a distribution-free alternative.

Chapters 7–9 were dealing with methods by which random samples were used to make inferences about populations and to estimate the parameters of their distributions. This chapter introduces tools used to verify—from the statistical viewpoint—whether a population model is acceptable or not [1].

10.1 Basic Concepts

To test the validity of a model we use *hypotheses* about the properties of a population or its probability distribution, for example, "the coin is fair", implying a probability of $p = 0.5$ for heads or tails. The basic hypothesis being tested is called the *null hypothesis* and is denoted by H_0, for instance

$$H_0 : p = 0.5.$$

According to the result of a statistical test the null hypothesis may be accepted or rejected—although "hypothesis accepted" should always be interpreted as "from the statistical perspective the available data is insufficient to reject it"; in the following we should keep in mind this subtle difference. Strictly speaking, one never tests the null hypothesis by itself, but always against its *alternative hypothesis* denoted by H_1,

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_10

for instance,

$$H_1 : p > 0.5.$$

Namely, hypotheses need not be exclusive: one can test, for example, the hypothesis $H_0 : p > 0.4$ with respect to $H_1 : p > 0.6$.

Testing hypothesis also brings about the question of inference *errors*. Imagine a blood test used to determine the presence of a disease in two populations, healthy and sick (see also Problem 1.5.6). Most often it happens that based on the test (values of x) the populations can not be clearly separated (Fig. 10.1 (left)): no matter what $x = x_*$ we choose, for $x > x_*$ a fraction of the sick population will be identified as sick (true positives, TP), and its remainder as healthy (false negatives, FN), while with $x < x_*$ a part of the healthy population will be seen as healthy (true negatives, TN), and the rest as sick (false positives, FP).

Let us discuss only a continuous random variable X, whose distribution is specified by a single-parameter probability density $f(x; \theta)$. Suppose that the experiment yields a value $X = x$, for which we wish to ascertain whether it is consistent with one or another hypothesis. Let the null and alternative hypotheses correspond to the distributions with parameters θ_0 and θ_1, respectively:

$$\begin{aligned} H_0 &: \ \theta = \theta_0, \\ H_1 &: \ \theta = \theta_1, \end{aligned}$$

as shown in Fig. 10.2. If the observed x exceeds the critical value x_*, we reject H_0, otherwise it may be accepted. (Recall our initial warning.) The interval $[x_*, \infty)$ in Fig. 10.2 (left) is therefore called the *rejection region*, while $(-\infty, x_*]$ is the *acceptance region*. Setting the value of x_* is our primary job—we choose in advance the probability α such that the observed x falls in the rejection region. This probability is called the *statistical significance* of the test,

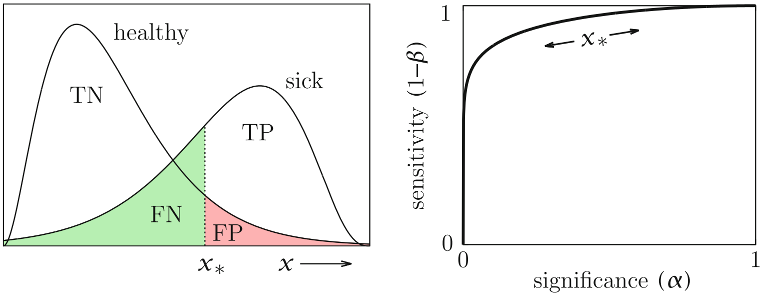


Fig. 10.1 [Left] Distribution with respect to blood test results in healthy and sick populations. One can have true positive (TP), false negative (FN), true negative (TN) and false positive (FP) outcomes. [Right] Comparison of significance (α) and sensitivity ($1 - \beta$) of the test as a measure of its reliability—known as the *Receiver Operating Characteristic* (ROC) curve. See also Example on p. 262

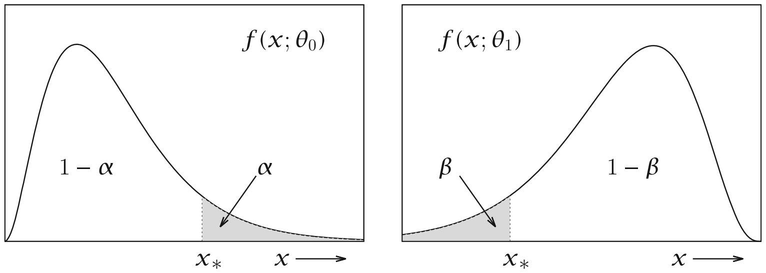


Fig. 10.2 Probability densities corresponding to [Left] Null hypothesis with parameter θ_0 and [Right] Alternative hypothesis with parameter θ_1. Shading indicates the rejection regions for significances α and β. The critical point is denoted by x_*

$$\alpha = \int_{x_*}^{\infty} f(x;\theta_0)\,\mathrm{d}x. \tag{10.1}$$

If H_0 is rejected with significance α, this means that with probability α we have made a wrong conclusion, as the observed x can also take values above x_*: we have rejected a hypothesis when in fact it should have been accepted. We say that we have rejected it at *confidence level* $1-\alpha$. In the language of blood tests this value is also called *specificity*, so that

$$\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} = \alpha \ \text{(significance)}, \qquad \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1-\alpha \ \text{(specificity)}.$$

We can also make a different error, namely, accept H_0 as true although it is, in fact, false and H_1 should have been accepted instead. The probability for this type of error depends on H_1 (Fig. 10.2 (right)) and is denoted by β:

$$\beta = \int_{-\infty}^{x_*} f(x;\theta_1)\,\mathrm{d}x.$$

The *power* of a statistical test or its *sensitivity* is defined as the probability that a hypothesis is rejected when it is indeed false. The power of testing the null hypothesis H_0 against the alternative hypothesis H_1 is

$$1-\beta = \int_{x_*}^{\infty} f(x;\theta_1)\,\mathrm{d}x, \tag{10.2}$$

therefore, in the blood-test example,

$$\frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} = \beta, \qquad \frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{TP}} = 1-\beta \ \text{(sensitivity)}.$$

How does a test attain large power? The integral (10.2) can be written as

$$\int_{x_*}^{\infty} f(x;\theta_1)\,\mathrm{d}x = \int_{x_*}^{\infty} \frac{f(x;\theta_1)}{f(x;\theta_0)}\, f(x;\theta_0)\,\mathrm{d}x = \left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right)_{x=\xi} \underbrace{\int_{x_*}^{\infty} f(x;\theta_0)\,\mathrm{d}x}_{\alpha},$$

where $\xi \in [x_*, \infty)$. Hence the power is large if $f(x;\theta_1)/f(x;\theta_0)$ is large or—in the case of a sequence of observations $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$—where the ratio

$$\frac{L(\boldsymbol{x};\theta_1)}{L(\boldsymbol{x};\theta_0)} = \frac{\prod_{i=1}^{n} f(x_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)} \tag{10.3}$$

exceeds a prescribed constant that depends on α. Usually $1-\beta$ rapidly increases with α and is close to 1 above $\alpha \gtrsim 0.10$ (see Fig. 10.1 (right)). The more the curve approaches the top left corner, the larger the predictive power of the test; in the extreme case (point $(\alpha, 1-\beta) = (0, 1)$) the populations are completely separated. Conventionally one chooses $0.01 \lesssim \alpha \lesssim 0.10$. How this works in practice is demonstrated by the following Example and Problem 10.5.1.

Example Photo-disintegration of ^{3}He nuclei below the pion production threshold involves two decay channels, two-body (2bbu) and three-body (3bbu) breakup:

$$\begin{aligned} H_0(\text{2bbu}):&\quad \gamma + {}^3\text{He} \longrightarrow \text{p} + \text{d},\\ H_1(\text{3bbu}):&\quad \gamma + {}^3\text{He} \longrightarrow \text{p} + \text{p} + \text{n}. \end{aligned}$$

Experimentally they can be distinguished by calculating the so-called missing energy, i.e. the difference between the photon energy and the kinetic energy of the final-state proton, $E_\mathrm{m} = E_\gamma - T_\mathrm{p}$, where both E_γ and T_p have some measurement uncertainties. An example of a measured spectrum is shown in Fig. 10.3 (left). The peak at $E_\mathrm{m} \approx$

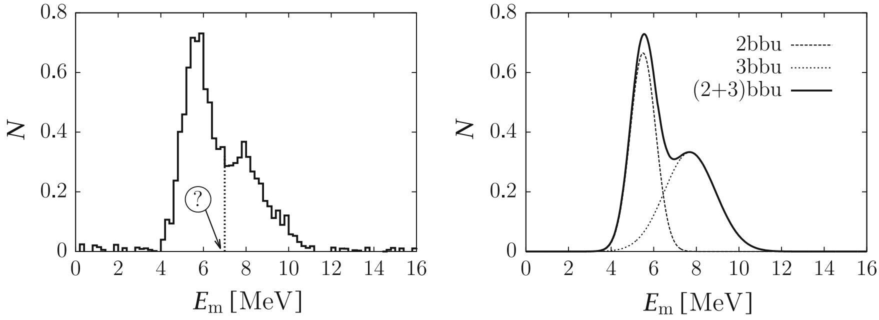


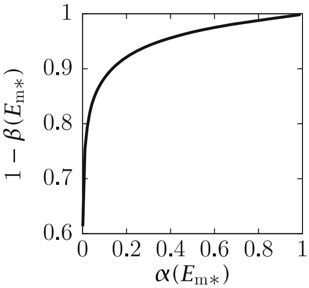
Fig. 10.3 Normalized distribution of events with respect to the missing energy in the processes $\gamma + {}^3\text{He} \longrightarrow \text{p} + \text{d}$ (two-body breakup, 2bbu) and $\gamma + {}^3\text{He} \longrightarrow \text{p} + \text{p} + \text{n}$ (three-body breakup, 3bbu). [Left] Measured spectrum. [Right] Theoretical spectrum

5.5 MeV corresponds to the separation energy of the proton in ^{3}He, while the bump above $E_\mathrm{m} \approx 7.7$ MeV is a witness to the additional 2.2 MeV required to split the remaining deuteron to a proton and a neutron.

How do we choose the critical $E_{\mathrm{m}*}$ to test the hypothesis that a detected proton comes from the two-body process? The higher we set it, the more certain we can be that the sample will contain all "true" protons from 2bbu, but at the same time more an more "false" protons from 3bbu will be identified as belonging to 2bbu. If $E_{\mathrm{m}*}$ is set very low, we may reduce the contamination of the 2bbu sample by protons from 3bbu, but at the same time we discard a significant fraction of true 2bbu events. Ideally we wish to minimize the probability α for the rejection of H_0 when it is actually correct, and minimize the probability β for the acceptance of H_0, when it is actually false.

How can this be accomplished? Suppose we have a theoretical model of the breakup processes that fits the data well (Fig. 10.3 (right)). The probability densities $f(E_\mathrm{m}; H_0)$ and $f(E_\mathrm{m}; H_1)$ describe the protons from 2bbu and 3bbu, respectively. (Both densities can be normalized to the number of counts, or vice-versa.) For the chosen statistical significance of the test, α, we first establish $E_{\mathrm{m}*}$ for which

$$\int_{E_{\mathrm{m}*}}^{\infty} f(E_\mathrm{m}; H_0)\,\mathrm{d}E_\mathrm{m} = \alpha(E_{\mathrm{m}*}).$$



With this $E_{\mathrm{m}*}$ we calculate the power of the test

$$1 - \beta(E_{\mathrm{m}*}) = \int_{E_{\mathrm{m}*}}^{\infty} f(E_\mathrm{m}; H_1)\,\mathrm{d}E_\mathrm{m}.$$

The model in Fig. 10.3 (right) is a sum of normal distributions $N(\mu_0, \sigma_0^2)$ for 2bbu and $N(\mu_1, \sigma_1^2)$ for 3bbu, with parameters $\mu_0 = 5.5$ MeV, $\sigma_0 = 0.6$ MeV, $\mu_1 = 7.7$ MeV and $\sigma_1 = 1.2$ MeV. The critical value $E_{\mathrm{m}*}$ at chosen α is therefore nothing but the corresponding quantile of the normal distribution, and the integral for $1 - \beta$ is its definite integral. In both cases we may use formula (3.9) and Tables D.1 and D.2. The dependence of $(1 - \beta)$ on α—the ROC curve—is shown in the above figure. ◁

10.2 Parametric Tests for Normal Variables

10.2.1 Test of Sample Mean

Let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ be a random sample drawn from a normally distributed population ($N(\mu, \sigma^2)$). "Testing the mean" implies that we shall use the sample mean $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$ to provide some sort of statement on the true (population) mean μ, which is unknown. Two cases must be distinguished: the population variance σ^2 is known or unknown. If σ^2 is known, testing the hypothesis

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis $H_1 : \mu \neq \mu_0$ can be accomplished by using the statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1),$$

which is distributed according to the standardized normal distribution. If σ^2 is unknown, we can replace it by the unbiased sample variance (7.10) and use the statistic

$$T = \frac{\overline{X} - \mu_0}{s_X/\sqrt{n}} \sim t(n-1),$$

which is distributed according to the Student's t distribution with $(n-1)$ degrees of freedom. Of course, if the sample variance is used in the biased form (7.8), one should replace $\sqrt{n}$ by $\sqrt{n-1}$. See also Sect. 7.3.1.

Example By using twelve ($n = 12$) identical thermometers we have measured the temperatures

$$\boldsymbol{x} = \{33.6,\ 34.3,\ 32.6,\ 32.8,\ 34.1,\ 34.9,\ 32.7,\ 33.9,\ 33.1,\ 32.5,\ 33.1,\ 33.4\}\,^\circ\mathrm{C}.$$

May we claim, with statistical significance $\alpha = 0.05$, that the true temperature during the measurement was higher than $\mu_0 = 32.8\,^\circ\mathrm{C}$?

The null hypothesis is $H_0 : \mu = \mu_0 = 32.8\,^\circ\mathrm{C}$, the alternative hypothesis is $H_1 : \mu > \mu_0$. We wish to reject H_0 if the sample mean $\overline{x}$ exceeds μ_0. From the data we calculate $\overline{x} = 33.42\,^\circ\mathrm{C}$ and $s_x^2 = 0.574\,(^\circ\mathrm{C})^2$. The value of the statistic T is

$$t = \frac{\overline{x} - \mu_0}{s_x/\sqrt{n}} = \frac{33.42\,^\circ\mathrm{C} - 32.8\,^\circ\mathrm{C}}{0.758\,^\circ\mathrm{C}/\sqrt{12}} \approx 2.83.$$

Table D.4, row $\nu = n - 1 = 11$, reveals that $t = 2.83$ lies between $t_{0.99} = 2.72$, corresponding to $\alpha = 0.01$, and $t_{0.995} = 3.11$, corresponding to $\alpha = 0.005$. The required α corresponds to the critical $t_* = t_{0.95} = 1.80$. Since $t > t_*$, we may reject H_0 and accept the hypothesis H_1 that the actual temperature exceeded 32.8 °C, with significance $\alpha = 0.05$ (confidence level $1 - \alpha = 0.95$). ◁

10.2.2 Test of Sample Variance

Testing the variance based on the sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ from a normally distributed population ($X \sim N(\mu, \sigma^2)$) again requires us to distinguish two cases: that the true mean μ is known or unknown. If μ is known, the hypothesis

$$H_0 : \sigma^2 = \sigma_0^2$$

against $H_1 : \sigma^2 \neq \sigma_0^2$ can be tested by using the statistic

$$X^2 = \frac{ns_X^2}{\sigma_0^2} = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} \bigl(X_i - \mu\bigr)^2 \sim \chi^2(n),$$

which is distributed according to the χ^2 distribution with n degrees of freedom. (Here s_X^2 is taken in its biased form (7.8).) If μ is unknown, it must be replaced by the sample mean $\overline{X} = (1/n) \sum_{i=1}^{n} X_i$ in the above formula. Then the statistic X^2 is also χ^2-distributed, but with $(n - 1)$ degrees of freedom. See also Sect. 7.2.2.

Example To construct a detector we need many wire electrodes of a specific length. The largest allowed length tolerance is $\sigma_0^2 = 100\,(\mu\mathrm{m})^2$. A precise measurement of the length is very demanding, so we can only afford a small sample of $n = 10$ electrodes, for which we establish a variance of $s_x^2 = 142\,(\mu\mathrm{m})^2$. Given a statistical significance of $\alpha = 0.05$, does the wire length in the unexplored "population" fluctuate exceedingly?

The null hypothesis is $H_0 : \sigma^2 = \sigma_0^2$, while the alternative hypothesis is $H_1 : \sigma^2 > \sigma_0^2$. We may reject H_0 if the sample variance exceeds the critical variance (at given α). The value of the test statistic is $x^2 = ns_x^2/\sigma_0^2 = 10 \cdot 142/100 = 14.2$. The critical x^2 can be read off from Table D.3, row for $\nu = n - 1 = 9$, column for $p = 1 - \alpha = 0.95$: it is $x_*^2 = \chi_{0.95}^2 = 16.9$. Since $x^2 < x_*^2$, we have no reason (at confidence level $1 - \alpha = 95\%$) to reject H_0. We may conclude that the variance of all electrodes is within the prescribed limits. ◁

10.2.3 Comparison of Two Sample Means, $\sigma_X^2 = \sigma_Y^2$

Assume we have *two* samples, $\boldsymbol{x} = \{x_i\}_{i=1}^{n_x}$ and $\boldsymbol{y} = \{y_i\}_{i=1}^{n_y}$, drawn from normally distributed populations with different means and equal, but unknown variances $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, that is, $X \sim N(\mu_X, \sigma^2)$ and $Y \sim N(\mu_Y, \sigma^2)$. A situation like this occurs, for instance, when we apply the same technique to measure a quantity that might have changed during the measurements. By comparing the samples $\boldsymbol{x}$ and $\boldsymbol{y}$ we test the hypothesis H_0 that they stem from populations with a specific difference between the true means,

$$H_0 : \mu_X - \mu_Y = (\mu_X - \mu_Y)_0,$$

against the alternative $H_1 : \mu_X - \mu_Y \neq (\mu_X - \mu_Y)_0$. The suitable statistic is

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{s_{XY}}, \qquad s_{XY} = \sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right) \frac{n_x s_X^2 + n_y s_Y^2}{n_x + n_y - 2}},$$

where $\overline{X}$ and $\overline{Y}$ are the sample means, while s_X^2 and s_Y^2 are biased sample variances (7.8). The statistic T is t-distributed, with $(n_x + n_y - 2)$ degrees of freedom.

Example A laboratory uses two chemicals to determine the concentration of a compound in blood. Presently a reliable, but expensive chemical is used, and it has yielded a sample of $n_x = 10$ concentrations

$$\boldsymbol{x} = \{7.41,\ 7.59,\ 7.56,\ 7.77,\ 8.06,\ 8.04,\ 7.57,\ 7.74,\ 7.47,\ 7.98\}$$

(in some units). With the cheaper chemical we obtain a sample of $n_y = 8$ values

$$\boldsymbol{y} = \{7.59,\ 7.83,\ 7.63,\ 7.94,\ 7.69,\ 7.46,\ 7.94,\ 7.38\}.$$

May we claim that there is a statistically significant difference between the average concentrations (with significance $\alpha = 0.05$)?

The null hypothesis is $H_0 : \mu_X - \mu_Y = 0$ (equal concentrations), and the alternative is $H_1 : \mu_X - \mu_Y \neq 0$. From the sample averages $\overline{x} = 7.719$ and $\overline{y} = 7.683$ and sample variances $s_x^2 = 0.0512$ and $s_y^2 = 0.0383$ we calculate $s_{xy} = 0.107$, then the value of the test statistic $t = (7.719 - 7.683 - 0)/0.107 \approx 0.34$. Table D.4, row $\nu = n_x + n_y - 2 = 16$, tells us that $t \approx 0.34$ lies between $t_{0.60} = 0.258$, corresponding to $1 - p = \alpha = 0.40$, and $t_{0.75} = 0.535$, corresponding to $\alpha = 0.25$. The required $\alpha/2 = 0.025$—we must perform a two-sided test, as the alternative hypothesis means either $\mu_X > \mu_Y$ or $\mu_X < \mu_Y$—corresponds to the critical value $t_* = 2.12$. Since $t < t_*$, there is no reason to reject the null hypothesis: both chemicals are equally effective, so in order to reduce costs, we may purchase the cheaper one. ◁

10.2.4 Comparison of Two Sample Means, $\sigma_X^2 \neq \sigma_Y^2$

A similar test can be performed for samples $\boldsymbol{x}$ and $\boldsymbol{y}$ of sizes n_x and n_y, presumably stemming from normal populations with different (and unknown) variances. The test statistic is

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{s_X^2/n_x + s_Y^2/n_y}},$$

where $\overline{X}$ and $\overline{Y}$ are the sample means, and s_X^2 and s_Y^2 are the *unbiased* variances (7.10). If H_0 is true, T is normally distributed ($N(0, 1)$) when $n_x, n_y \gg 1$.

10.2.5 Comparison of Two Sample Variances

Comparing the variances of two samples $\boldsymbol{x}$ and $\boldsymbol{y}$ is another classic: we thereby test the hypothesis whether the corresponding population variances are equal,

$$H_0 : \sigma_X^2 = \sigma_Y^2.$$

The test statistic is the ratio of *unbiased* sample variances,

$$F = s_X^2/s_Y^2.$$

If the null hypothesis is valid, this ratio is distributed according to the F distribution with $(\nu_x, \nu_y) = (n_x - 1, n_y - 1)$ degrees of freedom. Given the significance α three alternative hypotheses H_1 can be formulated:

$$\begin{aligned}
H_1 &: \sigma_X^2 \neq \sigma_Y^2 \;;\; H_0 \text{ rejected if } \; F < F_{\alpha/2} \;\text{ or }\; F > F_{1-\alpha/2},\\
H_1 &: \sigma_X^2 > \sigma_Y^2 \;;\; H_0 \text{ rejected if } \; F > F_{1-\alpha},\\
H_1 &: \sigma_X^2 < \sigma_Y^2 \;;\; H_0 \text{ rejected if } \; F < F_{\alpha}.
\end{aligned}$$

Here $F_{\alpha/2}$, F_α, $F_{1-\alpha}$ and $F_{1-\alpha/2}$ are the F-distribution quantiles (Tables D.5 and D.6).

Example In sputtering of thin metal layers on semiconductor substrate wafers we strive to minimize the variance of the layer thickness. The variance in the sample of $n_x = 16$ layers fabricated with oven X is $s_x^2 = 0.058\,(\text{nm})^2$, while the variance in the second sample of $n_y = 10$ layers made with oven Y is found to be $s_y^2 = 0.079\,(\text{nm})^2$. Can we claim that *any* of the two ovens makes layers whose thickness is more precise? Let the significance of the test be $\alpha = 0.10$.

The null hypothesis H_0 may be rejected if the sample f (value of statistic F) satisfies $f < F_{*-} = F_{\alpha/2}(n_x - 1, n_y - 1)$ or $f > F_{*+} = F_{1-\alpha/2}(n_x - 1, n_y - 1)$. From Table D.5 and from the symmetry of the F distribution (see Fig. 7.3) we get

$$\begin{aligned}
F_{*-} &= F_{0.05}(15, 9) = 1/F_{0.95}(9, 15) = 1/2.59 = 0.386,\\
F_{*+} &= F_{0.95}(15, 9) = 3.01.
\end{aligned}$$

The samples give $f = s_x^2/s_y^2 = 0.734$. Since neither $f < F_{*-}$ nor $f > F_{*+}$ apply, H_0 can not be rejected at confidence level $1 - \alpha = 0.90$. The ovens are equally fine. (However, the samples *are* small, thus the conclusion is a bit risky.) ◁

Example The F test can also be applied to determine the maximum sensible degree of the polynomial used to fit the data $\{(x_i, y_i \pm \sigma_i)\}_{i=1}^n$, given a confidence level $1 - \alpha$. The test can be used to distinguish among so-called *nested* models, in which a richer function inherits all parameters of the subordinate one and adds its own: the model $f_2(x; \boldsymbol{\theta}) = \theta_1 + \theta_2 x$, for instance, is nested within $f_3(x; \boldsymbol{\theta}) = \theta_1 + \theta_2 x + \theta_3 x^2$. In general the more modest ansatz has q, while the richer has p parameters, $p > q$. Is the former sufficient to describe the data ($H_0 : \theta_{q+1} = \theta_{q+2} = \cdots = \theta_p = 0$) or should new terms be included (H_1: at least one of the enumerated parameters $\neq 0$)? For both models we calculate the sum of squared residuals X_q^2 and X_p^2 by using (9.6) and form the statistic

$$F = \frac{(X_q^2 - X_p^2)/(p-q)}{X_p^2/(n-p-1)},$$

which is distributed according to $F(p-q,\ n-p-1)$. If the calculated F exceeds the critical value $F_{1-\alpha}(p-q, n-p-1)$, H_0 can be rejected; "further parameters are needed". In polynomial regression, $f(x; \boldsymbol{\theta}) = \sum_{j=1}^p \theta_j x^{j-1}$, the statistic

$$F = \frac{X_{p-1}^2 - X_p^2}{X_p^2/(n-p-1)} \sim F(1, n-p-1) \qquad (10.4)$$

therefore "measures" whether the inclusion of a new (pth) parameter—hence the next, $(p-1)$th polynomial degree—is justified or not [2].

Figure 10.4 (left) shows the data ($n = 77$) and polynomial fits of various orders, while Fig. 10.4 (right) shows $X_p^2/(n-p)$ and the statistic F (10.4). Let us choose $\alpha = 0.05$. Since $n \gg 1$, $F_* = F_{1-\alpha}(1, n-p-1) \approx 4$ for all shown p (Table D.5).

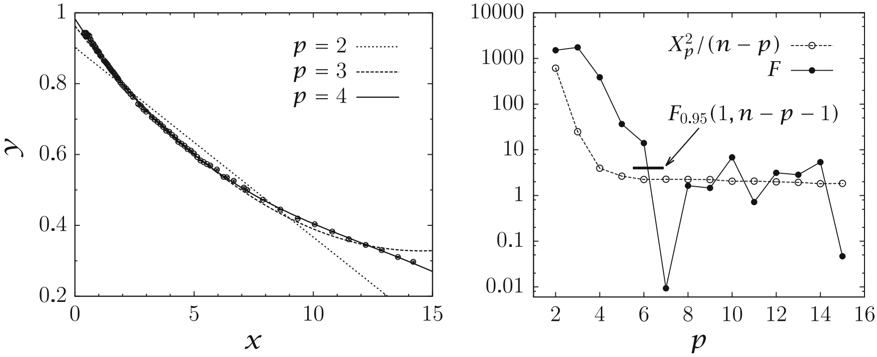


Fig. 10.4 Using the F test to determine the maximum degree of the regression polynomial. [Left] Data (see website of the book). [Right] Dependence of $X_p^2/(n-p)$ and the statistic F on the number of parameters p (polynomial of degree $p-1$)

At $p = 7$ the calculated F falls below F_* for the first time. Hence, from $p = 6$ upwards—even though X_p^2 keeps on dropping—the data do not offer sufficient statistical support to keep on "inflating" the model. ◁

10.3 Pearson's χ^2 Test

In Sect. 9.2 we have seen how the observations $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$, classified in N mutually exclusive bins, can be fitted by a chosen function. Now we are interested in the *goodness* of this fit. If x_i are the values of a random variable with the probability density f, ascertaining the quality of the fit amounts to testing the hypothesis

$$H_0 : f(x) = f_0(x),$$

where f_0 is the chosen density. The same applies in the discrete case, where f and f_0 are probability functions. In the ith bin one expects f_i counts, while in fact n_i are observed. The null hypothesis is $H_0 : f_1 = f_{01},\ f_2 = f_{02}, \ldots, f_N = f_{0N}$, and the test statistic is already familiar from (9.25):

$$X^2 = \sum_{i=1}^{N} \frac{\left(n_i - f_{0i}\right)^2}{f_{0i}}. \tag{10.5}$$

If H_0 is true, X^2 is distributed according to the χ^2 distribution with $N - p$ degrees of freedom, where p is the number of parameters estimated *from the sample* and taken into account in the formulation of the hypothesis. This is an important detail; namely, if (10.5) is written such that the ith bin corresponds to a theoretical probability p_i, while it actually contains n_i counts, the null hypothesis is $H_0 : p_1 = p_{01}, p_2 = p_{02}, \ldots, p_N = p_{0N}$, where $\sum_{i=1}^{N} p_{0i} = 1$ or, equivalently, $H_0 : n_1 = n_{01},\ n_2 = n_{02}, \ldots, n_N = n_{0N}$, where $\sum_{i=1}^{N} n_{0i} = n$. The appropriate test statistic is then

$$X^2 = \sum_{i=1}^{N} \frac{\left(n_i - np_{0i}\right)^2}{np_{0i}} \sim \chi^2(N - p - 1).$$

Example Let us revisit the bombing of London during World War II (Example on pp. 133–134). All 576 quadrants with 0, 1, 2, 3, 4 or ≥ 5 hits in each were classified in $N = 6$ bins $\{n_0, n_1, n_2, n_3, n_4, n_5\} = \{229, 211, 93, 35, 7, 1\}$. From 537 hits we calculated the expected number of hits in any quadrant, $\overline{n} = 537/576 \approx 0.9323$. If the hit distribution were Poissonian, with average $\overline{n}$, we would expect $\{f_{01}, f_{02}, f_{03}, f_{04}, f_{05}\} = \{226.74, 211.39, 98.54, 30.62, 7.14, 1.57\}$ quadrants with 0, 1, 2, 3, 4 or ≥ 5 hits, respectively. Let us test the hypothesis $H_0 : f_{\text{measured}} = f_0 = f_{\text{Poisson}}$ with confidence level $1 - \alpha = 0.90$!

In the formulation of the hypothesis *two* parameters have been fixed: the average $\overline{n}$ and the normalization $n = \sum_{i=1}^{N} f_{0i} = 576$, hence the statistic (10.5) is χ^2-distributed, with $N - 2 = 4$ degrees of freedom. We may reject H_0 if the observed x^2 exceeds the critical value $\chi^2_* = 7.78$ (see Table D.3 for $p = 0.90$ and $\nu = 4$). From the data we get $x^2 = 1.17 < \chi^2_*$. Hence H_0 can not be rejected: the observed distribution is consistent with the Poisson distribution. ◁

Example In an experiment we measure the distribution of events with respect to x that takes the values on the interval $[-2.75, 2.75]$. Due to instrumental restrictions we are able to measure only on a restricted range, $x \in [-1.35, 1.35]$. A total of $n = 838$ events are classified in $N = 27$ bins, as shown in Fig. 10.5. Is the measured distribution consistent with a uniform or, rather, normal distribution? Consider $\alpha = 0.05$ and $\alpha = 0.01$.

There are $n = 838$ counts in all bins. Let us first check the consistency of the data with the uniform distribution, where each bin is expected to contain $f_{0i} = n/N = 31.04$ counts. By using (10.5) we get $x^2 = 59.38$. Dividing this by the number of degrees of freedom $\nu = N - 1 = 26$ yields the so-called *reduced value* $\chi^2_{\mathrm{red}} = x^2/\nu = 2.28$. This should be compared to the critical χ^2_*/ν at chosen α (Fig. 10.6 or Table D.3). For $\alpha = 0.05$ we read off $\chi^2_*/\nu \approx 1.50$, while for $\alpha = 0.01$ we see that $\chi^2_*/\nu \approx 1.76$. In either case $\chi^2_{\mathrm{red}} = x^2/\nu > \chi^2_*/\nu$, indicating that the measured distribution does not match the uniform distribution.

We have narrowed down the acceptance region in Fig. 10.5, as this often happens in practice, forcing us to see the data as nothing but "a constant". What do we obtain with f_{0i} corresponding to the normal distribution? Assume that only its standard deviation has been determined from the data, so $\nu = 26$. Now we obtain $x^2 = 35.66$ or $\chi^2_{\mathrm{red}} = x^2/\nu = 1.37$, which is less than $\chi^2_*(\alpha = 0.05)/\nu$ and less than $\chi^2_*(\alpha = 0.01)/\nu$. We can therefore claim, with confidence level at least 99%, that the measured and normal distribution are mutually consistent. ◁

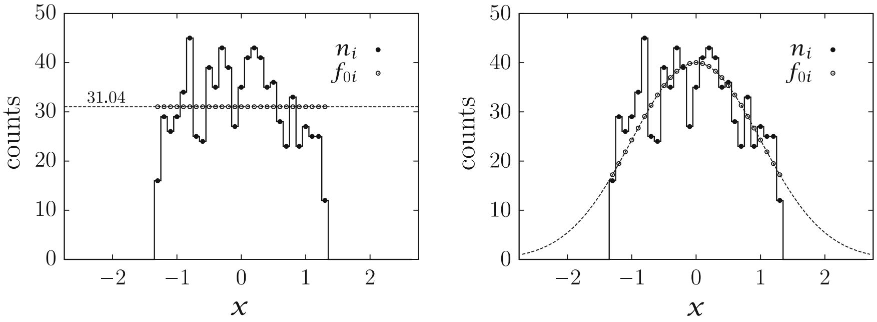


Fig. 10.5 Pearson's χ^2 test for checking the consistency of the binned data with an assumed theoretical model. [Left] Comparison of data to the uniform distribution. [Right] Comparison of data to the normal distribution

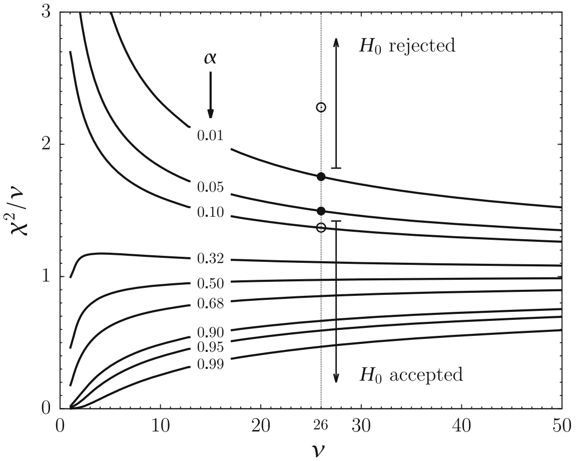


Fig. 10.6 Reduced value of χ^2/ν for Pearson's test as a function of the number of degrees of freedom ν at chosen significance α. The ● symbols at $\nu = 26$ denote the critical points $\chi^2_*(0.05)/\nu = 1.50$ and $\chi^2_*(0.01)/\nu = 1.76$ for the Example in Fig. 10.5, while ○ indicate the calculated values of $x^2/\nu = 2.28$ and $x^2/\nu = 1.37$ corresponding to the comparison of the data to the uniform and normal distribution, respectively

10.3.1 Comparing Two Sets of Binned Data

The same test can be used to ascertain the mutual consistency of two sets of histogrammed data of size m and n in N bins. The appropriate test statistic is

$$X^2 = \sum_{i=1}^{N} \frac{\left(\sqrt{n/m}\, m_i - \sqrt{m/n}\, n_i\right)^2}{m_i + n_i}, \qquad m = \sum_{i=1}^{N} m_i, \quad n = \sum_{i=1}^{N} n_i. \tag{10.6}$$

In general $m \neq n$. The χ^2-test with the chosen significance α is performed as before, by using the χ^2 distribution with $\nu = N - 1$ degrees of freedom. The values $x^2/\nu > \chi^2_*/\nu$ indicate that the observations m_i and n_i do not come from the same distribution law. An example is given in Problem 10.5.2.

10.4 Kolmogorov–Smirnov Test

Kolmogorov–Smirnov (KS) test [3, 4] is a non-parametric test used to establish the probability that the observed data (sample) stems from a population distributed according to the chosen continuous theoretical distribution, or that one sample comes from the same population as the other. Of course both the data and the model distribution can be binned and compared by Pearson's test, but a direct comparison has several advantages.

First we sort the sample $\{x_i\}_{i=1}^{n}$ so that $x_1 \leq x_2 \leq \cdots \leq x_n$. For this sorted set we define the *empirical distribution function*

$$\widetilde{F}_n(x) = \begin{cases} 0 & ; \; x < x_1, \\ i/n & ; \; x_i \le x < x_{i+1}, \quad i = 1, 2, \ldots, n-1, \\ 1 & ; \; x \ge x_n. \end{cases}$$

This is a monotonously increasing function that jumps upwards by $1/n$ at each point x_i. For the data

$$\begin{aligned} \boldsymbol{x} = \{&0.22, \; -0.87, \; -2.39, \; -1.79, \; 0.37, \; -1.54, \; 1.28, \; -0.31, \; -0.74, \; 1.72, \\ &0.38, \; -0.17, \; -0.62, \; -1.10, \; 0.30, \; 0.15, \; 2.30, \; 0.19, \; -0.50, \; -0.09\} \end{aligned} \tag{10.7}$$

it is shown by the "staircase" curve in Fig. 10.7 (left).

In the basic version of the KS test the empirical distribution $\widetilde{F}_n$ is compared to the model distribution F, shown by the smooth curve in the figure. The null hypothesis is

$$H_0 : \widetilde{F}_n(x) = F(x).$$

The test statistic is the maximum distance between these distributions,

$$D_n = \sup_x \left| \widetilde{F}_n(x) - F(x) \right| = \max_{1 \le i \le n} \left\{ \frac{i}{n} - F\big(x_i\big), \; F\big(x_i\big) - \frac{i-1}{n} \right\}. \tag{10.8}$$

The smaller the distance d_n (value of statistic D_n), the better the agreement between $\widetilde{F}_n$ and F, pointing to the acceptance of the null hypothesis. If, however, the calculated d_n is larger than the critical value $d_*(n; \alpha) \equiv d(\alpha)/\sqrt{n}$ at chosen α, H_0 may be rejected. The critical values are tabulated; a method (and a MATLAB code) to compute them can be found in [5]. The symbols in Fig. 10.7 (right) represent $d_*(n; \alpha)$ for a

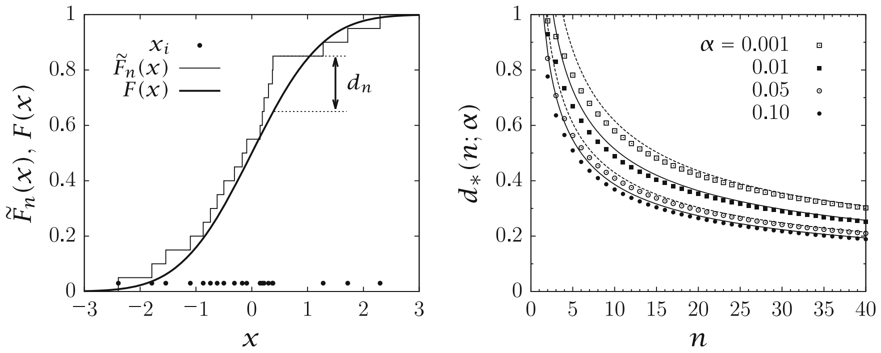


Fig. 10.7 Kolmogorov–Smirnov test. [Left] Data sample (•), the corresponding empirical distribution function $\widetilde{F}_n$, model distribution function F and the greatest distance between them, d_n. [Right] Critical values d_* as a function of n for various statistical significances α. The curves correspond to the asymptotic formulas (10.9)

few typical α. The figure also contains the asymptotic curves

$$d_*(n;\,\alpha) \approx \frac{1}{\sqrt{n}}\sqrt{-0.5\log(\alpha/2)}, \qquad n \gtrsim 35. \tag{10.9}$$

It is a most charming property of the KS test that the distribution of the statistic D_n is known and, moreover, *does not depend on the distribution* F. In the $n \gg 1$ limit it holds that

$$F_{\mathrm{KS}}(z) = \lim_{n\to\infty} P\left(D_n \le \frac{z}{\sqrt{n}}\right) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}\mathrm{e}^{-2k^2z^2}, \tag{10.10}$$

which is suitable for the calculation at large z, while the form

$$F_{\mathrm{KS}}(z) = \frac{\sqrt{2\pi}}{z}\sum_{k=1}^{\infty}\exp\left(-\frac{(2k-1)^2\pi^2}{8z^2}\right)$$

is preferable for small z. In either case

$$P\big(D_n > z\big) = 1 - P\big(D_n \le z\big) = 1 - F_{\mathrm{KS}}\big(\sqrt{n}\,z\big). \tag{10.11}$$

Everyday work is made simpler by the approximation

$$P\big(D_n > z\big) \approx 1 - F_{\mathrm{KS}}\big[\sqrt{n_{\mathrm{eff}}}\,z\big], \qquad \sqrt{n_{\mathrm{eff}}} = \sqrt{n} + 0.12 + 0.11/\sqrt{n}, \tag{10.12}$$

which works well already for $n \gtrsim 5$ and has the correct asymptotics. This can be exploited for the calculation of the critical $d_*(n;\,\alpha)$ for arbitrary, even small n, if tables are not at hand. Namely, one can insert $z = d_*$ in (10.10) to obtain

$$\alpha \approx 2\sum_{k=1}^{\infty}(-1)^{k-1}\mathrm{e}^{-2n_{\mathrm{eff}}k^2d_*^2}.$$

With given n we must figure out d_* such that the sum on the right equals the chosen α on the left. When we succeed, we have found $d_* = d_*(n_{\mathrm{eff}};\,\alpha)$. (Do this as an exercise! Replace guesswork by bisection.)

Example At significance $\alpha = 0.05$ we wish to test the null hypothesis that the sample (10.7) has been drawn from a standard normal population with distribution function F (see (3.11), $F = \Phi$). By using (10.8) we obtain $d_n = 0.202$ indicated in Fig. 10.7. The sample size is $n = 20$, and the exact critical value is $d_*(20;\,0.05) = 0.294$. By formula (10.9), not expected to apply at such low n, one gets $d_*(20;\,0.05) = 0.304$. In either case $d_n < d_*$, the hypothesis can not be rejected. The data are consistent with the normal distribution. ◁

The advantage that the distribution of D_n is independent of the model distribution F can be exploited for the determination of confidence regions for the true distribution function of the population, F_0. Namely, realizing that

$$P\big(D_n = \sup_x \big|\widetilde{F}_n(x) - F(x)\big| > d_*(n;\alpha)\big) = \alpha,$$

at chosen significance α, this also means that

$$P\big(\widetilde{F}_n(x) - d_*(n;\alpha) < F_0(x) < \widetilde{F}_n(x) + d_*(n;\alpha)\big) = 1 - \alpha \qquad \forall x.$$

Therefore, at any x the value of the true distribution function F_0 lies between $\widetilde{F}_n(x) - d_*$ and $\widetilde{F}_n(x) + d_*$ with probability $1 - \alpha$. In other words, if we use the calculated d_* to create a band about the empirical distribution, the curve of $F_0(x)$ lies within this band with probability $1 - \alpha$.

10.4.1 Comparison of Two Samples

Instead of comparing sample and model distributions the KS test can also be applied to two samples, which may even have different sizes [6]. Let the samples of size m and n have the empirical distribution functions $\widetilde{F}_m(x)$ and $\widetilde{F}_n(x)$, respectively. The null hypothesis is that the samples originate in the population with the same distribution function. In this case the test statistic is

$$D_{mn} = \sup_x \big|\widetilde{F}_m(x) - \widetilde{F}_n(x)\big|.$$

The critical values $d_*(m,n;\alpha)$ that the value of the statistic D_{mn} should exceed in order for the null hypothesis to be rejected at chosen α, are tabulated for small m and n. Fortunately, it turns out that (10.11) can be replaced by

$$P\big(D_{mn} > z\big) \approx 1 - F_{\mathrm{KS}}\left(\sqrt{\frac{mn}{m+n}}\,z\right), \qquad m, n \gg 1,$$

so the critical values for the two-sample test in the limit of large m and n are just suitably rescaled critical values of the one-sample test:

$$d_*(m,n;\alpha) = \sqrt{1 + \frac{n}{m}}\,d_*(n;\alpha) \approx \sqrt{\frac{m+n}{mn}}\sqrt{-0.5\log(\alpha/2)}. \tag{10.13}$$

For small m and n one may again use the empirical parameterization (10.12) with the replacement $n \longrightarrow mn/(m+n)$.

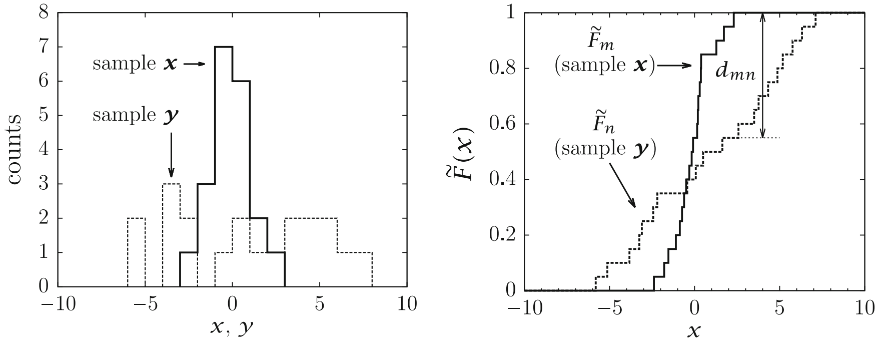


Fig. 10.8 Comparison of two samples by the Kolmogorov–Smirnov test. [Left] Histogrammed data (10.7) and (10.14). [Right] Corresponding empirical distribution functions and the maximum distance between them

Example We wish to examine the null hypothesis that sample (10.7) and sample

$$\boldsymbol{y} = \{-5.13,\ -2.19,\ -2.43,\ -3.83,\ 0.50,\ -3.25,\ 4.32,\ 1.63,\ 5.18,\ -0.43,\\ 7.11,\ 4.87,\ -3.10,\ -5.81,\ 3.76,\ 6.31,\ 2.58,\ 0.07,\ 5.76,\ 3.50\} \qquad (10.14)$$

originate in the population with the same distribution function. The samples have equal sizes, $m = n = 20$. Their histograms are shown in Fig. 10.8 (left), and their empirical distribution functions are shown in Fig. 10.8 (right).

The maximum distance between $\widetilde{F}_m$ and $\widetilde{F}_n$ is $d_{mn} = 9/20 = 0.45$. The critical values computed by formula (10.13) are $d_*(0.10) = 0.387$, $d_*(0.05) = 0.429$, $d_*(0.01) = 0.515$ (the exact ones are 0.350, 0.400, 0.500). Hence the null hypothesis can be rejected at significance α between 1 and 5%. ◁

10.4.2 Other Tests Based on Empirical Distribution Functions

The KS test is perhaps the most popular test based on empirical distributions, but others exist [7]. Let us name some general considerations regarding their use. Instead of the distance between $\widetilde{F}_n$ and F, for example, one could also measure the average square of the deviation of $\widetilde{F}_n$ from F by calculating the integral

$$\int_{-\infty}^{\infty} \left[\widetilde{F}_n(x) - F(x)\right]^2 w(x)\, \mathrm{d}F(x),$$

where $w(x)$ is a weight function. By choosing different $w(x)$ greater emphasis can be given to certain portions of the definition domain of a distribution or its specific aspects. Setting

$$w(x) = \frac{n}{F(x)(1-F(x))},$$

for instance, yields the Anderson-Darling (AD) test [8, 9], which is more sensitive to the distribution tails, where $F(x)$ and $1-F(x)$ are small. The corresponding statistic is

$$W_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\bigl[\log F(x_i) + \log\bigl(1-F(x_{n+1-i})\bigr)\bigr],$$

where x_i must be sorted, $x_i \le x_{i+1}$. The values of the statistic in this and other tests is not hard to calculate, but two essential questions always remain: how is this statistic distributed and how can we compute the critical values for a given distribution function F. For the AD test, critical values are available for the uniform, normal, log-normal, exponential and Pareto distributions [10–14].

A final warning: the KS test is strictly applicable only if the model distribution F being compared to the empirical distribution $\widetilde{F}_n$ is independent of the observations. This means that the sample being tested should not be used to estimate the parameters of F, say, its expected value or variance. In this case the test works, but both the cumulative distributions (10.10) and the critical values are modified. A possible solution of this problem is the so-called *bootstrap resampling*: the measured sample is used to generate a large set of new samples, and the test is performed with the entire ensemble. Introductory reading on bootstrap methods is offered by [15].

10.5 Problems

10.5.1 Test of Mean Decay Time

(Adapted from [16].) By a well-established theory (hypothesis H_0) a certain quantum-mechanical state should decay with decay time τ_0, while according to a competing theory (hypothesis H_1) it should have decay time τ_1,

$$H_0 : \tau = \tau_0 = 1\,\mathrm{ns}, \qquad H_1 : \tau = \tau_1 = 2\,\mathrm{ns}.$$

The actually observed decay times are $\boldsymbol{t} = \{t_1, t_2, \ldots, t_n\}$ and their average is $\overline{t} = (1/n)\sum_{i=1}^{n} t_i$. Find the region where the power of rejecting the null hypothesis with variable $\overline{t}$ at $\alpha = 0.05$ is largest, for ① $n = 1$ and ② $n \gg 1$! Assume that the density has the form $f(t;\tau) = \tau^{-1}\exp(-t/\tau)$ for both hypotheses.

✎ The power of the test of H_0 against H_1 is large where the ratio (10.3) is larger than some constant,

$$\frac{L(\boldsymbol{t};\tau=\tau_1)}{L(\boldsymbol{t};\tau=\tau_0)}=\frac{\prod_{i=1}^n \tau_1^{-1}\mathrm{e}^{-t_i/\tau_1}}{\prod_{i=1}^n \tau_0^{-1}\mathrm{e}^{-t_i/\tau_0}}=\left(\frac{\tau_0}{\tau_1}\right)^n \exp\left(-\left(\frac{1}{\tau_1}-\frac{1}{\tau_0}\right)\sum_{i=1}^n t_i\right)>C$$

or

$$\bar{t}>\frac{\tau_0\tau_1}{\tau_1-\tau_0}\left(\frac{1}{n}\log C+\log\frac{\tau_1}{\tau_0}\right)=T_n.$$

The best rejection region is the interval of values $\bar{t}$ satisfying this inequality, where T_n is a constant depending on α. The statistic $\overline{T}$ (with values $\bar{t}$) is thus an appropriate statistic to test the true mean τ, but we must also know *its own* distribution.

① In the case $n=1$ the probability density for $\overline{T}$ is simply

$$f_1(\bar{t};\tau)=\tau^{-1}\exp(-\bar{t}/\tau),$$

so by (10.1) the rejection region is defined as

$$\alpha=\int_{T_1}^{\infty}\frac{1}{\tau_0}\,\mathrm{e}^{-\bar{t}/\tau_0}\,\mathrm{d}\bar{t}=\mathrm{e}^{-T_1/\tau_0}.$$

The critical value is $T_1/\tau_0=-\log\alpha\approx 3.00$, and the power of the test of H_0 against H_1 is

$$1-\beta=\int_{T_1}^{\infty}\frac{1}{\tau_1}\,\mathrm{e}^{-\bar{t}/\tau_1}\,\mathrm{d}\bar{t}=\mathrm{e}^{-T_1/\tau_1}=\alpha^{\tau_0/\tau_1}\approx 0.224.$$

Therefore, by a single observation, $\bar{t}=t_1$, one may reject H_0 if $t_1>T_1\approx 3\tau_0$. Conversely, the probability that H_0 is accepted when actually H_1 is true, is $\beta\approx 0.776$.

② In the case $n\gg 1$ the distribution of the test statistic $\overline{T}$ can be approximated by the normal distribution with average τ and variance τ^2/n,

$$f_n\big(\bar{t};\tau\big)=\frac{1}{\sqrt{2\pi}\,\tau/\sqrt{n}}\exp\left(-\frac{1}{2}\frac{(\bar{t}-\tau)^2}{\tau^2/n}\right).$$

The critical value T_n of the rejection region for $\bar{t}$ at given α is then defined by

$$\alpha=\int_{T_n}^{\infty}f_n\big(\bar{t};\tau_0\big)\,\mathrm{d}\bar{t}=1-\int_{-\infty}^{T_n}f_n\big(\bar{t};\tau_0\big)\,\mathrm{d}\bar{t}=1-\Phi\left(\frac{T_n-\tau_0}{\tau_0/\sqrt{n}}\right),$$

where Φ is the distribution function of the standardized normal distribution. At chosen α this means $(T_n-\tau_0)/(\tau_0/\sqrt{n})=z_*=z_{0.95}\approx 1.645$ (see Table D.1) or

$$T_n=\tau_0\left(1+\frac{z_*}{\sqrt{n}}\right),$$

and the power of the test of H_0 against H_1 is

$$1-\beta = \int_{T_n}^{\infty} f_n\big(\bar{t};\tau_1\big)\,\mathrm{d}\bar{t} = 1 - \int_{-\infty}^{T_n} f_n\big(\bar{t};\tau_1\big)\,\mathrm{d}\bar{t} = 1 - \Phi\left(\frac{T_n-\tau_1}{\tau_1/\sqrt{n}}\right).$$

Increasing the sample size dramatically increases the power of the test: with $n = 100$, for instance, we get $T_{100} = 1.1645\,\mathrm{ns}$ and $1-\beta = 0.99999$.

10.5.2 Pearson's Test for Two Histogrammed Samples

The same quantity is measured in two laboratories. We have obtained a sample of $m = 100$ observations from laboratory A and $n = 200$ data points from laboratory B, classified in a histogram with $N = 10$ bins shown in Fig. 10.9. The numbers of counts in each bin are represented by the samples

$$\begin{aligned}\boldsymbol{x}\ (\text{lab A}) &= \{1,\ 3,\ 6,\ 14,\ 27,\ 20,\ 14,\ 10,\ 2,\ 3\},\\ \boldsymbol{y}\ (\text{lab B}) &= \{3,\ 2,\ 6,\ 20,\ 45,\ 63,\ 31,\ 24,\ 5,\ 1\}.\end{aligned}$$

Are these two samples mutually consistent from the perspective of the Pearson's χ^2 test with statistical significance $\alpha = 0.10$?

✎ The statistic (10.6) is χ^2-distributed, with $\nu = N - 1 = 9$ degrees of freedom. At chosen $\alpha = 0.10$ we need its 90. percentile, available in Table D.3: $\chi^2_{0.90}(\nu)/\nu = 1.63$. By using (10.6) with the observed data we get $x^2/\nu = 1.27$. Since $x^2/\nu < \chi^2_{0.90}(\nu)/\nu$, we can not reject the hypothesis that the observed distributions are consistent. What if we are a bit less demanding and assume $\alpha = 0.25$? In this case we obtain $\chi^2_{0.75}/\nu = 1.26$, which is a tad below the observed x^2/ν: at confidence level $1-\alpha = 0.75$ we can just claim that the distributions are mutually inconsistent.

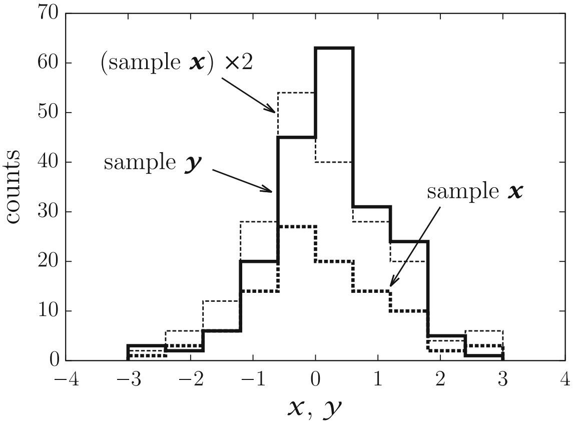


Fig. 10.9 Histograms of $m = 100$ and $n = 200$ observations obtained in two laboratories, compared to each other by the Pearson's χ^2 test

10.5.3 Flu Medicine

A classical pharmaceutical problem is the test of a drug intended to shorten the duration of flu symptoms. In untreated patients their average duration is $\mu_0 = 7$ days, with standard deviation $\sigma = 1.5$ days. The medicine is given to $n = 100$ random patients when they develop first symptoms. In this sample the average symptom duration is $\overline{x} = 5.5$ days. Is this outcome statistically significant at significance $\alpha = 0.01$?

✎ We are testing the null hypothesis $H_0 : \mu = \mu_0$ (medicine has no statistically significant effect) against $H_1 : \mu < \mu_0$ (medicine shortens the duration of symptoms). Assume that the distribution of $\overline{X}$ is normal; by (7.6) and (7.7) it can be assigned the mean μ_0 and variance σ^2/n. Since the alternative hypothesis has the form $\mu < \mu_0$, the rejection region is at the lower end of the real axis, strictly speaking $(-\infty, x_*]$, but practically $[0, x_*]$, as the duration of symptoms can not be negative. We therefore seek x_* such that $P\big(\overline{X} \le x_*\big) = \alpha$ if H_0 is true. In terms of the standardized variable $Z = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$ this means

$$P\left(Z \le \frac{x_* - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{x_* - \mu_0}{\sigma/\sqrt{n}}\right) = \alpha,$$

where Φ is the distribution function of the standardized normal distribution. Hence

$$\frac{x_* - \mu_0}{\sigma/\sqrt{n}} = \Phi^{-1}(\alpha) = \sqrt{2}\,\mathrm{erf}^{-1}\big(2\alpha - 1\big) \approx -2.33 \quad\Longrightarrow\quad x_* \approx 6.65\,\text{days}.$$

Because $\overline{x} < x_*$, the hypothesis H_0 (i.e. that a relatively small average duration has only been observed by chance) can be rejected. The efficacy of the medicine is indeed statistically significant.

10.5.4 Exam Grades

The eternal professor (not student) question: are exam grades normally distributed? ① The grades of a small sample of $n = 10$ students (in %) are $\boldsymbol{x} = \{14, 20, 22, 48, 55, 57, 63, 74, 88, 97\}$. Is this result compatible with a normal distribution with mean $\mu_0 = 50$ and standard deviation $\sigma_0 = 30$, at significance $\alpha = 0.10$? ② From a large population we take a sample of $m = 6$ exams of students majoring in A (grades $\boldsymbol{x} = \{24, 33, 56, 65, 77, 94\}$) and $n = 8$ exams of students majoring in B (grades $\boldsymbol{y} = \{27, 30, 34, 55, 69, 73, 88, 93\}$). Is there a statistically significant difference between the A and B students, at $\alpha = 0.01$?

✎ We use the Kolmogorov–Smirnov test. ① The null hypothesis is $H_0 : \widetilde{F}_n(x) - F(x)$, where F is the distribution function of $N(\mu_0, \sigma_0^2)$. The maximum distance between $\widetilde{F}_n$ and F is $d_n = 0.173$. The exact critical value is $d_*(n; \alpha) = 0.369$, while the asymptotic formula (10.9) gives 0.387. Since $d_n < d_*$, H_0 can not be rejected, which speaks in favor of the normal distribution of grades.

② The null hypothesis is that the samples stem from the same population, thus within statistical fluctuations their empirical distribution functions should also be the same, $H_0 : \widetilde{F}_m(x) = \widetilde{F}_n(x)$. The maximum distance between $\widetilde{F}_m$ and $\widetilde{F}_n$ is $d_{mn} = 0.167$. The exact (tabulated) critical value is $d_*(6, 8; 0.01) = 0.8$, while the asymptotic formula (10.13) gives 0.879. Since $d_{mn} < d_*$, H_0 can not be rejected.

References

1. W.J. Conover, *Practical Nonparametric Statistics*, 3rd edn. (Wiley, New York, 1999)
2. F. James, *Statistical Methods in Experimental Physics*, 2nd edn. (World Scientific, Singapore, 2006)
3. A. Kolmogorov, *Sulla determinazione empirica di una legge di distribuzione*, Giornalo dell'Istituto Italiano degli Attuari **4**, 461 (1933). Translated in A.N. Shiryayev (Ed.), *Selected works of A.N. Kolmogorov*, Vol. II (Springer Science+Business Media, Dordrecht, 1992) p. 139
4. N. Smirnov, Sur les écarts de la courbe de distribution empirique. Rec. Math. **6**, 3 (1939)
5. S. Facchinetti, A procedure to find exact critical values of Kolmogorov–Smirnov test, Statistica Applicata—Ital. J. Appl. Stat. **21**, 337 (2009)
6. J.W. Pratt, J.D. Gibbons, *Concepts of Nonparametric Theory* (Springer, New York, 1981)
7. M.A. Stephens, Tests based on EDF statistics, in *Goodness of Fit Techniques*, ed. by R.B. D'Agostino, M.A. Stephens (Marcel Dekker, New York, 1986), pp. 97–194
8. T.W. Anderson, D.A. Darling, Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann. Math. Statist. **23**, 193 (1952)
9. T.W. Anderson, D.A. Darling, A test of goodness of fit. J. Am. Stat. Assoc. **49**, 765 (1954)
10. M.A. Stephens, EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. **69**, 730 (1974)
11. M.A. Stephens, Asymptotic results for goodness-of-fit statistics with unknown parameters. Annals Stat. **4**, 357 (1976)
12. M.A. Stephens, Goodness of fit for the extreme value distribution. Biometrika **64**, 583 (1977)
13. M.A. Stephens, Goodness of fit with special reference to tests for exponentiality, Technical Report No. 262, Department of Statistics, Stanford University, Stanford, 1977
14. M.A. Stephens, Tests of fit for the logistic distribution based on the empirical distribution function. Biometrika **66**, 591 (1979)
15. A.C. Davison, D.V. Hinkley, *Bootstrap Methods and their Application* (Cambridge University Press, Cambridge, 1997)
16. A.G. Frodesen, O. Skjeggestad, H. Tøfte, *Probability and statistics in particle physics* (Universitetsforlaget, Bergen, 1979)

Part III
Special Applications of Probability

Chapter 11
Entropy and Information ⋆

Abstract Entropy is introduced as a concept that quantifies the amount of information contained in a signal or in its corresponding probability distribution. It is defined for discrete and continuous distributions, along with its relative counterpart, the Kullback-Leibler divergence that measures the "distance" between two distributions. The principle of maximum entropy is stated, paving the way to the derivation of several discrete maximum-entropy distributions by means of Lagrange multiplier formalism: the Maxwell-Boltzmann, Bose-Einstein and Fermi-Dirac distributions. The relation between information and thermodynamic entropy is elucidated. A brief discussion of continuous maximum-entropy distributions is followed by presenting the method of maximum-entropy spectral analysis.

11.1 Measures of Information and Entropy

One of the possible paths to the definition of entropy leads through the concept of *information*. This connection can be elucidated by studying a discrete random variable X that can take finitely many values $\{x_1, x_2, \ldots, x_n\}$ with probabilities $\{p_1, p_2, \ldots, p_n\}$, where $p_i = P(X = x_i)$ and $\sum_{i=1}^n p_i = 1$. Imagine that each outcome of the experiment with this variable, say, the event $(X = x)$ occurring with probability $p = P(X = x)$, brings some information $I(p)$. The value x can be seen as a "signal" or "message" carrying information $I(p)$.

How can its quantity be measured? Intuitively it is clear that any measure of information must have logarithmic nature [1, 2]. If events with probabilities p_1 and p_2 occur independently (probability $p_1 p_2$), the information of such a combined outcome should equal the information supplied by single outcomes: the sentences "it snows" and "it is Friday" together carry as much information as "it snows and it is Friday". Hence, a measure of information should be additive,

$$I(p_1 p_2) = I(p_1) + I(p_2).$$

Besides, we wish the function $I(p)$ to be non-negative, $I(p) \geq 0$, monotonous, $p_1 < p_2 \Longrightarrow I(p_1) > I(p_2)$, and continuous: small changes in p imply small changes in $I(p)$. An obvious candidate is the function

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_11

$$I(p) = -C \log_b p$$

and in fact it can be shown that it is the only possible [3]. A measure defined in this way has the sensible properties $I(1) = 0$ (a certain event carries no information) and $\lim_{p \to 0} I(p) = \infty$ (a highly improbable event brings lots of information). The arbitrary real constant C can be hidden in the base of the logarithm by using $\log_b x = \log x / \log b$, and is therefore irrelevant. If we adopt $b = 2$ and $C = 1$, information is measured in *bits*. If we choose $b = \mathrm{e}$ and $C = 1$, it is measured in *nats*, differing from bits only by the factor $\log 2$.

There is only one step from information to information entropy. If individual values x_i occur with probabilities p_i, $i = 1, 2, \ldots, n$, the average quantity of received or “created” information is

$$H(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i I(p_i) = -\sum_{i=1}^{n} p_i \log p_i. \tag{11.1}$$

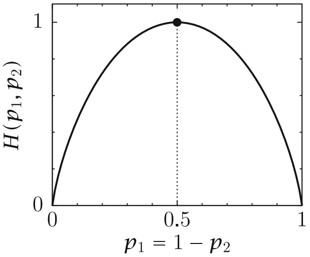
This “weighting scale” of information is called the entropy of a finite probability distribution due to Shannon [3]. How can it be interpreted?

The essence of any random process is *uncertainty*. The outcomes are not predictable, but each received signal (a single value x_i) reduces the uncertainty we had prior to receiving it. The expression (11.1) can therefore be understood as a measure of such uncertainty. We must realize that H measures information entropy that should not be confused with the thermodynamic entropy S. In the following ‘entropy’ means information entropy.[1]

The measure (11.1) has many convenient properties. The uncertainty of a certain event is zero, $H(p = 1) = 0$. The uncertainty of an impossible event is also zero, $H(p = 0) = 0$, as nothing is unclear in an event that never occurs, besides, formally $\lim_{p \to 0} p \log p = 0$. The value of the entropy depends only on the probability distribution $\{p_i\}$ and no other properties that might be assigned to the signal. It is independent of the permutations among p_i and does not change if n events are augmented by an impossible event, $H(p_1, p_2, \ldots, p_n) = H(p_1, p_2, \ldots, p_n, 0)$. Entropy is maximal when we are “maximally uncertain”, i.e. when all outcomes are equally likely: $p_1 = p_2 = \cdots = p_n = 1/n$ (uniform distribution). For any other distribution or under any condition imposed on p_i the entropy decreases (see Example on p. 288 and [5]).

[1] Shannon had second thoughts on introducing the concept of entropy to information theory. It is said that his decision was stimulated by the mathematician John Neumann who said [4]: “Firstly, you have got the same expression $-\sum_i p_i \log p_i$ as is used for entropy in thermodynamics and, secondly and more importantly, since even after one hundred years, nobody understands what entropy is, if you use the word entropy you will always win in an argument!” See also Sect. 11.3.4.

Example In tossing a fair coin heads and tails are equally probable: $p_1 = p_2 = \frac{1}{2}$.



The entropy of the random variable with such distribution is

$$H = -\left(\tfrac{1}{2}\,\log_2 \tfrac{1}{2} + \tfrac{1}{2}\,\log_2 \tfrac{1}{2}\right) = 1.$$

Hence, tossing a coin supplies one bit of information on average. If the coin is unfair such that, for instance, $p_1 = \frac{9}{20}$ and $p_2 = \frac{11}{20}$, we get

$$H = -\left(\tfrac{9}{20}\,\log_2 \tfrac{9}{20} + \tfrac{11}{20}\,\log_2 \tfrac{11}{20}\right) \approx 0.9928 < 1.$$

The entropy has decreased as the coin prefers a specific side, reducing the uncertainty. The dependence of H on $p_1 = 1 - p_2$ is shown in the above figure. ◁

Example Throwing a die has the probability distribution $p_i = \frac{1}{6}$, $1 \le i \le 6$. (We "know" that. How exactly this follows from the principle of maximum entropy is discussed in Sect. 11.2.) The entropy of this uniform distribution is

$$H = -6\tfrac{1}{6}\log_2 \tfrac{1}{6} = \log_2 6 \approx 2.585.$$

Therefore, throwing a die yields about 2.585 bits of information on average. But if someone tells us, say, that the number of dots is odd, the sample space shrinks since there are only three possible outcomes, hence $p_1 = p_3 = p_5 = \frac{1}{3}$ and

$$H = -3\tfrac{1}{3}\log_2 \tfrac{1}{3} = \log_2 3 \approx 1.585.$$

The entropy has diminished as three outcomes instead of six imply less "uncertainty", less "indefiniteness". The restriction to odd number of points on average means $\log_2 6 - \log_2 3 = \log_2 2 = 1$ bit of acquired information. ◁

11.1.1 Entropy of Infinite Discrete Probability Distribution

If the partial sums $\sum_{i=1}^{n} p_i \log p_i$ converge when $n \to \infty$, then

$$H(p_1, p_2, \ldots) = -\sum_{i=1}^{\infty} p_i \log p_i \qquad (11.2)$$

represents the entropy of an infinite discrete probability distribution.

Double example For the geometric distribution $p_i = 1/2^i$ $(i = 1, 2, \ldots)$ the sum (11.2) is not difficult to calculate: $-\sum_{i=1}^{\infty} p_i \log p_i = -\log 2 \sum_{i=1}^{\infty} p_i \log_2 p_i = \log 2 \sum_{i=1}^{\infty} i/2^i = 2 \log 2$. So its entropy is 2 bits.

With $p_i = 1/(a\, i \log^2 i)$, where $i = 2, 3, \ldots$ and $a = \sum_{k=2}^{\infty} 1/(k \log^2 k)$, we are not that fortunate: even though $a \approx 2.10974 < \infty$ and the distribution is normalized, $\sum_{i=2}^{\infty} p_i = 1$, we realize that $-\sum_{i=1}^{\infty} p_i \log p_i = \infty$. The entropy of such a distribution does not exist (or we say that it has infinite entropy). ◁

11.1.2 Entropy of a Continuous Probability Distribution

The entropy of a continuous probability distribution is defined by analogy to the discrete formulation (11.1). If X is a continuous random variable with the probability density f, the entropy of its distribution is

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, \mathrm{d}x. \tag{11.3}$$

Note that we are using the notation $H(X)$ instead of, say, $H(f)$. Information and entropy may be assigned both to the random variable itself or to its probability distribution. There is no general consensus about that, so we will adopt the notations $I(X)$ and $I(p_1, p_2, \ldots)$ as well as $H(X)$ and $H(p_1, p_2, \ldots)$ in the discrete case—just as $H(X)$ and $H(f)$ in the continuous case—as equivalent.

Example The entropy of the uniform distribution $U(a, b)$ or the uniformly distributed continuous random variable $X \sim U(a, b)$ is

$$H(X) = -\frac{1}{b-a} \log \frac{1}{b-a} \int_a^b \mathrm{d}x = \log(b-a). \tag{11.4}$$

The result clearly depends only on the *difference* $b - a$. This means that all uniform distributions with the same spacing have the same entropy. ◁

Example The distribution with the probability density

$$f(x) = \begin{cases} 0 & ;\ x < \mathrm{e}, \\ 1/(x \log^2 x) & ;\ x \geq \mathrm{e}, \end{cases}$$

is normalized, $\int_{-\infty}^{\infty} f(x)\, \mathrm{d}x = 1$, but $-\int_{-\infty}^{\infty} f(x) \log f(x)\, \mathrm{d}x = \infty$. Its entropy is infinite (or: does not exist). ◁

11.1.3 Kullback–Leibler Distance

Imagine a time series (signal, sequence) with values distributed according to a discrete probability distribution $\boldsymbol{p} = \{p_1, p_2, \ldots\}$. The recorded signal contains a certain information. We measure another sequence with values corresponding to the distribution $\boldsymbol{q} = \{q_1, q_2, \ldots\}$. Has the new sample brought any additional information with respect to the original data set? In other words, how "distant" to each other—in the entropy sense—are the two distributions? A measure of this "remoteness" is the Kullback-Leibler distance or *divergence*, sometimes also called *relative entropy* [6]. For discrete distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ it is defined as

$$D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) = \sum_i p_i \log \frac{p_i}{q_i},$$

while for continuous distributions with densities p and q it is formulated as

$$D_{\mathrm{KL}}(p\|q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\,\mathrm{d}x.$$

The concept of distance refers to the property $D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) \geq 0$, where the equality sign applies precisely when $\boldsymbol{p} = \boldsymbol{q}$. Moreover, $D_{\mathrm{KL}}(\boldsymbol{p}\|\boldsymbol{q}) \neq D_{\mathrm{KL}}(\boldsymbol{q}\|\boldsymbol{p})$.

Example A seismological study [7] resulted in the distribution of vertical components of seismic velocities. The signals (348 time series lasting 68 hours each) were frequency-filtered, so that they corresponded to oscillations with periods (5–10) s and (10–20) s on the time scale. The obtained distributions $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ shown in Fig. 11.1 (left) were fitted by normal distributions $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$. For all 348 samples,

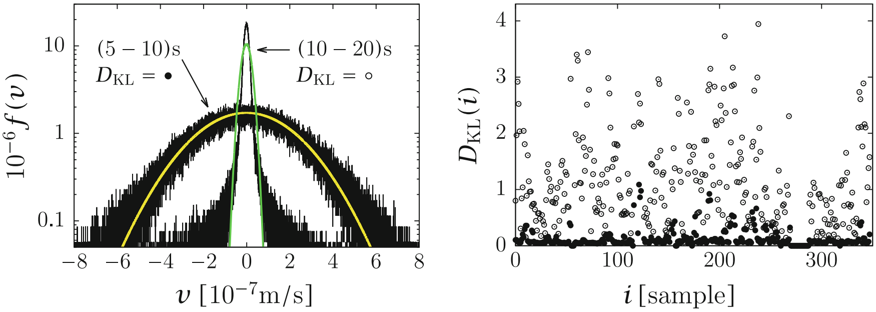


Fig. 11.1 [Left] Distributions of vertical seismic velocities (histograms) and the corresponding normal distributions as best fits to the data (*brighter curves*). [Right] The Kullback-Leibler distance between the measured distributions and the normal distributions from the *left panel*, $D_{\mathrm{KL}}(\boldsymbol{p}_1\|\boldsymbol{q}_1) = \bullet$, $D_{\mathrm{KL}}(\boldsymbol{p}_2\|\boldsymbol{q}_2) = \circ$

relative entropies $D_{\mathrm{KL}}(\boldsymbol{p}_1 \| \boldsymbol{q}_1)$ and $D_{\mathrm{KL}}(\boldsymbol{p}_2 \| \boldsymbol{q}_2)$ were calculated. As shown in Fig. 11.1 (right), the "distance" of the second distribution from the normal is much larger, i.e. it is "much less Gaussian". ◁

11.2 Principle of Maximum Entropy

We have seen that in a random process the "uncertainty"— i.e. its information entropy—is largest when all of its outcomes are equally probable. This realization has been succinctly expressed already by Laplace in his *principle of insufficient reason* or *principle of indifference*: in the absence of a specific reason to distinguish between two or more outcomes, the best strategy is to treat them as equally probable.

The *principle of maximum entropy* builds upon and upgrades this guideline: in any circumstance where incomplete information is available—for instance, in a sample of observations—we strive to quantitatively describe the data by a probability distribution that is consistent with all *known* information, yet at the same time as "nonrestrictive" as possible, "uncertain", "free" with respect to the *unknown* information.[2] Laplace's argument offers only the negative lever-arm "in the absence of ...", while the principle of maximum entropy offers clearly defined, positive tool in the sense of determining the distribution that is "as nonrestrictive as possible". It is precisely this aspect that removes the flavor of arbitrariness from Laplace's principle [8, 9].

We are using the words like "lever-arm", "aspect", "flavor"—all loose, non-mathematical concepts! The principle of maximum entropy can not be strictly "proven", yet *de facto* practically all known probability distributions follow from it in a very natural manner. In contrast to Laplace, it has the important property that each outcome not absolutely excluded by a known piece of information is assigned a non-zero contribution. Initial reading on maximum entropy is offered by the review article [10]; for a very mathematically tinted discussion see [11].

Example According to Laplace, throwing a fair die corresponds to the uniform distribution $p_i = 1/6$, $i = 1, 2, \ldots, 6$. One can reach the same conclusion by invoking the principle of maximum entropy. We maximize $-\sum_{i=1}^{6} p_i \log p_i$ with the condition $\sum_{i=1}^{6} p_i = 1$. This can be done by the classical method of Lagrange multipliers. We calculate the first derivative of the Lagrange function

$$\mathcal{L} = -\sum\nolimits_{i=1}^{6} p_i \log p_i - \lambda \left(\sum\nolimits_{i=1}^{6} p_i - 1 \right)$$

with respect to p_i and set it to zero. It follows that $\partial\mathcal{L}/\partial p_i = -\log p_i - 1 - \lambda = 0$ and $p_i = \mathrm{e}^{-\lambda-1}$. Hence p_i do not depend on i (that is, they are all equal) and their

[2] As shown below, the simplest example is the uniform distribution: if no additional condition is imposed on the distribution apart from normalization, the distribution with the maximum entropy is precisely the uniform distribution. (This applies in both the discrete and continuous cases.)

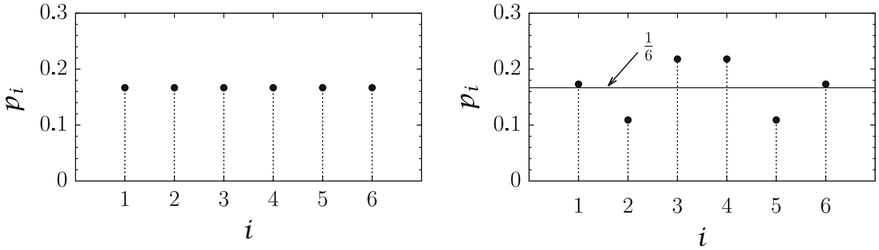


Fig. 11.2 Discrete distribution of values in throwing a fair die and the principle of maximum entropy. [Left] Normalized distribution with no restrictions (constraints). [Right] Normalized distribution with constraints (11.11) and (11.12)

sum must be 1. Therefore $p_1 = p_2 = \cdots = p_6 = 1/6$, see Fig. 11.2 (left). In the following this Example will be expanded into a more general tool based on Lagrange multipliers. ◁

11.3 Discrete Distributions with Maximum Entropy

11.3.1 Lagrange Formalism for Discrete Distributions

On any discrete probability distribution we may impose additional conditions or *constraints*. The constraint we indeed *must* impose is the normalization requirement $\sum_{i=1}^{n} p_i = 1$. But we may also include conditions of the form

$$\sum_{i=1}^{n} p_i f_j(x_i) = \overline{f}_j, \qquad j = 1, 2, \ldots, m. \tag{11.5}$$

An example of such constraint is the requirement that the average number of dots in throwing a die is 4, which is expressed as $\sum_{i=1}^{6} i p_i = 4$. The distribution that maximizes the entropy at given constraints can be calculated by the general method of Lagrange multipliers. To each of the $m + 1$ constraints (normalization plus m conditions of the form (11.5)) we assign its multiplier and minimize

$$\mathcal{L} = -\sum_{i=1}^{n} p_i \log p_i - (\lambda_0 - 1)\left(\sum_{i=1}^{n} p_i - 1\right) - \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{n} p_i f_j(x_i) - \overline{f}_j\right). \tag{11.6}$$

In the second term we have subtracted 1 from λ_0 in order to cancel the 1 from the derivative of the first sum. No harm has been done: λ_0 is just as good a multiplier as $\lambda_0 - 1$. We calculate the derivative of $\mathcal{L}$ with respect to p_i and set it to zero:

$$\frac{\partial \mathcal{L}}{\partial p_i} = -\log p_i - 1 - \lambda_0 + 1 - \sum_{j=1}^{m} \lambda_j f_j(x_i) = 0. \tag{11.7}$$

It follows that

$$p_i = \exp\left(-\lambda_0 - \sum_{j=1}^{m} \lambda_j f_j(x_i) \right). \tag{11.8}$$

We insert this in $\sum_{i=1}^{n} p_i = 1$ and extract from it the factor $\exp(\lambda_0)$ called the *phase sum* or the *partition function*:

$$Z = \mathrm{e}^{\lambda_0} = \sum_{i=1}^{n} \exp\left(-\sum_{j=1}^{m} \lambda_j f_j(x_i) \right). \tag{11.9}$$

We take the logarithm of the phase sum,

$$\log Z = \lambda_0 = \log\left[\sum_{i=1}^{n} \exp\left(-\sum_{j=1}^{m} \lambda_j f_j(x_i) \right) \right],$$

calculate its derivative with respect to λ_j and, finally, multiply the expressions in the numerator and denominator of the obtained fraction by $\mathrm{e}^{-\lambda_0}$:

$$\frac{\partial(\log Z)}{\partial \lambda_j} = \frac{-\sum_{i=1}^{n} f_j(x_i) \exp\left(-\lambda_0 - \sum_{k=1}^{m} \lambda_k f_k(x_i)\right)}{\sum_{i=1}^{n} \exp\left(-\lambda_0 - \sum_{k=1}^{m} \lambda_k f_k(x_i)\right)} = \frac{-\sum_{i=1}^{n} f_j(x_i) p_i}{\sum_{i=1}^{n} p_i} = -\overline{f}_j.$$

This is a system of m equations for m unknowns $\lambda_j, j = 1, 2, \ldots, m$ that needs to be solved for better or worse. The calculated multipliers yield the final formula for individual probabilities corresponding to the maximum-entropy distribution:

$$p_i = \frac{1}{Z} \exp\left(-\sum_{j=1}^{m} \lambda_j f_j(x_i) \right), \qquad i = 1, 2, \ldots, n. \tag{11.10}$$

Example A die has been tweaked such that the probability of obtaining three dots is twice the probability of getting two, and the probability of observing four dots is twice the probability of finding five. Calculate, with these restrictions, the probability distribution of the number of dots $\boldsymbol{p} = \{p_1, p_2, p_3, p_4, p_5, p_6\}$ that is consistent with the maximum-entropy assumption!

Normalization and the two specified conditions introduce the constraints

$$p_1+p_2+p_3+p_4+p_5+p_6-1=0,$$
$$-2p_2+p_3=0, \tag{11.11}$$
$$p_4-2p_5=0. \tag{11.12}$$

We see that $f_1(x_2)=f_2(x_5)=-2$, $f_1(x_3)=f_2(x_4)=1$ and $\overline{f}_1=\overline{f}_2=0$. The appropriate Lagrange function is

$$\mathcal{L}=-\sum\nolimits_{i=1}^6 p_i\log p_i-(\lambda_0-1)\left(\sum\nolimits_{i=1}^6 p_i-1\right)-\lambda_1\bigl(-2p_2+p_3\bigr)-\lambda_2\bigl(p_4-2p_5\bigr),$$

but there is no need to calculate its derivative (11.7) again, since (11.9) immediately gives us the partition function

$$Z=2+\mathrm{e}^{2\lambda_1}+\mathrm{e}^{-\lambda_1}+\mathrm{e}^{-\lambda_2}+\mathrm{e}^{2\lambda_2}, \tag{11.13}$$

and thence a system of two equations for λ_1 and λ_2:

$$\frac{\partial(\log Z)}{\partial\lambda_1}=\frac{1}{Z}\left(2\mathrm{e}^{2\lambda_1}-\mathrm{e}^{-\lambda_1}\right)=0,\qquad \frac{\partial(\log Z)}{\partial\lambda_2}=\frac{1}{Z}\left(-\mathrm{e}^{-\lambda_2}+2\mathrm{e}^{2\lambda_2}\right)=0.$$

Its solution is $\lambda_1=\lambda_2=-\frac{1}{3}\log 2$, so that (11.13) yields $Z\approx 5.77976$. The final result then follows from (11.10):

$$p_1=p_6\approx 0.1730,\quad p_2=p_5\approx 0.1090,\quad p_3=p_4\approx 0.2180.$$

See Fig. 11.2 (right) and think: we certainly have not anticipated the distribution $p_i=1/6\,(1\le i\le 6)$ after all this rattle; but why is the answer not simply $p_1=p_2=p_3/2=p_4/2=p_5=p_6=1/8$? Why did the probabilities p_1 and p_6 change from their "Laplacian" values of $1/6$ even though the constraints (11.11) and (11.12) do not address them at all? ◁

11.3.2 *Distribution with Prescribed Mean and Maximum Entropy*

Among all finite discrete distributions with probabilities p_i $(1\le i\le n)$ and prescribed arithmetic mean μ $(1\le\mu\le n)$ the one with the maximum entropy is the power distribution. One can see that if one maximizes the entropy $-\sum_{i=1}^n p_i\log p_i$ with the constraints

$$\sum_{i=1}^n p_i=1,\qquad \sum_{i=1}^n ip_i=\mu, \tag{11.14}$$

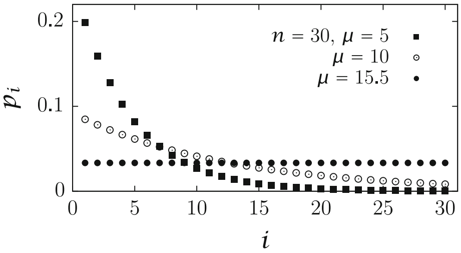


Fig. 11.3 Finite discrete distributions ($n = 30$) with prescribed arithmetic means $\mu = 5$, 10 and 15.5 and maximum entropy. In the last case the power distribution has degenerated into the uniform distribution

i.e. the Lagrange function

$$\mathcal{L} = -\sum_{i=1}^{n} p_i \log p_i - (\lambda_0 - 1)\left(\sum_{i=1}^{n} p_i - 1\right) - \lambda_1\left(\sum_{i=1}^{n} i p_i - \mu\right).$$

From $\partial\mathcal{L}/\partial p_i = 0$ it follows that $p_i = \mathrm{e}^{-\lambda_0}\mathrm{e}^{-\lambda_1 i} \equiv ab^i$, where $i = 1, 2, \ldots, n$. When this is inserted in (11.14), two equations for the unknowns a and b follow:

$$a\sum_{i=1}^{n} b^i = ab\frac{1-b^n}{1-b} = 1, \qquad a\sum_{i=1}^{n} ib^i = ab\left[\frac{1-b^n}{(1-b)^2} - \frac{nb^n}{1-b}\right] = \mu.$$

The system is solved numerically. Taking $n = 30$ and $\mu = 5$, 10, 15.5 yields $(a, b) \approx (0.2479, 0.8016)$, $(0.09181, 0.9229)$, $(0.03333, 1.0000)$, respectively. The obtained power distributions with calculated parameters are shown in Fig. 11.3.

11.3.3 Maxwell–Boltzmann Distribution

Assume that a physical system possesses energy levels with single-particle energies $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ that particles occupy with probabilities $p_1, p_2, \ldots, p_n$. Let the expected value of the energy, ε, be prescribed. What is the probability distribution of particles that is consistent with the assumption of maximum entropy? We must maximize the entropy with the constraints

$$\sum_{i=1}^{n} p_i = 1, \qquad \sum_{i=1}^{n} p_i\varepsilon_i = \varepsilon, \tag{11.15}$$

that is, the Lagrange function

$$\mathcal{L} = -\sum\nolimits_{i=1}^{n} p_i \log p_i - (\lambda_0 - 1)\left(\sum\nolimits_{i=1}^{n} p_i - 1\right) - \lambda_1\left(\sum\nolimits_{i=1}^{n} p_i\varepsilon_i - \varepsilon\right).$$

This means $\partial\mathcal{L}/\partial p_i = -\log p_i - \lambda_0 - \lambda_1\varepsilon_i = 0$ or $p_i = \mathrm{e}^{-\lambda_0}\mathrm{e}^{-\lambda_1\varepsilon_i} \equiv a\,\mathrm{e}^{-\lambda_1\varepsilon_i}$, where $i = 1, 2, \ldots, n$. When this is inserted in (11.15), it follows that

$$p_i = \frac{\mathrm{e}^{-\lambda_1\varepsilon_i}}{\sum_{i=1}^{n}\mathrm{e}^{-\lambda_1\varepsilon_i}}, \qquad \varepsilon = \frac{\sum_{i=1}^{n}\varepsilon_i\,\mathrm{e}^{-\lambda_1\varepsilon_i}}{\sum_{i=1}^{n}\mathrm{e}^{-\lambda_1\varepsilon_i}}. \tag{11.16}$$

If we set $1/\lambda_1 = k_\mathrm{B}T$, where k_B is the Boltzmann constant, these expressions specify the Maxwell-Boltzmann distribution (see also Sect. 11.3.4).

Example We discuss a system with three ($n = 3$) discrete energy levels ε_1, $\varepsilon_2 = 4\varepsilon_1$ and $\varepsilon_3 = 9\varepsilon_1$, shown in Fig. 11.4 (left). We are interested in the level occupation probabilities p_i at three inverse values of the λ_1 parameter, say, $1/\lambda_1 = \varepsilon_1$, $3\,\varepsilon_1$ and $10\,\varepsilon_1$. By using (11.16) we obtain

$$\begin{aligned} 1/\lambda_1 = \varepsilon_1 : \{p_1, p_2, p_3\} &\approx \{0.9523,\ 0.0474,\ 0.0003\},\ \varepsilon \approx 1.145\,\varepsilon_1, \\ 3\,\varepsilon_1 : \qquad &\ \{0.6957,\ 0.2559,\ 0.0483\},\qquad 2.155\,\varepsilon_1, \\ 10\,\varepsilon_1 : \qquad &\ \{0.4566,\ 0.3383,\ 0.2052\},\qquad 3.656\,\varepsilon_1. \end{aligned}$$

The probabilities p_i are shown in Fig. 11.4 (right). The average energy $\varepsilon \approx 3.656\,\varepsilon_1$ in the case of $1/\lambda_1 = k_\mathrm{B}T = 10\,\varepsilon_1$, lying just slightly below ε_2, is denoted by the dashed line in the level scheme (left part of Figure).

The reverse task is also interesting: to what temperature must the system be heated that the average energy will equal a specific value? If, for example, we wish to attain $\varepsilon = 3\,\varepsilon_1$, we must set $1/\lambda_1 = 5.5455\,\varepsilon_1$, i.e. crank up the heater to $T = 5.5455\,\varepsilon_1/k_\mathrm{B}$, where $\{p_1, p_2, p_3\} \approx \{0.5499,\ 0.3201,\ 0.1299\}$. ◁

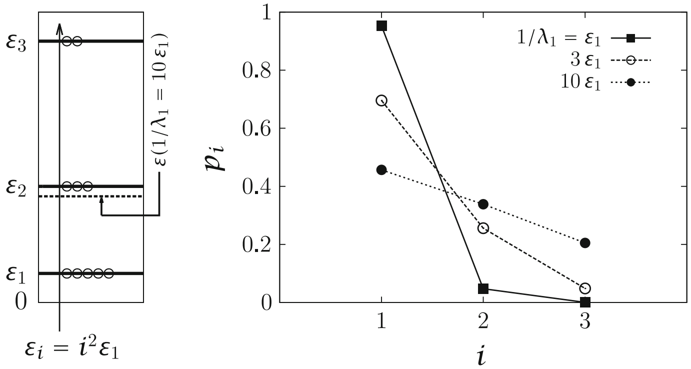


Fig. 11.4 Maxwell-Boltzmann statistics in a three-level system. [Left] The circles approximately denote the distribution of $N = 10$ particles corresponding to $1/\lambda_1 = 10\,\varepsilon_1$ ($N_1 = p_1N \approx 5$, $N_2 = p_2N \approx 3$, $N_3 = p_3N \approx 2$), and the *dashed line* indicates the average energy $\varepsilon \approx 3.656\,\varepsilon_1$. [Right] The maximum-entropy probability distribution for various values of $1/\lambda_1 = k_\mathrm{B}T$. At high T the distribution approaches the uniform distribution!

11.3.4 Relation Between Information and Thermodynamic Entropy

The mathematical form of the theory of information entropy is identical to the formulas for entropy obtained in the framework of statistical mechanics. In other words: the rules of statistical mechanics are principles of statistical inference in physical garb. Let $\varepsilon_i(\alpha_1, \alpha_2, \ldots)$ be the energy levels of a physical system with parameters α_i specifying quantities like volume, external electro-magnetic field, gravitational potential, and so on. At given mean energy ε the probabilities p_i for the occupation of levels ε_i are given by a special form of (11.8), readily identified as the Maxwell-Boltzmann distribution if one identifies $\lambda_1 = 1/(k_\mathrm{B}T)$. Similar arguments [8, 9] can be used to accommodate free energy

$$F(T, \alpha_1, \alpha_2, \ldots) = U - TS = -k_\mathrm{B}T \log Z(T, \alpha_1, \alpha_2, \ldots) \qquad (11.17)$$

in the framework of statistical inference, as well as thermodynamic entropy,

$$S = -\frac{\partial F}{\partial T} = -k_\mathrm{B} \sum_i p_i \log p_i,$$

formally differing from the information entropy only by the Boltzmann constant providing the appropriate units.

From (11.17) it also becomes clear why in the three-level system in Fig. 11.4 at high temperatures particles fail to accumulate at the highest level as one might intuitively expect, but rather their distribution approaches the uniform distribution (p_1, p_2 and p_3 all tend to $1/3$). Minimizing the free energy $F = U - TS$ at $T = 0$ means minimizing the internal energy U, so at $T = 0$ indeed all particles occupy the lowest level. But at high T one has $U \ll TS$, so in this limit minimizing F implies *maximizing* S—thus the uniformity of the distribution.

Example A pair of elementary magnetic dipoles (e.g. electrons with magnetic moments μ_0 treated classically) is exposed to a homogeneous external field $\vec{B}$. Each dipole can only be oriented along $\vec{B}$ or opposite to it, so four configurations are possible, shown in Fig. 11.5 together with their magnetic energies $\varepsilon_\mathrm{m} = -\vec{\mu}_0\vec{B}$. At what temperature the average energy of this system (at given magnetic field density $B = |\vec{B}|$) is equal to $\varepsilon = -\mu_0 B$ and the entropy maximal?

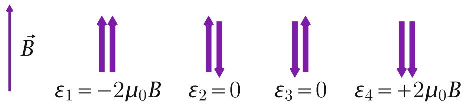


Fig. 11.5 Configurations of a pair of magnetic dipoles (with individual magnetic moments μ_0) in an external field and the corresponding magnetic energies

From (11.16) we obtain $\sum_{i=1}^{4} \varepsilon_i \,\mathrm{e}^{-\lambda_1 \varepsilon_i} = \varepsilon \sum_{i=1}^{4} \mathrm{e}^{-\lambda_1 \varepsilon_i}$ or

$$-2\mu_0 B \,\mathrm{e}^{-\lambda_1(-2\mu_0 B)} + 2\mu_0 B \,\mathrm{e}^{-\lambda_1(2\mu_0 B)} = -\mu_0 B \left(2 + \mathrm{e}^{-\lambda_1(-2\mu_0 B)} + \mathrm{e}^{-\lambda_1(2\mu_0 B)}\right).$$

Solving this equation for λ_1 we get $\lambda_1 = 1/(k_\mathrm{B} T) = \frac{1}{2}\log 3/(\mu_0 B) \approx 0.549/(\mu_0 B)$ or $T \approx 1.82\,\mu_0 B/k_\mathrm{B}$. At this temperature the expected occupation probabilities are $p_1 = 0.5625$, $p_2 = p_3 = 0.1875$ and $p_4 = 0.0625$. ◁

11.3.5 Bose–Einstein Distribution

Let us discuss a more complex problem of N particles that may occupy n energy levels with energies ε_i, $i = 1, 2, \ldots, n$. Let p_{ij} be the conditional probability that the ith level contains j particles, $j = 0, 1, 2, \ldots$ (the number of particles on the individual level is not restricted). The condition—namely that the system is in the ith state—is given by the prior probability q_i, so that $P_{ij} = q_i p_{ij}$. If the probabilities q_i are unknown, we may recall Laplace and simply set $q_i = 1/n$. The distribution of particles p_{ij}, consistent with the requirement of maximum entropy, is then found by maximizing the entropy with the constraints

$$\sum_{i=1}^{n} q_i = 1, \quad \sum_{j=0}^{\infty} p_{ij} = 1, \quad \sum_{i=1}^{n} q_i \sum_{j=0}^{\infty} j p_{ij} = N, \quad \sum_{i=1}^{n} q_i \varepsilon_i \sum_{j=0}^{\infty} j p_{ij} = \varepsilon, \qquad (11.18)$$

where ε is the prescribed average system energy. When the system is in the ith state, its entropy is $-\sum_{j=0}^{\infty} p_{ij} \log p_{ij}$, so the total entropy is

$$-\sum_{i=1}^{n} q_i \sum_{j=0}^{\infty} p_{ij} \log p_{ij}.$$

This can also be seen if the expression for entropy is rewritten as

$$-\sum_{ij} P_{ij} \log P_{ij} = -\sum_{ij} q_i p_{ij} \log\left(q_i p_{ij}\right) = -\sum_i q_i \log q_i - \sum_i q_i \sum_j p_{ij} \log p_{ij},$$

where we have used $\sum_j p_{ij} = 1$, so the first term plays no role in taking the derivative with respect to p_{ij}. The Lagrange function to be minimized is then

$$\begin{aligned}\mathcal{L} = &-\sum_{i=1}^{n} q_i \sum_{j=0}^{\infty} p_{ij} \log p_{ij} - \sum_{i=1}^{n} (\lambda_i - q_i) \left(\sum_{j=0}^{\infty} p_{ij} - 1\right) \\ &-\alpha \left(\sum_{i=1}^{n} q_i \sum_{j=0}^{\infty} j p_{ij} - N\right) - \beta \left(\sum_{i=1}^{n} q_i \varepsilon_i \sum_{j=0}^{\infty} j p_{ij} - \varepsilon\right), \qquad (11.19)\end{aligned}$$

where α and β are additional Lagrange multipliers. Taking the derivative with respect to p_{ij} we get $\partial\mathcal{L}/\partial p_{ij} = -q_i(1 + \log p_{ij}) - (\lambda_i - q_i) - \alpha j q_i - \beta j q_i \varepsilon_i = 0$. From here we express p_{ij} and insert it in (11.18). A brief calculation [5] then leads to the occupation probabilities

$$p_{ij} = A_i \,\mathrm{e}^{-(\alpha+\beta\varepsilon_i)j}, \qquad A_i = 1 - \mathrm{e}^{-(\alpha+\beta\varepsilon_i)}, \tag{11.20}$$

as well as to the formulas for the number of particles and system energy,

$$N = \sum_{i=1}^{n} \frac{q_i}{\exp(\alpha + \beta\varepsilon_i) - 1}, \qquad \varepsilon = \sum_{i=1}^{n} \frac{q_i \varepsilon_i}{\exp(\alpha + \beta\varepsilon_i) - 1}, \tag{11.21}$$

where $\beta = 1/(k_\mathrm{B}T)$. We also set $\alpha = \beta\mu = \mu/(k_\mathrm{B}T)$, where μ is the chemical potential. The expected number of particles on the ith level is

$$N_i = \sum_{j=0}^{\infty} j p_{ij} = \sum_{j=0}^{\infty} j A_i \,\mathrm{e}^{-(\alpha+\beta\varepsilon_i)j} = \frac{1}{\exp(\alpha + \beta\varepsilon_i) - 1}, \tag{11.22}$$

where α and β must be determined from (11.21) at known N and ε. One also has

$$N = \sum_{i=1}^{n} q_i N_i, \qquad \varepsilon = \sum_{i=1}^{n} q_i \varepsilon_i N_i.$$

The obtained distribution is suitable for the description of particles with integer spin (bosons), e.g. photons, atoms with even numbers of electrons, and nuclei with even numbers of nucleons.

If we are dealing with bosons whose number is not conserved (virtual photons, phonons, magnons), there is no restriction on the particle number N, hence the third constraint in (11.18) and the third term in (11.19) are superfluous. In this case $\alpha = 0$ and therefore the chemical potential also vanishes: $\mu = 0$.

11.3.6 Fermi–Dirac Distribution

The Fermi-Dirac distribution describes fermionic systems, i.e. systems of particles with non-integer spins $\left(\frac{1}{2}, \frac{3}{2}, \ldots\right)$. For such particles Fermi's rule says that the same energy level can not be occupied by two (or more) particles: the level can be vacant or inhabited by precisely one particle. To determine the corresponding maximum-entropy distribution we can exploit our previous derivation, where we restrict $j = 0, 1$ in (11.18). The expression for occupation probabilities still has the form (11.20), while (11.21) is replaced by

$$N = \sum_{i=1}^{n} \frac{q_i}{\exp(\alpha + \beta\varepsilon_i) + 1}, \qquad \varepsilon = \sum_{i=1}^{n} \frac{q_i \varepsilon_i}{\exp(\alpha + \beta\varepsilon_i) + 1},$$

and (11.22) by

$$N_i = \frac{1}{\exp(\alpha + \beta\varepsilon_i) + 1}.$$

11.4 Continuous Distributions with Maximum Entropy

Maximum-entropy *continuous* distributions with imposed additional constraints can also be handled by the Lagrange multiplier method. We discuss a single paradigmatic case: a distribution whose variance σ^2 is prescribed and has maximum entropy. One must maximize (11.3) with the constraints $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$ and $\int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,\mathrm{d}x = \sigma^2$, where μ is a free parameter. By analogy to the discrete case (11.6) the Lagrange function is

$$\mathcal{L} = -\int_{-\infty}^{\infty} f(x) \log f(x)\,\mathrm{d}x - (\lambda_0 - 1)\left[\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x - 1\right] - \lambda_1\left[\int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,\mathrm{d}x - \sigma^2\right].$$

The variation of the first term is

$$\delta\left[-\int_{-\infty}^{\infty} f \log f\,\mathrm{d}x\right] = -\int_{-\infty}^{\infty}\left[\delta f \log f + f\frac{1}{f}\,\delta f\right]\mathrm{d}x = \int_{-\infty}^{\infty}(-\delta f)\Big[\log f + 1\Big]\,\mathrm{d}x.$$

By the variation of the whole Lagrange function, which is set to zero,

$$\delta\mathcal{L} = \int_{-\infty}^{\infty}\big(-\delta f(x)\big)\left[\log f(x) + \lambda_0 + \lambda_1 (x - \mu)^2\right]\,\mathrm{d}x = 0,$$

we get $\log f(x) + \lambda_0 + \lambda_1 (x - \mu)^2 = 0$ or $f(x) = \exp\big(-\lambda_0 - \lambda_1 (x - \mu)^2\big)$. We insert this function into the constraint equations:

$$1 = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \mathrm{e}^{-\lambda_0}\int_{-\infty}^{\infty} \mathrm{e}^{-\lambda_1 (x-\mu)^2}\,\mathrm{d}x = \frac{\mathrm{e}^{-\lambda_0}}{\sqrt{\lambda_1}}\int_{-\infty}^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t = \sqrt{\frac{\pi}{\lambda_1}}\,\mathrm{e}^{-\lambda_0},$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,\mathrm{d}x = \frac{\mathrm{e}^{-\lambda_0}}{\sqrt{\lambda_1^3}}\int_{-\infty}^{\infty} t^2\,\mathrm{e}^{-t^2}\,\mathrm{d}t = \frac{1}{2}\sqrt{\frac{\pi}{\lambda_1^3}}\,\mathrm{e}^{-\lambda_0}.$$

It follows that $\lambda_1 = 1/(2\sigma^2)$, then the first equation gives $\mathrm{e}^{-\lambda_0} = 1/(\sqrt{2\pi}\,\sigma)$. Hence the desired density f corresponds to the normal distribution (3.7).

Two further interesting results refer to case when the mean is prescribed and the case that the definition domain of X is a finite interval. The following theorems on maximum-entropy distributions applicable to continuous random variables X with density f and entropy $H(X)$ are given without proof:

1. If $\mathrm{var}[X] = \sigma^2 \neq \infty$, then $H(X)$ exists and $H(X) \leq \log\sqrt{2\pi\,\mathrm{e}\,\sigma^2}$ holds true, where the equality applies only if $X \sim N(\,\cdot\,, \sigma^2)$—in this case the mean can be anything, as only an offset along the x-axis is involved.
2. If X is a non-negative random variable ($f(x) = 0$ for $x < 0$) with finite mean $E[X] = \mu$, then $H(X)$ exists and $H(X) \leq \log(\mu\mathrm{e})$ holds true, where the equality applies only if $X \sim \mathrm{Exp}(1/\mu)$.
3. If the variable X is restricted to the interval $[a, b]$, i.e. $f(x) = 0$ for $x < a$ and $x > b$, then $H(X)$ exists and $H(X) \leq \log(b - a)$ holds true, where the equality applies only if $X \sim U(a, b)$—see (11.4).

Example Let us fix the variance of a continuous distribution, σ^2, and check that the exponential and uniform distributions with such variance have lower entropy than the normal with the same variance. By Theorem 1 the normal distribution has entropy $\log\sqrt{2\pi\,\mathrm{e}\,\sigma^2} = \log\sigma + \log\sqrt{2\pi\mathrm{e}} \approx \log\sigma + 1.42$. The exponential distribution has $\sigma = \sqrt{\mathrm{var}[X]} = E[X] = \mu$ (see (3.4) and Table 4.1 and set $\lambda = 1/\mu$). By Theorem 2 its entropy is $\log(\mu\mathrm{e}) = \log(\sigma\mathrm{e}) = \log\sigma + 1$. The uniform distribution on $[a, b]$ has $\sigma^2 = (b - a)^2/12$ (Table 4.1), so by Theorem 3 its entropy is $\log(b - a) = \log\bigl(\sigma\sqrt{12}\bigr) \approx \log\sigma + 1.24$. ◁

11.5 Maximum-Entropy Spectral Analysis

The maximum-entropy principle also leads to a powerful tool for spectral analysis of time series. Suppose we have a set of (generally complex) observations $\boldsymbol{x} = \{x_0, x_1, \ldots, x_T\}$ at times $0, 1, \ldots, T$. We are interested in the probability distribution of these values, f, that minimizes the entropy $H = -\int f(\boldsymbol{x}) \log f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. From the data x_i (values of X_i) we first form our fundamental observables, the temporal autocorrelations

$$A_k(\boldsymbol{X}) = \frac{1}{T+1}\sum_{i=0}^{T-k} X_i^* X_{i+k}, \qquad k = 0, 1, \ldots, m, \qquad m < T.$$

Their expected values

$$E[A_k(\boldsymbol{X})] = \int A_k(\boldsymbol{x}) f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \equiv \overline{A}_k$$

may be understood as continuous analogues of discrete constraints (11.5). If X_i are complex, A_k are also complex in general, but it is readily noticed that

$$A_k = \frac{1}{T+1}\sum_{i=0}^{T-k} X_i^* X_{i+k} = \frac{1}{T+1}\sum_{i=k}^{T} X_{i-k}^* X_i = \left(\frac{1}{T+1}\sum_{i=k}^{T} X_i^* X_{i-k}\right)^* := A_{-k}^*.$$

As we shall see, this symmetry is essential if we wish that certain quantities—for instance, the frequency spectrum of the measured signal—is purely real. We must maximize the Lagrange function

$$\mathcal{L} = H - \sum_{k=0}^{m} \alpha_k \,\mathrm{Re}\, A_k - \sum_{k=1}^{m} \beta_k \,\mathrm{Im}\, A_k,$$

where α_k and β_k are unknown Lagrange multipliers. The second sum runs from $k = 1$, since the auto-correlation $A_0 = (T+1)^{-1}\sum_i |X_i|^2$ is purely real, hence $\mathrm{Im}\, A_0 = 0$. Therefore

$$\begin{aligned}\mathcal{L} &= H - \sum_{k=0}^{m} \frac{\alpha_k}{2}\left(A_k + A_{-k}\right) - \sum_{k=1}^{m} \frac{\beta_k}{2\,\mathrm{i}}\left(A_k - A_{-k}\right)\\ &= H - \alpha_0 A_0 - \sum_{k=1}^{m} \frac{\alpha_k - \mathrm{i}\,\beta_k}{2} A_k - \sum_{k=1}^{m} \frac{\alpha_k + \mathrm{i}\,\beta_k}{2} A_{-k} = H - \sum_{k=-m}^{m} \lambda_k A_k,\end{aligned}$$

where we have denoted $\lambda_k = (\alpha_k - \mathrm{i}\,\beta_k)/2$ and $\lambda_{-k} = (\alpha_k + \mathrm{i}\,\beta_k)/2$, so that $\lambda_{-k} = \lambda_k^*$. We know how to solve the problem of maximizing such a Lagrange function from the discrete case: see formulas (11.9) and (11.10). The probability density and the partition function have the form

$$f(\boldsymbol{x}) = Z^{-1} \exp\left(-\sum_{k=-m}^{m} \lambda_k A_k(\boldsymbol{x})\right), \qquad Z = \int \exp\left(-\sum_{k=-m}^{m} \lambda_k A_k(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x},$$

while the constraint equations are

$$\frac{\partial(\log Z)}{\partial \lambda_k} = -\overline{A}_k, \qquad k = -m, \ldots, m. \tag{11.23}$$

The sum $\sum_k \lambda_k A_k$ in the arguments of the exponentials can be written as

$$\frac{1}{T+1}\left(\sum_{k=0}^{m} \lambda_k \sum_{i=0}^{T-k} x_i^* x_{i+k} + \sum_{k=1}^{m} \lambda_{-k} \sum_{i=k}^{T} x_i^* x_{i-k}\right) = \frac{1}{2}\sum_{i,j=0}^{T} x_i^* B_{ij} x_j = \frac{1}{2}\boldsymbol{x}^\dagger B \boldsymbol{x},$$

where B is a banded ($(2m+1)$-diagonal), $(T+1) \times (T+1)$ Toeplitz matrix with the elements

$$B_{ij} = \frac{2}{T+1} C_{ij}, \qquad C_{ij} = \begin{cases} \lambda_{j-i} & ; \; |i-j| \le m, \\ 0 & ; \; |i-j| > m. \end{cases} \tag{11.24}$$

The properly normalized probability density we are seeking is therefore

$$f(\boldsymbol{x}) = \frac{|\det B|^{1/2}}{(2\pi)^{(T+1)/2}} \exp\left(-\frac{1}{2}\,\boldsymbol{x}^\dagger B\boldsymbol{x}\right), \tag{11.25}$$

which is the density of the multivariate normal distribution (4.23). Note that B is Hermitian, $B = B^\dagger$. It must also be positive definite, otherwise f does not exist.

11.5.1 Calculating the Lagrange Multipliers

The entropy corresponding to the obtained probability density is

$$\begin{aligned} H &= -\int f(\boldsymbol{x}) \log f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = -\int f(\boldsymbol{x}) \left[\log \frac{|\det B|^{1/2}}{(2\pi)^{(T+1)/2}} - \frac{1}{2}\boldsymbol{x}^\dagger B\boldsymbol{x}\right] \mathrm{d}\boldsymbol{x} \\ &= -\log \frac{|\det B|^{1/2}}{(2\pi)^{(T+1)/2}} + \frac{1}{2}E\left[\boldsymbol{X}^\dagger B\boldsymbol{X}\right] \\ &= -\frac{1}{2}\log|\det B| + \frac{T+1}{2}\big(1+\log 2\pi\big). \end{aligned} \tag{11.26}$$

Here we have used the relation $E\big[\boldsymbol{X}^\dagger B\boldsymbol{X}\big] = T+1$ which is easy to prove.[3] The phase sum is nothing but the multi-dimensional Gauss integral

$$Z = \int \exp\left(-\frac{1}{2}\boldsymbol{x}^\dagger B\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x} = \frac{(2\pi)^{(T+1)/2}}{|\det B|^{1/2}},$$

so that

$$\log Z = \frac{T+1}{2}\log 2\pi - \frac{1}{2}\log|\det B|.$$

Using the relation between the determinants $|\det B| = (2/(T+1))^{T+1}|\det C|$ or

$$\log|\det B| = (T+1)\log(2/(T+1)) + \log|\det C|, \tag{11.27}$$

[3] Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^{\mathrm{T}}$ be a d-dimensional vector of complex random variables with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^{\mathrm{T}}$ and covariance matrix Σ. Then any constant symmetric matrix M satisfies $E\big[\boldsymbol{X}^\dagger M\boldsymbol{X}\big] = E\big[\mathrm{tr}\big(\boldsymbol{X}^\dagger M\boldsymbol{X}\big)\big] = E\big[\mathrm{tr}\big(M\boldsymbol{X}\boldsymbol{X}^\dagger\big)\big] = \mathrm{tr}\big(M\,E\big[\boldsymbol{X}\boldsymbol{X}^\dagger\big]\big) = \mathrm{tr}\big(M\big(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\dagger\big)\big) = \mathrm{tr}\big(M\Sigma\big) + \mathrm{tr}\big(M\boldsymbol{\mu}\boldsymbol{\mu}^\dagger\big) = \mathrm{tr}\big(M\Sigma\big) + \boldsymbol{\mu}^\dagger M\boldsymbol{\mu}$. In our case (see (11.25)) B in the density $f_{\boldsymbol{X}}(\boldsymbol{x})$ is Hermitian, but by decomposing $B = K^\dagger K$ and transforming $\boldsymbol{Y} = K\boldsymbol{x}$ we can write $\boldsymbol{X}^\dagger B\boldsymbol{X} = \boldsymbol{Y}^\dagger\boldsymbol{Y}$, so that $f_{\boldsymbol{Y}}(\boldsymbol{y}) \propto \exp\big(-\boldsymbol{y}^\dagger\boldsymbol{y}/2\big)$. This is the density of the multivariate normal distribution with mean $\boldsymbol{\mu} = \boldsymbol{0}$ and covariance matrix $\Sigma = I$. It follows that $E\big[\boldsymbol{X}^\dagger B\boldsymbol{X}\big] = E\big[\boldsymbol{Y}^\dagger\boldsymbol{Y}\big] = d = T+1$.

(11.23) can be written as

$$\overline{A}_k = -\frac{\partial(\log Z)}{\partial\lambda_k} = \frac{1}{2}\frac{\partial}{\partial\lambda_k}\bigl(\log|\det C|\bigr), \qquad k = -m, \ldots, m.$$

By Szegő's theorem [12] at fixed m we have

$$\lim_{T\to\infty}\frac{1}{T+1}\log|\det C| = \frac{1}{2\pi}\int_0^{2\pi}\log p(\phi)\,\mathrm{d}\phi, \tag{11.28}$$

where

$$p(\phi) = \sum_{k=-m}^{m}\lambda_k\,\mathrm{e}^{\mathrm{i}\,k\phi}.$$

For $T \gg m$ this leads to the approximation

$$\overline{A}_k = \frac{T+1}{2}\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{p(\phi)}\frac{\partial p(\phi)}{\partial\lambda_k}\,\mathrm{d}\phi = \frac{T+1}{2}\frac{1}{2\pi}\int_0^{2\pi}\frac{\mathrm{e}^{\mathrm{i}\,k\phi}}{p(\phi)}\,\mathrm{d}\phi.$$

This is the key formula connecting the auto-correlations $\overline{A}_k$ to the Lagrange multipliers λ_k contained in $p(\phi)$. Let us denote $\eta_k = (2/(T+1))\lambda_k$ and write

$$\widetilde{p}(\phi) = \frac{2}{T+1}p(\phi) = \sum_{k=-m}^{m}\eta_k\,\mathrm{e}^{\mathrm{i}\,k\phi} = \sum_{k=-m}^{m}\eta_k z^k = g(z), \tag{11.29}$$

where $z = \mathrm{e}^{\mathrm{i}\,\phi}$. Calculating $\overline{A}_k$ requires an integration along the unit circle in the complex plane:

$$\overline{A}_k = \frac{1}{2\pi}\int_0^{2\pi}\frac{\mathrm{e}^{\mathrm{i}\,k\phi}}{\widetilde{p}(\phi)}\,\mathrm{d}\phi = \frac{1}{2\pi\,\mathrm{i}}\oint\frac{z^{k-1}}{g(z)}\,\mathrm{d}z, \qquad k = -m, \ldots, m. \tag{11.30}$$

It turns out [13] that $g(z)$ can be factorized as

$$g(z) = G(z)\bigl[G(1/z^*)\bigr]^*, \qquad G(z) = \sum_{k=0}^{m} g_k z^{-k}. \tag{11.31}$$

The first factor, $G(z)$, has zeros only within the unit circle, and the second factor only outside of it; the function $g(z)$ has m zeros *within* the unit circle, m zeros *outside* and one *on the circle*. By the convolution of g_j and $\overline{A}_k$ we get

$$\sum_{j=0}^{m} g_j\overline{A}_{k-j} = \frac{1}{2\pi\,\mathrm{i}}\oint\frac{\sum_{j=0}^{m} g_j\,z^{k-j-1}}{G(z)\bigl[G(1/z^*)\bigr]^*}\,\mathrm{d}z = \frac{1}{2\pi\,\mathrm{i}}\oint\frac{z^{k-1}}{\bigl[G(1/z^*)\bigr]^*}\,\mathrm{d}z = \frac{1}{g_0^*}\,\delta_{k,0}.$$

This can be recast as the *Yule-Walker* system of equations

$$\sum_{j=0}^{m} h_j \overline{A}_{k-j} = \delta_{k,0}, \qquad h_j = g_0^* g_j, \qquad k = 0, 1, \ldots, m,$$

or

$$\begin{pmatrix} \overline{A}_0 & \overline{A}_{-1} & \cdots & \overline{A}_{-m} \\ \overline{A}_1 & \overline{A}_0 & \cdots & \overline{A}_{-m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{A}_m & \overline{A}_{m-1} & \cdots & \overline{A}_0 \end{pmatrix} \begin{pmatrix} h_0 \\ h_1 \\ \vdots \\ h_m \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

where $\overline{A}_{-k} = \overline{A}_k^*$ for all matrix elements. The last step is to use the obtained h_k to calculate the Lagrange multipliers

$$\lambda_k = \frac{T+1}{2}\,\eta_k, \qquad \eta_k = \sum_{j=j_{\min}}^{j_{\max}} g_j g_{j+k}^* = \frac{1}{|g_0|^2} \sum_{j=j_{\min}}^{j_{\max}} h_j h_{j+k}^*, \qquad k = -m, \ldots, m,$$

where $j_{\min} = \max\{0, -k\}$ and $j_{\max} = \min\{m, m-k\}$, so that $\eta_{-k} = \eta_k^*$ and $\lambda_{-k} = \lambda_k^*$. The matrix B from the definition (11.24) is thereby uniquely determined, and with it the probability density (11.25).

11.5.2 Estimating the Spectrum

The *power spectral density* (PSD) of a signal is defined as

$$S(\omega) = \sum_{k=-\infty}^{\infty} \overline{A}_k \, \mathrm{e}^{-\mathrm{i}\,k\omega},$$

where $\omega = 2\pi\nu$ and $\overline{A}_k$ are the auto-correlations of an infinite signal. In the true world we usually only know its finite sample, so the obtained formulas will offer just an estimate of the true frequency spectrum. With the calculated h_k ($k = 0, 1, \ldots, m$), the solutions of the Yule-Walker system, we define the function

$$H(z) = \sum_{k=0}^{m} h_k z^{-k} = \sum_{k=0}^{m} g_0^* g_k z^{-k} = g_0^* G(z).$$

From (11.31) it follows that

$$g\big(\mathrm{e}^{\mathrm{i}\,\omega}\big) = G\big(\mathrm{e}^{\mathrm{i}\,\omega}\big)\big[G\big(\mathrm{e}^{\mathrm{i}\,\omega}\big)\big]^* = \big|G\big(\mathrm{e}^{\mathrm{i}\,\omega}\big)\big|^2 = \frac{1}{|g_0|^2}\,\big|H\big(\mathrm{e}^{\mathrm{i}\,\omega}\big)\big|^2.$$

On the other hand, (11.30) can be used to write

$$\sum_{k=-\infty}^{\infty} \overline{A}_k \, \mathrm{e}^{-\mathrm{i}\,k\omega} = \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{\widetilde{p}(\phi)} \underbrace{\left(\sum_{k=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\,k(\phi-\omega)} \right)}_{2\pi\delta_{[0,2\pi]}(\omega)} \mathrm{d}\phi = \frac{1}{\widetilde{p}(\omega)} = \frac{1}{g\bigl(\mathrm{e}^{\mathrm{i}\,\omega}\bigr)}.$$

The power spectral density can therefore be estimated as

$$\widehat{S}(\omega) = \frac{|g_0|^2}{\bigl|H\bigl(\mathrm{e}^{\mathrm{i}\,\omega}\bigr)\bigr|^2} = \frac{h_0}{\bigl|\sum_{k=0}^{m} h_k \, \mathrm{e}^{-\mathrm{i}\,k\omega}\bigr|^2}, \qquad |\omega| \le \pi.$$

This formula (up to a multiplicative constant) is usually seen in the form

$$\widehat{S}(\nu) = \frac{\sigma^2}{\bigl|1 + \sum_{k=1}^{m} a_k \, \mathrm{e}^{-\mathrm{i}\,2\pi k\nu\Delta t}\bigr|^2}, \qquad |\nu| \le \frac{1}{2\Delta t}, \tag{11.32}$$

where $\Delta t = t_{i+1} - t_i$ $(i = 0, 1, \ldots, T-1)$, $\sigma^2 = 1/h_0$ and $a_k = h_k/h_0$. In signal-processing theory—see, for instance, [14]—the parameter σ^2 represents the variance of the Gaussian noise generated by the signal through the feedback loop with filter $F(z) = \sum_k a_k z^{-k}$.

The described spectral estimation tool is called the *auto-regression model;* if the process (the signal) has a Gaussian nature, it is also known as *Maximum-Entropy Spectral Analysis* (MESA) [15]. How MESA works in practice is demonstrated by the following Example.

Example We use the auto-regression method to analyze the signal

$$x_i = \sin\Bigl(i\,\tfrac{2\pi}{10}\Bigr) + 2\sin\Bigl(i\,\tfrac{4\pi}{10}\Bigr) + 3\sin\Bigl(i\,\tfrac{6\pi}{10}\Bigr) + 5\Bigl(\mathcal{R} - \tfrac{1}{2}\Bigr), \tag{11.33}$$

where $\mathcal{R} \sim U(0,1)$ and $i = 0, 1, \ldots, 1023$, thus $T + 1 = 1024$. The time series contains three frequency components with frequencies $\nu = 0.1$, 0.2 and 0.3 with amplitudes 1, 2 and 3, respectively. Besides, it is very noisy (last term in (11.33)). The sample of the first 500 values of the signal is shown in Fig. 11.6 (left).

The estimate of the spectrum, calculated by (11.32) for two different m, is shown in Fig. 11.6 (right). With increasing m the spikes become sharper, but spurious peaks start to appear that are not expected to be present in the spectrum and represent noise. ◁

By using the entropy expression and the relation between the determinants of B and C one can derive an interesting formula relating entropy to power spectral density. At large enough T Szegő's theorem (11.28) can be understood as

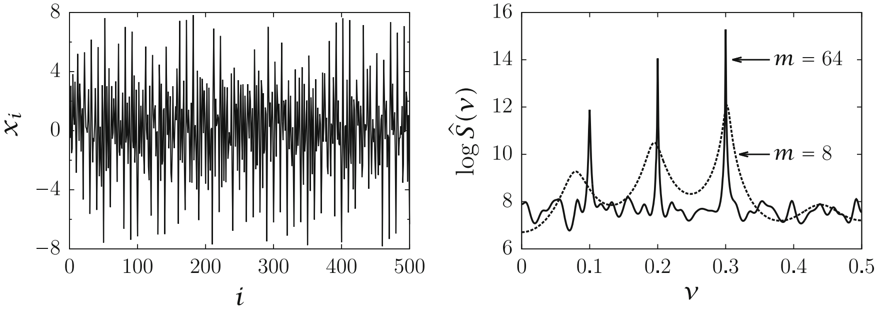


Fig. 11.6 [Left] The first 500 values of the signal (11.33). [Right] The estimate of the spectrum by using the auto-regression method for two different m

$$\log|\det C| \approx \frac{T+1}{2\pi}\int_0^{2\pi} \log p(\phi)\,\mathrm{d}\phi = \frac{T+1}{2\pi}\left[-2\pi\log\frac{2}{T+1} + \int_0^{2\pi}\log\widetilde{p}(\phi)\,\mathrm{d}\phi\right],$$

where we have used the relation (11.29) between p and $\widetilde{p}$. We insert this in (11.26) and consider (11.27); see also definition (11.24). It follows that

$$\begin{aligned} H &\approx -\frac{1}{2}\left\{(T+1)\log\frac{2}{T+1} + \frac{T+1}{2\pi}\left[-2\pi\log\frac{2}{T+1} + \int_0^{2\pi}\log\widetilde{p}(\phi)\,\mathrm{d}\phi\right]\right\} \\ &\quad + \frac{T+1}{2}\,(1+\log 2\pi) \\ &= \frac{T+1}{2}\left[(1+\log 2\pi) - \frac{1}{2\pi}\int_0^{2\pi}\log\widetilde{p}(\phi)\,\mathrm{d}\phi\right]. \end{aligned}$$

Since $S(\omega) = 1/\widetilde{p}(\omega)$, this also means

$$\frac{H}{T+1} \approx \frac{1}{2}\log 2\pi\mathrm{e} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log S(\omega)\,\mathrm{d}\omega.$$

where $H/(T+1)$ is the change of entropy per unit time (*entropy rate*). This is the average "speed" of acquiring information in the measurement of a spectrum.

References

1. H. Nyquist, Certain factors affecting telegraph speed. Bell Syst. Tech. J. **3**, 324 (1924)
2. R.V.L. Hartley, Transmission of information. Bell Syst. Tech. J. **7**, 535 (1928)
3. C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27** (1948) 379 (Part I) and 623 (Part II)
4. M. Tribus, Thirty years of information theory, in *Maximum Entropy Formalism*, ed. by R.D. Levine, M. Tribus (MIT Press, Cambridge, 1979), pp. 1–19

5. J.N. Kapur, *Maximum Entropy Models in Science and Engineering* (Wiley Eastern Ltd., New Delhi, 1989)
6. S. Kullback, R.A. Leibler, On information and sufficiency. Ann. Math. Stat. **22**, 79 (1951)
7. S.M. Hanasoge, M. Branicki, Interpreting cross-correlations of one-bit filtered seismic noise. Geophys. J. Int. **195**, 1811 (2013)
8. E.T. Jaynes, Information theory and statistical mechanics. Phys. Rev. **106**, 620 (1957)
9. E.T. Jaynes, Information theory and statistical mechanics. II. Phys. Rev. **108**, 171 (1957)
10. S. Pressé, K. Ghosh, J. Lee, K.A. Dill, Principles of maximum entropy and maximum caliber in statistical physics. Rev. Mod. Phys. **85**, 1115 (2013)
11. P. Harremoës, F. Topsøe, Maximum entropy fundamentals. Entropy **3**, 191 (2001)
12. G. Szegő, Ein Grenzwertsatz über die Toeplitzschen Determinanten einer reellen positiven Funktion. Math. Ann. **76**, 490 (1915)
13. E.T. Jaynes, On the rationale of maximum-entropy methods. IEEE Proc. **70**, 939 (1982)
14. P. Stoica, R. Moses, *Spectral Analysis of Signals* (Prentice-Hall Inc, New Jersey, 2005)
15. J.P. Burg, Maximum entropy spectral analysis, lecture at the 37th annual international meeting, Soc. Explor. Geophys., Oklahoma City, Oklahoma, 31 Oct 1967

Chapter 12
Markov Processes ⋆

Abstract Markov processes are introduced as memoryless stochastic processes and classified in four classes based on whether the time parameter is continuous or discrete and whether the sample space is continuous or discrete. Two of them are treated in more detail: discrete-time ("classical") Markov chains and continuous-time, continuous-state Markov processes. Long-time behavior of the chains is discussed, establishing the conditions for the formation of equilibrium distributions. In the continuous case, the Markov propagator is defined along with a discussion of moment functions, characterizing functions, and time evolution of the moments. Two particular Markov processes, the Wiener and the Ornstein–Uhlenbeck process, are given special attention due to their relevance for the study of diffusion.

Imagine a sequence of random variables $X(t)$ describing the state of a dynamical system at times t. In Sect. 6.7 such a sequence was called a random or stochastic process [1]. A special class of random processes consists of processes in which the state of the system at current t depends only on its state just prior to t, while all earlier states are irrelevant for its time evolution: the process is *memoryless*. Such processes are called *Markov processes* after the Russian mathematician A.A. Markov (1856–1922).

Markov harnessed statistical methods to analyze letter sequences in Pushkin's poem *Eugene Onegin:* he was seeking probabilities of a vowel preceding a consonant, a vowel appearing after the consonant, and so on, as well as the answer to the question whether such estimates change with the length of the analyzed text and whether Pushkin's "statistical profile" is perhaps unique. In 1913 he presented his findings to the Imperial Academy of Sciences in St. Petersburg [2] and thereby initiated a completely novel field of research [3, 4].

Markov processes are divided in four families based on whether the time parameter is continuous or discrete and whether the values of X are continuous or discrete. In the following we shall discuss two combinations: discrete variables with discrete time steps—such processes are known as discrete-time Markov chains—and continuous variables with a continuous time evolution.

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_12

12.1 Discrete-Time (Classical) Markov Chains

A classical Markov chain is a random process $X(t)$ in a finite discrete state space $\Omega = \{i_0, i_1, \ldots, i_m\}$ with discrete time $t = 0, 1, 2, \ldots$ For simplicity we denote $X(t) = X_t$. The "memoryless" feature of the process is expressed by the relation

$$P\bigl(X_{t+1} = j_{t+1} \mid X_t = j_t, X_{t-1} = j_{t-1}, \ldots, X_0 = j_0\bigr) = P\bigl(X_{t+1} = j_{t+1} \mid X_t = j_t\bigr).$$

In plain words: the probability of arriving to the state-space point j_{t+1} at time $t+1$ is independent of all previous points except j_t in which the system dwelled at time t. Abbreviating $j_t = i$ and $j_{t+1} = j$, the right-hand side expresses the conditional probability for the transition from $X_t = i$ to $X_{t+1} = j$ in a single time step, the so-called *single-step transition probability*,

$$p_{ij} \equiv P\bigl(X_{t+1} = j \mid X_t = i\bigr), \qquad i, j \in \Omega, \tag{12.1}$$

which is at the heart of any Markov chain. The basic properties of probability demand $\sum_{j\in\Omega} p_{ij} = 1, \forall i \in \Omega$. In the following we shall only discuss *time-homogeneous* chains, in which the transition probabilities do not depend on time,

$$p_{ij} = P\bigl(X_{t+1} = j \mid X_t = i\bigr) = P\bigl(X_1 = j \mid X_0 = i\bigr).$$

Analogously one defines the probability for the transition from state i to state j after n time steps, known as the *n-step transition probability*,

$$p_{ij}^{(n)} = P\bigl(X_n = j \mid X_0 = i\bigr), \qquad p_{ij}^{(1)} = p_{ij}.$$

For different n these are related by the Chapman–Kolmogorov equation [5]

$$p_{ij}^{(n+m)} = \sum_{k\in\Omega} p_{ik}^{(n)} p_{kj}^{(m)}. \tag{12.2}$$

How can it be elucidated? The transition from state i to state j in $n+m$ steps occurs in n steps from the initial state i to the intermediate state k with probability $p_{ik}^{(n)}$, and thence in m steps to the final state j with probability $p_{kj}^{(m)}$. The events "go from i to k in n steps" and "go from k to j in m steps" are independent. The probability for the whole transition is then obtained by the total probability formula (1.15) by summing over all intermediate states k.

According to (12.1) the transition probabilities can be organized in the so-called *stochastic matrix* $\mathcal{P} = \bigl[p_{ij}\bigr]$, while the distribution of states which the system occupies at time t, can be summarized by the row-vector

$$\boldsymbol{p}(t) = \bigl(P(X_t = j_0),\ P(X_t = j_1),\ \ldots,\ P(X_t = j_m)\bigr).$$

Since

$$p_i(1) = \sum_k p_k(0) p_{ki},$$

the distribution $\boldsymbol{p}(1)$ at time $t = 1$ can be calculated from $\boldsymbol{p}(0)$ at time $t = 0$ by simply multiplying $\boldsymbol{p}(1) = \boldsymbol{p}(0)\mathcal{P}$. Equation (12.2) then also tells us that the mapping between $\boldsymbol{p}(0)$ and the distribution $\boldsymbol{p}(t)$ at an arbitrary later time is as simple as it gets, namely

$$\boldsymbol{p}(t) = \boldsymbol{p}(0)\mathcal{P}^t, \tag{12.3}$$

where t is the power of the matrix $\mathcal{P}$. Therefore the dynamics of the probability distribution of the chain is completely determined by the probability distribution of the initial state X_0 and the one-step transition probabilities p_{ij}.

12.1.1 Long-Time Characteristics of Markov Chains

If there is a non-zero probability of arriving to any state in Ω from any other state in Ω we say that the Markov chain is *irreducible*. It is also important whether one can return to the initial state or not. A state is *periodic* if it can be revisited by paths with the numbers of steps whose greatest common divisors are greater than 1. In the opposite case, the state is aperiodic. A state is *reproducible* if we certainly return to it in finite time. If the chain is irreducible on the whole Ω and all states are aperiodic and reproducible, we call it *ergodic*.

The bombardment with all these definitions has more than just academic purpose, as it leads to the important concept of *equilibrium* distributions; see also Sect. 6.4. The equilibrium distribution is defined by $\boldsymbol{\pi} = \boldsymbol{\pi}\mathcal{P}$ or

$$\pi_j = \sum_k \pi_k p_{kj}, \qquad \sum_j \pi_j = 1. \tag{12.4}$$

Ergodic chains possess a limit distribution which is equal to the equilibrium distribution:

$$\lim_{t\to\infty} p_{ij}^{(t)} = \pi_j \qquad \forall i, j \in \Omega. \tag{12.5}$$

No condition is imposed on the initial distribution (index i), so the attribute "equilibrium" is justified. In finite Ω all states are reproducible,[1] hence eigenvectors and eigenvalues of $\mathcal{P}$ can be found: the vector $\boldsymbol{\pi}$ representing the equilibrium distribution

[1] If the chain is irreducible and the states are reproducible (in finite Ω they always are), the equilibrium distribution does not exist. If the chain is irreducible and its states are periodic, the limit (12.5) may not exist or it may depend on i: an example is the matrix $\mathcal{P} = ((0, 1), (1, 0))$ with the equilibrium distribution $\boldsymbol{\pi} = (1/2, 1/2)$, as $\boldsymbol{\pi} = \boldsymbol{\pi}\mathcal{P}$, but $\lim_{t\to\infty} \mathcal{P}^t$ does not exist.

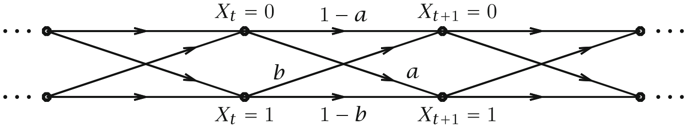
is the left eigenvector of $\mathcal{P}$ with the largest possible eigenvalue $\lambda_1 = 1$. Each initial distribution $\boldsymbol{p}$ converges to the equilibrium as $\boldsymbol{\pi} - \boldsymbol{p}\mathcal{P}^t = \mathcal{O}\bigl(|\lambda_2|^t\bigr)$, $t \to \infty$, where λ_2 ($|\lambda_2| < 1$) is the second largest eigenvalue of $\mathcal{P}$.

In finite spaces Ω the Perron-Frobenius theorem [6] guarantees that for irreducible chains there exists a vector $\boldsymbol{\pi} = (\pi_j > 0)_{j\in\Omega}$ with components

$$\pi_j = \lim_{t\to\infty} \frac{1}{t} \sum_{n=1}^{t} p_{ij}^{(n)} \qquad \forall j \in \Omega, \tag{12.6}$$

representing the equilibrium distribution *regardless of the initial state i*. For finite Ω the methods (12.4) and (12.6) to compute the equilibrium distribution are equivalent.

Example Imagine a binary communication channel shown in the figure.



Each node receives a signal (bit) and passes it on to the next node with some probability, or there may be an error in the process so that the opposite bit is forwarded. What happens at each node depends on its state: if the node receives bit 0, it is forwarded correctly with probability $P(X_{t+1}=0 \mid X_t=0) = 1 - a$, while the probability of forwarding the wrong bit, 1, is $P(X_{t+1}=1 \mid X_t=0) = a$. If it receives bit 1, the probability of correct transmittal is $P(X_{t+1}=1 \mid X_t=1) = 1 - b$, while the probability of passing on the wrong value, 0, is $P(X_{t+1}=0 \mid X_t=1) = b$. (Topologically it all looks like Pushkin's vowels and consonants!) Such error-prone communication can be modeled by a discrete-time Markov chain on the state-space $\Omega = \{0, 1\}$. The variable X_t represents the bits 0 or 1 leaving the t-th node of the channel. Let us choose $a = 0.1$ and $b = 0.2$, so the stochastic matrix is

$$\mathcal{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}.$$

Let $\boldsymbol{p}(0) = \bigl(P(X_0 = 0), P(X_0 = 1)\bigr) = (0.5, 0.5)$ be the initial state—the channel input is a symmetric mixture of bits 0 and 1. We are interested in the behavior of the chain at large "times", i.e. the distribution of states 0 and 1 at the output of the channel containing very many nodes. We could use formula (12.3),

$$\boldsymbol{p}(t) = \boldsymbol{p}(0)\,\mathcal{P}^t = (0.5, 0.5) \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}^t,$$

but it requires us to compute the t-th power of $\mathcal{P}$. Fortunately, our linear algebra professor tells us that $\mathcal{P}^t$ can be written as $\mathcal{P}^t = \lambda_1^t \mathcal{P}_1 + \lambda_2^t \mathcal{P}_2$, where

$$\mathcal{P}_1 = \frac{1}{\lambda_1 - \lambda_2}\left(\mathcal{P} - \lambda_2 I\right), \qquad \mathcal{P}_2 = \frac{1}{\lambda_2 - \lambda_1}\left(\mathcal{P} - \lambda_1 I\right),$$

and λ_1, λ_2 are the eigenvalues of $\mathcal{P}$. They can be calculated by solving the secular equation $\det\left(\lambda I - \mathcal{P}\right) = 0$, whence $\lambda_1 = 1$ and $\lambda_2 = 1 - a - b$. Therefore

$$\mathcal{P}^t = \mathcal{P}_1 + (1-a-b)^t \mathcal{P}_2 = \frac{1}{a+b}\left\{\begin{pmatrix} b & a \\ b & a \end{pmatrix} + (1-a-b)^t \begin{pmatrix} a & -a \\ -b & b \end{pmatrix}\right\}.$$

By using the specified parameters a and b we get

$$\mathcal{P}^t = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}^t = \frac{1}{3}\begin{pmatrix} 2 + 0.7^t & 1 - 0.7^t \\ 2 - 2\cdot 0.7^t & 1 + 2\cdot 0.7^t \end{pmatrix}.$$

Hence the distribution of states at the tth node is

$$\boldsymbol{p}(t) = \big(P(X_t = 0),\ P(X_t = 1)\big) = \left(\frac{2}{3} - \frac{0.7^t}{6}, \frac{1}{3} + \frac{0.7^t}{6}\right)$$

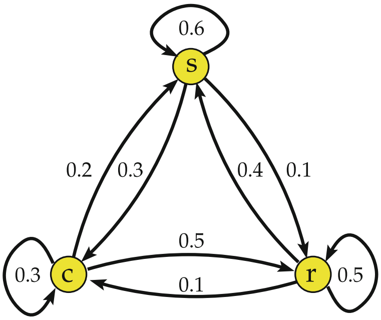
All we need to do now is $\lim_{t\to\infty} \boldsymbol{p}(t) = (2/3, 1/3)$, amounting to

$$P(X_\infty = 0) = \frac{2}{3}, \qquad P(X_\infty = 1) = \frac{1}{3}. \tag{12.7}$$

It is no wonder that the chain "drifts" to a regime where the probability for output 0 is larger than the probability for 1, since $b > a$. We could have even guessed the values (12.7) by reasoning that $P(X_\infty = 0) : P(X_\infty = 1) = b : a = 2 : 1$ and $P(X_\infty = 0) + P(X_\infty = 1) = 1$. ◁

Example (Adapted from [7].) Can the long-time analysis of a Markov chain be used to predict weather? Imagine a simple model that knows only three weather conditions: sunny (s), cloudy (c) and rainy (r), so the state space is $\Omega = \{\mathrm{s}, \mathrm{c}, \mathrm{r}\}$. Let the time step be one day. Assume that the probabilities that *tomorrow* will be sunny, cloudy or rainy if the weather *today* is sunny, are 0.6, 0.3 and 0.1; the probabilities that the next day will be sunny, cloudy or rainy if the weather today is cloudy, are 0.2, 0.3 and 0.5; the probabilities of having sun, clouds or rain tomorrow if it is raining today, are 0.4, 0.1 and 0.5, as shown by the graph. We arrange all nine conditional probabilities in the stochastic matrix where the lines and rows correspond to today's and tomorrow's weather conditions, respectively:

$$\mathcal{P} = \begin{array}{c} \begin{array}{ccc} \mathrm{s}\downarrow & \mathrm{c}\downarrow & \mathrm{r}\downarrow \end{array} \\ \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.3 & 0.5 \\ 0.4 & 0.1 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} \\ \leftarrow \mathrm{s} \\ \leftarrow \mathrm{c}\cdot \\ \leftarrow \mathrm{r} \end{array} \tag{12.8}$$



What is the probability of having rain in two days if it is cloudy today? It is given by the Chapman–Kolmogorov equation (12.2):

$$\left(\mathcal{P}^2\right)_{\mathrm{cr}} = p_{\mathrm{cr}}^{(2)} = \sum_{x\in\Omega} p_{\mathrm{cx}}\, p_{\mathrm{xr}} = p_{\mathrm{cs}}\, p_{\mathrm{sr}} + p_{\mathrm{cc}}\, p_{\mathrm{cr}} + p_{\mathrm{cr}}\, p_{\mathrm{rr}} = 0.42.$$

And what is the probability of rain three days, five days ... from now if it is cloudy today? The answer always sits at the same spot: in the matrix element at the second-row, third-column crossing of the matrices $\mathcal{P}^3$, $\mathcal{P}^5$, and so on:

$$\mathcal{P}^3 = \begin{pmatrix} 0.436 & 0.248 & 0.316 \\ 0.436 & 0.216 & \underline{0.348} \\ 0.452 & 0.232 & 0.316 \end{pmatrix}, \quad \mathcal{P}^5 = \begin{pmatrix} 0.440 & 0.235 & 0.325 \\ 0.443 & 0.235 & \underline{0.322} \\ 0.441 & 0.236 & 0.322 \end{pmatrix}, \quad \ldots$$

We see that the columns of ever higher powers of $\mathcal{P}$ become more and more constant (independent of the rows): at long times we approach the equilibrium distribution regardless of the initial state (Fig. 12.1). But formula (12.6) is impractical as it requires

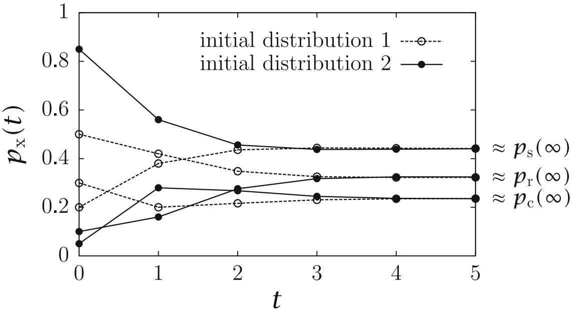


Fig. 12.1 Time evolution of three components of the probability distribution $\boldsymbol{p}(t)$ at two different initial distributions, $\boldsymbol{p}(0) = (0.2, 0.3, 0.5)$ and $\boldsymbol{p}(0) = (0.85, 0.05, 0.1)$. In both cases the Markov chain with the stochastic matrix (12.8) converges to the stable distribution (12.9) in just a few steps

us to compute high powers of $\mathcal{P}$. Besides, the series has a slow convergence. It is therefore preferable to solve the system (12.4):

$$\begin{aligned}\pi_\mathrm{s} &= 0.6\pi_\mathrm{s} + 0.2\pi_\mathrm{c} + 0.4\pi_\mathrm{r},\\ \pi_\mathrm{c} &= 0.3\pi_\mathrm{s} + 0.3\pi_\mathrm{c} + 0.1\pi_\mathrm{r},\\ \pi_\mathrm{r} &= 0.1\pi_\mathrm{s} + 0.5\pi_\mathrm{c} + 0.5\pi_\mathrm{r},\end{aligned}$$

together with $\pi_\mathrm{s} + \pi_\mathrm{c} + \pi_\mathrm{r} = 1$. (The normalization condition is needed since only two of these equations are independent. Thus we get three equations for three unknowns.) The solution of the system is the equilibrium distribution

$$\boldsymbol{\pi} = \boldsymbol{p}(\infty) \approx (0.441, 0.235, 0.324). \tag{12.9}$$

Of course the probabilities $p_\mathrm{s}(\infty)$, $p_\mathrm{c}(\infty)$ and $p_\mathrm{r}(\infty)$ also express the probabilities for sunny, cloudy or rainy weather over a longer time period. If the values (12.9) were in fact measured, say, over a period of one year, they could even be used to calibrate the model—i.e. the elements of the stochastic matrix—so that it would always converge to the desired end configuration. ◁

12.2 Continuous-Time Markov Processes

In continuous-time Markov processes the transitions between the states $X(t)$ in a dynamical system do not occur in discrete time jumps but rather in a continuous, smooth time evolution. In the following—the notation mostly follows [8]—we discuss continuous-time processes in which also the states themselves can be represented on the whole real axis, i.e. by a random variable $X(t) \in \mathbb{R}$, $t \geq t_0$.

As we are dealing with continuous random variables, the set of variables $X(t_1), X(t_2), \ldots, X(t_n)$ can be assigned a joint probability density

$$\begin{aligned} &f_n^{(1)}\big(x_n, t_n; x_{n-1}, t_{n-1}; \ldots; x_1, t_1 \,|\, x_0, t_0\big)\,\mathrm{d}x_n\,\mathrm{d}x_{n-1}\cdots\mathrm{d}x_1\\ &\qquad = P\big(X(t_i) \in [x_i, x_i + \mathrm{d}x_i), i = 1, 2, \ldots, n \,|\, X(t_0) = x_0\big),\end{aligned}$$

where the subscript n and the superscript (1) on f indicate that the n values $\{x_1, x_2, \ldots, x_n\}$ at times $\{t_1, t_2, \ldots, t_n\}$ on the left of the | sign depend on *one* value x_0 at time t_0 at its right. This applies to general time evolution; but Markov processes are "memoryless" and therefore

$$f_1^{(i)}\big(x_i, t_i \,|x_{i-1}, t_{i-1};\ \ldots;\ x_0, t_0\big) = f_1^{(1)}\big(x_i, t_i \,|x_{i-1}, t_{i-1}\big) \equiv f\big(x_i, t_i \,|x_{i-1}, t_{i-1}\big).$$

In general, a state x_i at time t_i may depend on i states $\{x_{i-1}, x_{i-2}, \ldots, x_0\}$ at all previous times $\{t_{i-1}, t_{i-2}, \ldots, t_0\}$, while in Markov processes only the state immediately preceding it is relevant, i.e. x_{i-1} at time t_{i-1}. Hence the joint probability density can be written as a product of densities for individual transitions:

$$f_n^{(1)}\bigl(x_n, t_n; x_{n-1}, t_{n-1}; \ldots; x_1, t_1 \,|\, x_0, t_0\bigr) = \prod_{i=1}^{n} f\bigl(x_i, t_i \,|x_{i-1}, t_{i-1}\bigr).$$

It follows that for *arbitrary* time t_2 on the interval $t_1 \le t_2 \le t_3$, Chapman–Kolmogorov equation applies:

$$f\bigl(x_3, t_3 \,|\, x_1, t_1\bigr) = \int_{-\infty}^{\infty} f\bigl(x_3, t_3 \,|\, x_2, t_2\bigr) f\bigl(x_2, t_2 \,|\, x_1, t_1\bigr)\,\mathrm{d}x_2.$$

In continuous language this equation conveys the same message as its discrete analogue (12.2): the probability $f(x_3, t_3 \,|\, x_1, t_1)\,\mathrm{d}x_3$ for the transition from the state x_1 at time t_1 to some state on the interval $[x_3, x_3 + \mathrm{d}x_3)$ at time t_3 is the sum of probabilities that this transition occurred through a state on any interval $[x_2, x_2 + \mathrm{d}x_2)$ at intermediate time t_2.

12.2.1 Markov Propagator and Its Moments

The key quantity embodying the actual step between two states in a very short time Δt is the *Markov propagator*

$$\Xi(\Delta t; x, t) = X(t + \Delta t) - X(t), \quad \text{given } X(t) = x. \tag{12.10}$$

The propagator tells us the state of the process at time $t + \Delta t$, if at time t it was in state x: the new state will be $x + \Xi(\Delta t; x, t)$. The propagator depends on three real parameters x, t and Δt, but conventionally the latter is given the most prominent spot: namely, Ξ may be independent of x and t, but it *must* depend on Δt. Since $X(t)$ is a random variable, the propagator is also a random variable, so it can be assigned its own *propagator density function* Π with the definition

$$\Pi\bigl(\xi \,|\, \Delta t; x, t\bigr)\,\mathrm{d}\xi = P\bigl(\Xi(\Delta t; x, t) \in [\xi, \xi + \mathrm{d}\xi)\bigr).$$

The density Π allows us to define the *propagator moment functions*

$$E\bigl[\Xi^n(\Delta t; x, t)\bigr] = \int_{-\infty}^{\infty} \xi^n \,\Pi(\xi \,|\, \Delta t; x, t)\,\mathrm{d}\xi = M_n(x, t)\,\Delta t + \mathcal{O}(\Delta t),$$

where $n = 1, 2, \ldots$ The moment functions M_n and the density f are related by [8]

$$\frac{\partial}{\partial t} f\bigl(x, t \,|\, x_0, t_0\bigr) = \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} \frac{\partial^n}{\partial x^n}\bigl[M_n(x, t) f(x, t \,|\, x_0, t_0)\bigr]. \tag{12.11}$$

This is the *Kramers–Moyal* partial differential equation of infinite order describing the time evolution of $f(x, t \mid x_0, t_0)$ at fixed x_0 and t_0, for which all moments M_n and the initial condition $f(x, t_0 \mid x_0, t_0) = \delta(x - x_0)$ must be known.

If we choose a short enough Δt, the propagator $\Xi(\Delta t; x, t)$ can be composed of n propagators Ξ_i with which the process proceeds in time from state $X(t_0) = x$ to states $X(t_1), X(t_2), \ldots, X(t_n)$ in steps of length $\Delta t/n$. These, however, can be made so small that during a step the value of x remains almost constant:

$$\Xi(\Delta t; x, t) = \sum_{i=1}^{n} \Xi_i\left(\frac{\Delta t}{n}; X(t_{i-1}), t_{i-1}\right) \approx \sum_{i=1}^{n} \Xi_i\left(\frac{\Delta t}{n}; x, t\right). \qquad (12.12)$$

In this approximation all these steps become mutually independent, therefore $E\big[\Xi(\Delta t; x, t)\big] = n\, E\big[\Xi(\Delta t/n; x, t)\big]$ and $\mathrm{var}\big[\Xi(\Delta t; x, t)\big] = n\, \mathrm{var}\big[\Xi(\Delta t/n; x, t)\big]$. It is easy to show[2] that in this case the expected value and variance of the Markov propagator must be proportional to the length of the time step:

$$E\big[\Xi(\Delta t; x, t)\big] = A(x, t)\, \Delta t + \mathcal{O}(\Delta t),$$
$$\mathrm{var}\big[\Xi(\Delta t; x, t)\big] = D(x, t)\, \Delta t + \mathcal{O}(\Delta t),$$

where the functions A and D do not depend on Δt. The key consideration follows. In the mentioned approximation the right-hand side of (12.12) is a sum of independent and identically distributed random variables, so by the central limit theorem the variable Ξ on its left is normally distributed,

$$\Xi(\Delta t; x, t) \sim N\big(A(x, t)\Delta t,\ D(x, t)\Delta t\big). \qquad (12.13)$$

The functions A and D are the *characterizing functions* of the Markov process. Do not confuse them with the *characteristic* functions of Sect. B.3! Due to obvious reasons A is called the *drift function* and D is known as the *diffusion function*. The propagator density of a continuous Markov process is therefore

$$\Pi\big(\xi \mid \Delta t; x, t\big) = \frac{1}{\sqrt{2\pi D(x, t)\, \Delta t}} \exp\left(-\frac{\big(\xi - A(x, t)\, \Delta t\big)^2}{2D(x, t)\, \Delta t}\right).$$

From here and from (12.13) we see[3] that A and D are equal to the moment functions M_1 and M_2, respectively, while it turns out that higher moments vanish:

$$M_1(x, t) = A(x, t), \qquad M_2(x, t) = D(x, t), \qquad M_n(x, t) = 0, \quad n \geq 3.$$

[2] We are referring to a simple lemma: is $g(z)$ is a smooth function of z satisfying $g(z) = ng(z/n)$ for any positive integer n, it holds that $g(z) = Cz$, where C does not depend on z.

[3] The expected value is $E[\Xi(\Delta t; x, t)] = M_1(x, t)\Delta t + \mathcal{O}(\Delta t)$, whence $M_1(x, t) = A(x, t)$. The variance is given by $\mathrm{var}[\Xi(\Delta t; x, t)] = E[\Xi^2(\Delta t; x, t)] - (E[\Xi(\Delta t; x, t)])^2 = M_2(x, t)\Delta t + \mathcal{O}(\Delta t) - (M_1(x, t)\Delta t + \mathcal{O}(\Delta t))^2 = M_2(x, t)\Delta t + \mathcal{O}(\Delta t)$, therefore $M_2(x, t) = D(x, t)$.

What is left of (12.11), then, is just the *Fokker–Planck equation*

$$\frac{\partial}{\partial t} f(x,t\,|\,x_0,t_0) = -\frac{\partial}{\partial x}\big[A(x,t)f(x,t\,|\,x_0,t_0)\big] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\big[D(x,t)f(x,t\,|\,x_0,t_0)\big],$$

which is a partial differential equation of the second order that needs to be solved with the initial condition $f(x,t_0\,|\,x_0,t_0) = \delta(x-x_0)$.

12.2.2 Time Evolution of the Moments

The time evolution of the variable X is determined by the Markov propagator. But how do its *expected value* and *variance* evolve? And what is the time evolution of the expected value and variance of the random variable

$$S(t) = \int_{t_0}^{t} X(t')\,\mathrm{d}t', \tag{12.14}$$

which is called the *integral of the Markov process?* The answers to both questions for any continuous-time Markov process are given by ordinary differential equations, which we list without proof: for their derivation see e.g. [8]. The time evolution of the expected value and variance of $X(t)$ is given by the equations

$$\frac{\mathrm{d}}{\mathrm{d}t} E\big[X(t)\big] = E\big[A\big(X(t),t\big)\big], \qquad t \ge t_0, \tag{12.15}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{var}\big[X(t)\big] = 2\Big(E\big[X(t)A\big(X(t),t\big)\big] - E\big[X(t)\big]E\big[A\big(X(t),t\big)\big]\Big) + E\big[D\big(X(t),t\big)\big], \qquad t \ge t_0, \tag{12.16}$$

with initial conditions $E\big[X(t_0)\big] = x_0$ and $\mathrm{var}\big[X(t_0)\big] = 0$. The time evolution of the corresponding moments of $S(t)$ for $t \ge t_0$ is given by

$$E\big[S(t)\big] = \int_{t_0}^{t} E\big[X(t')\big]\,\mathrm{d}t', \tag{12.17}$$

$$\mathrm{var}\big[S(t)\big] = 2\int_{t_0}^{t} \mathrm{cov}\big[S(t'),X(t')\big]\,\mathrm{d}t', \tag{12.18}$$

where the integrand in (12.18) is the solution of the auxiliary equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{cov}\big[S(t),X(t)\big] = \mathrm{var}\big[X(t)\big] + E\big[S(t)A\big(X(t),t\big)\big] - E\big[S(t)\big]E\big[A\big(X(t),t\big)\big]$$

with the initial condition $\mathrm{cov}[S(t_0),X(t_0)] = 0$.

12.2.3 Wiener Process

The Wiener process is a Markov process in which the drift and diffusion functions are constant, i.e. independent of x and t: $A(x,t) = A$ and $D(x,t) = D \geq 0$. In this case, not much is left of (12.15) and (12.16); their solutions are

$$E\big[X(t)\big] = x_0 + A(t - t_0), \qquad t \geq t_0, \tag{12.19}$$

$$\mathrm{var}\big[X(t)\big] = D(t - t_0), \qquad t \geq t_0, \tag{12.20}$$

while from (12.17) and (12.18) we obtain

$$E\big[S(t)\big] = x_0(t - t_0) + \tfrac{1}{2}A(t - t_0)^2, \qquad t \geq t_0, \tag{12.21}$$

$$\mathrm{var}\big[S(t)\big] = \tfrac{1}{3}D(t - t_0)^3, \qquad t \geq t_0. \tag{12.22}$$

The Fokker–Planck equation also simplifies significantly,

$$\frac{\partial}{\partial t} f(x,t\,|\,x_0,t_0) = -A\frac{\partial}{\partial x} f(x,t\,|\,x_0,t_0) + \frac{D}{2}\frac{\partial^2}{\partial x^2} f(x,t\,|\,x_0,t_0).$$

It is solved by the initial condition $f(x,t|x_0,t_0) = \delta(x - x_0)$, and its solution is

$$f(x,t\,|\,x_0,t_0) = \frac{1}{\sqrt{2\pi D(t - t_0)}}\exp\left(-\frac{\big(x - x_0 - A(t - t_0)\big)^2}{2D(t - t_0)}\right). \tag{12.23}$$

The Wiener process is therefore a Markov process described by a normally distributed variable with mean $x_0 + A(t - t_0)$ and variance $D(t - t_0)$, as shown in Fig. 12.2. The corresponding phenomenon in nature is *self-diffusion*, in which a particle with mass M diffuses among a large ensemble of equally heavy particles (see also Sect. 12.2.4 and Example on p. 320).

Let us check our understanding of the Wiener process by a simple computer simulation of the Fokker–Planck equation, where the exact functional form of $A(x,t)$ and $D(x,t)$ can be freely chosen. We set $t = t_0 = 0$, $x = x_0 = 0$ and $s = s_0 = 0$, and choose a small enough Δt. Then we repeat until desired:

1. Draw a value $\mathcal{N} \sim N(0,1)$.
2. $s \leftarrow s + x\,\Delta t$
3. $x \leftarrow x + A(x,t)\Delta t + \mathcal{N}\sqrt{D(x,t)\Delta t}$
4. $t \leftarrow t + \Delta t$
5. Write t, $x(t)$ and $s(t)$, and return to 1.

An example of such a simulation until $t = 10$ is shown in Fig. 12.3.

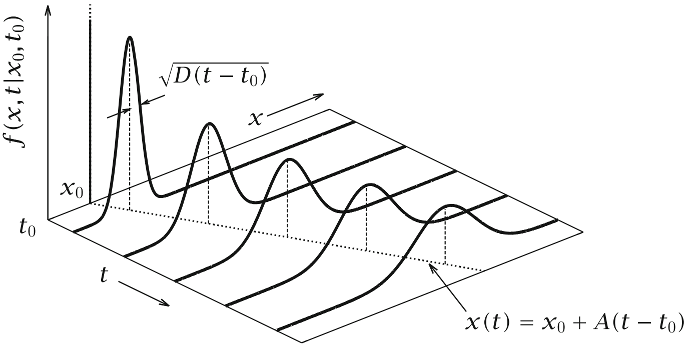


Fig. 12.2 Time evolution of the probability density of a continuous-time Markov process with constant drift and diffusion functions, $A(x,t) = A$ and $D(x,t) = D \geq 0$. The initial condition, the $\delta(x - x_0)$ "function", gradually spreads into an ever broader normal distribution which "drifts" along the straight line given by $x(t) = x_0 + A(t - t_0)$

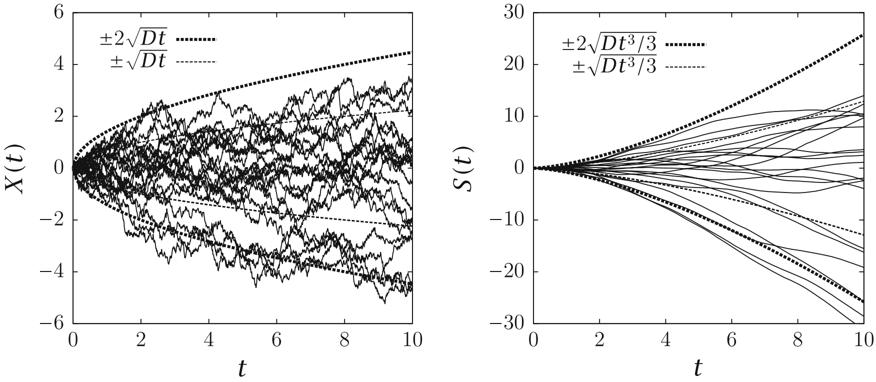


Fig. 12.3 Simulation of the Wiener process with initial value $x_0 = 0$, time step length $\Delta t = 0.01$, drift function $A(x,t) = 0$ and diffusion function $D(x,t) = 0.5$. [Left] Some realizations of the random process that does not "drift" anywhere on average, as $E[X(t)] = 0$ by (12.19). The thin and thick curves denote the boundaries of one and two standard deviations, as dictated by (12.20). [Right] Some integrals of the process with mean zero (see (12.21)) and cubic increase of variance (12.22)

12.2.4 Ornstein–Uhlenbeck Process

The Ornstein–Uhlenbeck process is a special case of a continuous Markov process where the drift function has the form $A(x,t) = -kx$, $k > 0$, while the diffusion function is independent of x and t, $D(x,t) = D \geq 0$. The evolution equation (12.15) therefore has the form $\dot{x} = -kx$ with the initial condition $x(t_0) = x_0$. Its solution is

$$E\big[X(t)\big] = x_0\,\mathrm{e}^{-k(t-t_0)}, \qquad t \geq t_0. \tag{12.24}$$

Equation (12.16) for the variance of X is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{var}\big[X(t)\big] &= 2\Big(E\big[X(t)\big(-kX(t)\big)\big] - E\big[X(t)\big]E\big[-kX(t)\big]\Big) + E[D] \\ &= -2k\Big(E\big[X^2(t)\big] - E\big[X(t)\big]^2\Big) + D = -2k\,\mathrm{var}\big[X(t)\big] + D,\end{aligned}$$

and the initial condition is $\mathrm{var}[X(t)] = 0$, thus

$$\mathrm{var}\big[X(t)\big] = \frac{D}{2k}\left(1 - \mathrm{e}^{-2k(t-t_0)}\right), \qquad t \geq t_0. \tag{12.25}$$

In this case the Fokker–Planck equation has the form

$$\frac{\partial}{\partial t} f(x, t \,|\, x_0, t_0) = k\frac{\partial}{\partial x}\big[x f(x, t \,|\, x_0, t_0)\big] + \frac{D}{2}\frac{\partial^2}{\partial x^2} f(x, t \,|\, x_0, t_0)$$

and needs to be solved with the initial condition $f(x, t \,|\, x_0, t_0) = \delta(x - x_0)$. The probability density $f(x, t \,|\, x_0, t_0)$ that solves this equation corresponds to a normally distributed random variable

$$X(t) \sim N\left(x_0\,\mathrm{e}^{-k(t-t_0)},\ \frac{D}{2k}\left(1 - \mathrm{e}^{-2k(t-t_0)}\right)\right).$$

(You can check this, with some effort, by writing the density of the normal distribution with specified mean and variance as in (12.23) and insert it in the Fokker–Planck equation.) In the large-t limit this means

$$X(t) \sim N\left(0,\ \frac{D}{2k}\right), \qquad t \to \infty,$$

in other words, convergence to a stable distribution:

$$\lim_{t\to\infty} f(x, t \,|\, x_0, t_0) = f_{\mathrm{stab}}(x) = \frac{1}{\sqrt{\pi D/k}}\exp\left(-\frac{x^2}{D/k}\right).$$

We are witnessing a process $X(t)$ with probability density $f(x, t \,|\, x_0, t_0)$ whose expected value exponentially approaches zero, while its variance in the limit $t \to \infty$ stabilizes at the value $D/(2k)$, as shown in Fig. 12.4.

By the same token we solve (12.17) for the expected value of the integral of the Markov process, $S(t)$, as well as (12.18) for its variance. For $t \geq t_0$ we get

$$E\big[S(t)\big] = \frac{x_0}{k}\left(1 - \mathrm{e}^{-k(t-t_0)}\right), \tag{12.26}$$

$$\mathrm{var}\big[S(t)\big] = \frac{D}{k^2}\left\{(t - t_0) - \frac{2}{k}\left(1 - \mathrm{e}^{-k(t-t_0)}\right) + \frac{1}{2k}\left(1 - \mathrm{e}^{-2k(t-t_0)}\right)\right\}. \tag{12.27}$$

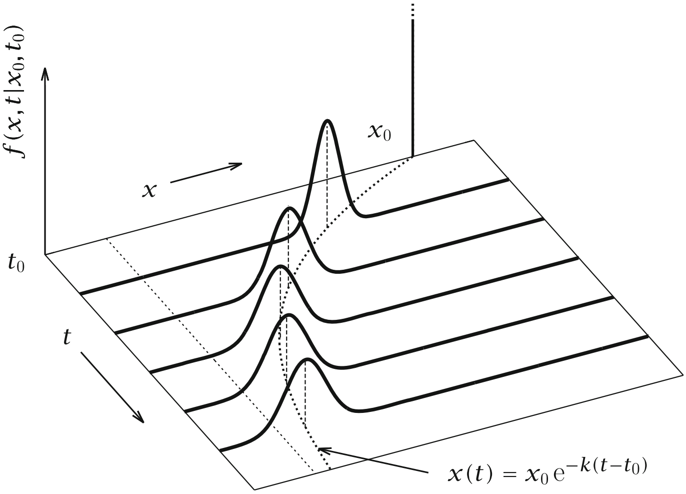


Fig. 12.4 Time evolution of the probability density of a continuous Markov process with the drift function $A(x,t) = -kx$, $k > 0$, and the diffusion function $D(x,t) = D \geq 0$. The "center of gravity" of the initial condition $\delta(x - x_0)$ gradually slips towards zero, yet the width of the distribution no longer increases: the last time slice is already $\approx f_{\text{stab}}(x)$

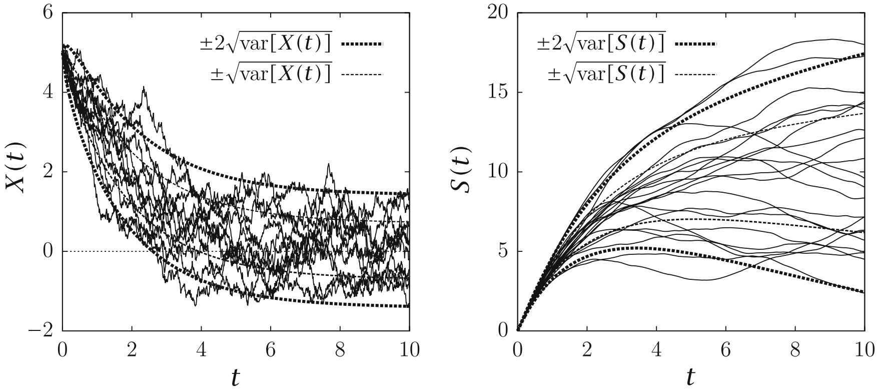


Fig. 12.5 Computer simulation of the Ornstein–Uhlenbeck process with initial value $x_0 = 5$, time step $\Delta t = 0.01$, drift function $A(x,t) = -kx$, $k = 0.5$ and diffusion function $D(x,t) = 0.5$. [Left] Some realizations of the process that (on average) exponentially "drifts" towards zero, as dictated by (12.24). The thin and thick dashed lines indicate the boundaries of one and two standard deviations according to (12.25). At large t the standard deviation settles at $\sqrt{D/(2k)} \approx 0.71$ (stable distribution). [Right] Some integrals of the process with a mean that attains the value $x_0/k = 10$ in the limit $t \to \infty$ according to (12.26), and the variance that changes according to (12.27)

Let us again try to understand what (12.24–12.27) are saying by a computer simulation, based on the algorithm from p. 317. We choose $k = D = 0.5$, $x_0 = 5$ and $\Delta t = 0.01$. Sample results are shown in Fig. 12.5.

Long example Wanderings of a heavy macroscopic particle in a gas (fluid) of smaller, lighter particles—Brownian motion—can be approximately treated as a continuous Markov process. In order to see this, we first reinterpret the propagator (12.10) as a small change

$$\Delta X(\Delta t; x, t) = X(t+\Delta t) - X(t), \quad \text{given } X(t) = x.$$

By (12.13), ΔX is a normally distributed random variable with mean $A(x,t)\Delta t$ and variance $D(x,t)\Delta t$. We use the relation $Y \sim N(0,1) \Longleftrightarrow aY + b \sim N(b, a^2)$ which is valid for arbitrary real constants a and b and can be proven by using methods of Sect. 2.7. Identifying $b = A(x,t)\Delta t$ and $a^2 = D(x,t)\Delta t$ leads to $\Delta X \sim \sqrt{D(x,t)\Delta t}\, N(0,1) + A(x,t)\Delta t$, therefore

$$X(t+\Delta t) = X(t) + A(x,t)\Delta t + \sqrt{D(x,t)\Delta t}\,\mathcal{N}, \tag{12.28}$$

where $\mathcal{N} \sim N(0,1)$. We have derived a specific form of the *Langevin equation*, which is easy to code as it explicitly expresses the random variable $X(t+\Delta t)$ in terms of the variables $X(t)$ and $\mathcal{N}$.

The analysis of Brownian motion, discovered in the early 19th century, has a long history. Einstein's approach was to treat the *coordinate* of the particles as a Wiener process [9], while Langevin [10] placed his bet on their *velocity* as the key quantity. A spherical particle with radius R and mass M, moving with velocity v in a fluid with viscosity η, obeys Newton's law $M\dot{v} = -\gamma v$ or

$$v(t+\Delta t) = v(t) - (\gamma/M)v(t)\Delta t,$$

where $\gamma = 6\pi\eta R$ is the linear drag coefficient. (In both Langevin equation (12.28) and Netwon's law the time interval Δt is assumed to be infinitesimally small.) The "driving" term $-\gamma v(t)\Delta t$ on the right represents the average linear momentum transferred to the wandering particle by the particles of the fluid (average force over Δt). But in general this momentum transfer fluctuates about its average. With this in mind we augment Newton's law by a term that provides the process with this kind of jitter:

$$V(t+\Delta t) = V(t) - (\gamma/M)V(t)\Delta t + \sqrt{c\Delta t}\,\mathcal{N}, \tag{12.29}$$

where $\mathcal{N} \sim N(0,1)$ and c is a positive constant which needs to be determined. We have denoted the velocity by an upper-case letter as we are dealing with a random variable. By comparing (12.29) and (12.28) we realize that Brownian motion can be understood as a continuous Markov process in which the role of the generic random variable $X(t)$ is played by the physical velocity, $V(t)$. The process has the Ornstein-Uhlenbeck form with the drift and diffusion functions

$$\begin{aligned} A(v,t) &= -(\gamma/M)\,v, \\ D(v,t) &= c. \end{aligned}$$

With the initial condition $V(t_0) = v_0$, (12.24) and (12.25) immediately give us the expected value of the velocity and its variance for $t \geq t_0$:

$$E\big[V(t)\big] = v_0\, \mathrm{e}^{-(\gamma/M)t}, \tag{12.30}$$

$$\mathrm{var}\big[V(t)\big] = \frac{cM}{2\gamma}\left(1 - \mathrm{e}^{-2(\gamma/M)t}\right). \tag{12.31}$$

This is precisely what we observe in Fig. 12.5 (left), where on the ordinate axis one should imagine the velocity $V(t)$ instead of the generic variable $X(t)$: a particle that starts moving in the fluid with velocity v_0 at time zero, *on average* slows down exponentially according to (12.30), but the velocity distribution settles into a stable form with the variance (12.31).

The path (coordinate) of the particle is the integral of its velocity over time, $s = \int v(t)\,\mathrm{d}t$, so at the level of random variables we may resort to (12.14), where $X(t)$ is replaced by $V(t)$, and formulas (12.26) and (12.27):

$$E\big[S(t)\big] = \frac{v_0 M}{\gamma}\left(1 - \mathrm{e}^{-(\gamma/M)t}\right), \tag{12.32}$$

$$\mathrm{var}\big[S(t)\big] = \frac{cM^2}{\gamma^2}\left(t - \frac{2M}{\gamma}\left(1 - \mathrm{e}^{-(\gamma/M)t}\right) + \frac{M}{2\gamma}\left(1 - \mathrm{e}^{-2(\gamma/M)t}\right)\right). \tag{12.33}$$

The content of these equations is expressed by Fig. 12.5 (right): the path traveled by the particle with non-zero initial velocity v_0 keeps increasing according to (12.32) at short times, but *on average* it attains the terminal value $v_0 M/\gamma$, about which is straggles with variance (12.33).

Asymptotically we have

$$\lim_{t\to\infty} E\big[V(t)\big] = 0, \qquad \lim_{t\to\infty} \mathrm{var}[V(t)\big] = \frac{cM}{2\gamma}, \tag{12.34}$$

which aids us in determining the constant c. After a long time the particle is in thermodynamic equilibrium with the fluid at temperature T. We can then expect—that was Langevin's key assumption—that $V(t \to \infty)$ is a normally distributed random variable with mean zero and variance $k_\mathrm{B}T/M$, since

$$\int_{-\infty}^{\infty} v_x^2 \sqrt{\frac{M}{2\pi k_\mathrm{B}T}} \exp\left(-\frac{Mv_x^2}{2k_\mathrm{B}T}\right) \mathrm{d}v_x = \frac{k_\mathrm{B}T}{M}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u^2\, \mathrm{e}^{-u^2/2}\,\mathrm{d}u = \frac{k_\mathrm{B}T}{M}.$$

According to (12.34) we may therefore equate

$$\frac{cM}{2\gamma} = \frac{k_\mathrm{B}T}{M}$$

or $c = 2\gamma k_{\mathrm{B}} T/M^2$. It follows that

$$\mathrm{var}\big[S(t)\big] = (2k_{\mathrm{B}}T/\gamma)\,t, \qquad t \gg M/\gamma.$$

We have obtained the famous result that the variance of the particle's position at long times linearly increases with time and that it depends on the temperature and viscosity of the fluid. The main deficiency of this approach, of course, is the treatment of collisions, in which velocity actually changes very rapidly, as continuous ("smooth") processes. This is why in the described approximation the effective deviation depends neither on the mass of the Brownian particle, M, nor on the mass of the fluid particles, m. A much improved calculation, in which Brownian motion is analyzed as a *discrete-time* Markov process with *continuous states* (see, for example, Sect. 4.5 in [8]) reveals these delicate dependencies as well. One then obtains

$$\mathrm{var}\big[S(t)\big] = \frac{1}{2\rho R^2}\left(\frac{\pi k_{\mathrm{B}}T}{2m}\right)^{1/2} t, \qquad t \gg \frac{M}{4\rho R^2}\sqrt{\frac{\pi}{2mk_{\mathrm{B}}T}},$$

where ρ is the average particle density of the gas. The dependence of the variance on m is particularly intriguing. The velocity of the gas particles has a Maxwell distribution and, as we have seen in the Example on p. 105, all its characteristic velocities (mode, expected value and effective deviation) exhibit the $1/\sqrt{m}$ dependence: the lighter the gas particles, the more efficient they are in "kicking" the heavier particle and in dispersing its position. ◁

References

1. E. Parzer, *Stochastic Processes* (Holden-Day, San Francisco, 1962)
2. A.A. Markov, An example of statistical investigation of the text "Eugene Onegin" concerning the connection of samples in chains. Sci. Context **19**, 591 (2006). English translation from original Russian
3. S. Meyn, R.L. Tweedie, *Markov Chains and Stochastic Stability* (Springer, Berlin, 1995)
4. D.W. Stroock, *An Introduction to Markov Processes*, 2nd edn. (Springer, Berlin, 2014)
5. A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 3rd edn. (McGraw-Hill, New York, 1991)
6. C. Meyer, *Matrix Analysis and Applied Linear Algebra* (SIAM, Philadelphia, 2000)
7. B. Hayes, First links in the Markov chain. Am. Sci. **101**, 92 (2013)
8. D.T. Gillespie, *Markov Processes. An Introduction for Physical Scientists* (Academic Press Inc, Boston, 1992)
9. A. Einstein, On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat. Ann. Phys. **17**, 549 (1905)
10. P. Langevin, Sur la théorie du mouvement brownien, C. R. Acad. Sci. (Paris) **146**, 530 (1908). See also the English translation D.S. Lemons, A. Gythiel, Paul Langevin's 1908 paper "On the theory of Brownian motion". Am. J. Phys. **65**, 1079 (1997)

Chapter 13
The Monte–Carlo Method

Abstract The Monte–Carlo method is introduced as a generic tool for the solution of mathematical physics models by means of computer simulation. These problems range from simple one-dimensional integration to sophisticated multi-dimensional models involving elaborate geometries and complex system states. A historical introduction and an exposition of the basic idea are followed by a basic treatment of numerical integration and discussing methods of variance reduction like importance sampling and use of quasi-random sequences. Markov-chain Monte Carlo is presented as a powerful method to generate random numbers according to arbitrary, even extremely complicated distributions. A specific implementation in the form of the Metropolis–Hastings algorithm is offered.

The Monte Carlo (MC) method or *simulation* is a generic name for any procedure in which drawing random numbers and statistical samples allows us to *approximately* evaluate some mathematical quantity or expression, for example, a definite integral or a system of equations, but it can also be applied to much more general problems of mathematical physics [1]. The emphasis is on the word 'approximately': the quality of the solution depends on the sample size one can afford. Yet from the viewpoint of feasibility and precision as compared to standard numerical methods—in particular in multi-dimensional integration with complicated integration boundaries and in handling complex mathematical models—the Monte–Carlo method offers the only reasonable approach.

13.1 Historical Introduction and Basic Idea

The French naturalist Georges–Louis Leclerc, count de Buffon (1707–1788), has shown how throwing a needle onto a mesh of uniformly spaced parallel lines allows one to estimate π. Let the needle length be L and the line spacing be $D \geq L$. Take y to be the shortest distance from the needle center to the closest line and ϕ the acute angle of the needle with respect to the lines. At each throw any distance $0 \leq y \leq D/2$ and any angle $0 \leq \phi \leq \pi/2$ are equally probable, which corresponds to the uniform

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_13

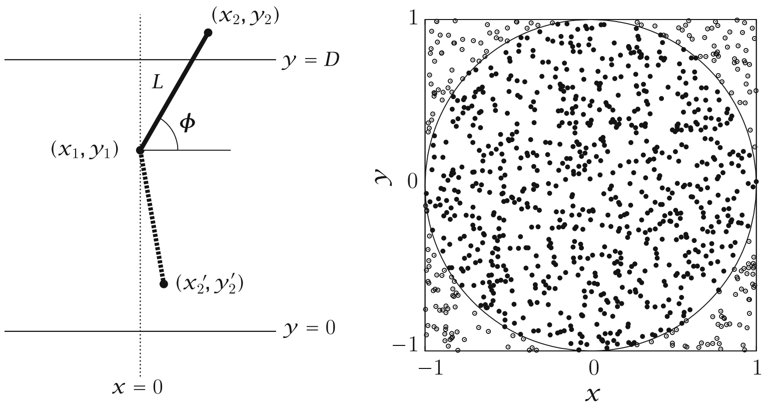


Fig. 13.1 Determining π by the Monte–Carlo method. [Left] Simulation of the Buffon's needle experiment (see text for explanation). [Right] Determination of π by drawing points uniformly distributed on the square $[0, 1] \times [0, 1]$ and checking whether the points fall within the inscribed unit circle. Shown are $N = 1000$ points, $n = 782$ of which lie within the circle, hence $\pi \approx 4n/N = 3.128$

probability densities $2/D$ and $2/\pi$ and the joint density $f(y, \phi) = 4/(\pi D)$. The probability of a needle crossing a line is then

$$P\left(y \leq \frac{L}{2} \sin \phi\right) = \int_0^{\pi/2} \mathrm{d}\phi \int_0^{(L/2)\sin\phi} f(y, \phi)\, \mathrm{d}y = \frac{2L}{\pi D}.$$

Let us check this result by a simple program, which is already our first Monte–Carlo simulation! We draw a random number y_1 with a uniform distribution between 0 and D determining the ordinate of one end of the needle (Fig. 13.1 (left)). The ordinate y_2 of the other end is obtained by drawing an angle ϕ from $[0, 2\pi]$ and calculating $y_2 = y_1 + L \sin \phi$. Then we check whether the needle crosses a line ($y_2 > D$ or $y_2 < 0$) or not. If the whole procedure is repeated N-times and we count n crossings, it holds that

$$\frac{2L}{\pi D} \approx \frac{n}{N}.$$

The estimate of π is then

$$\widehat{\theta} = \frac{2LN}{Dn} \approx \pi.$$

The relative error $(\widehat{\theta} - \pi)/\pi$ depends on the numbers of drawn and accepted events and, of course, on the ratio D/L. Its dependence on N for $D/L = 1.1$ is shown in Fig. 13.2 (left) indicating that in order to determine π to six-digit precision one needs approximately 10^{12} throws. The importance of this dependence on N will become evident shortly.

Let us try another way to "calculate" π. We draw N pairs of random numbers (x_i, y_i), uniformly distributed on $[-1, 1] \times [-1, 1]$. By testing the condition

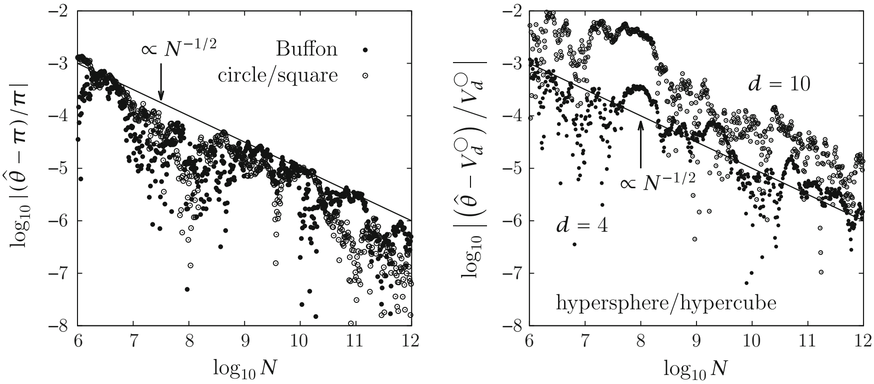


Fig. 13.2 Statistical error of the Monte–Carlo method in dependence of the number of drawn points, N. [Left] Relative error of the approximation for $\pi$when calculating the ratio of the areas of the square and inscribed circle, and in the simulation of Buffon's needle experiment ($D/L = 1.1$). [Right] Relative error of the approximation for the volume of a hypersphere with dimension 4 or 10

$$x_i^2 + y_i^2 \le 1$$

we check whether a pair is within the inscribed unit circle, which we take to be a "good" outcome (Fig. 13.1 (right)). The ratio between the number of good outcomes, n, and the number of all drawn pairs, N, is an approximation for the ratio between the area of the circle and the area of the square: $n/N \approx \pi R^2/(2R)^2 = \pi/4$, thus

$$\widehat{\theta} = \frac{4n}{N} \approx \pi.$$

The error $(\widehat{\theta} - \pi)/\pi$ as a function of N is also shown in Fig. 13.2 (left).

One can look at the problem from the opposite perspective. Assume we already know π, but are interested in the volume of a d-dimensional hypersphere with radius R, "inscribed" in the corresponding hypercube with side $2R$. The exact volumes are

$$V_d^{\bigcirc} = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} R^d, \qquad V_d^{\square} = (2R)^d.$$

Let us simply set $R = 1$ and make N draws of d random numbers $\{x_1, x_2, \ldots, x_d\}$, distributed according to $U(-1, 1)$. At each draw we check the validity of

$$x_1^2 + x_2^2 + \cdots + x_d^2 \le 1.$$

The ratio of the number of draws n for which the condition is fulfilled, to the number of all draws, N, is equal to the ratio of the volumes of the hypersphere and the hypercube, $n/N = V_d^{\bigcirc}/V_d^{\square}$, so we can estimate

$$\widehat{\theta} = \frac{n}{N}\, V_d^{\square} \approx V_d^{\bigcirc}. \tag{13.1}$$

Fig. 13.2 (right) shows how the statistical error of this calculation depends on N if $d = 4$ and $d = 10$. As in the circle-square case we notice the characteristic inverse-square-root decrease of the error.

13.2 Numerical Integration

Numerical integration is among the most important problems that can be solved by the Monte–Carlo method. We wish to calculate a definite integral of the form

$$\theta = \int_{\Omega} g(x)\, f(x)\, \mathrm{d}x, \tag{13.2}$$

where f is some probability density ($f \geq 0$ and $\int_{\Omega} f(x)\, \mathrm{d}x = 1$). We independently draw N values $\{x_1, x_2, \ldots, x_N\}$ of the random variable X, distributed according to the density f. (How this can be done for an arbitrary distribution, is discussed in Sect. C.2.) The value θ is then estimated as

$$\widehat{\theta} = \frac{1}{N} \sum_{i=1}^{N} g(x_i). \tag{13.3}$$

This estimator is unbiased, since its expected value is

$$E\big[\widehat{\theta}\big] = \frac{1}{N} \sum_{i=1}^{N} E\big[g(X_i)\big] = \frac{1}{N} N\, E\big[g(X)\big] = \theta.$$

Yet to control the integration we are interested in its variance, as it determines the quality of the numerical approximation:

$$\begin{aligned}
\mathrm{var}\big[\widehat{\theta}\big] &= \frac{1}{N^2} \sum_{i=1}^{N} \mathrm{var}\big[g(X_i)\big] = \frac{1}{N^2} N\, \mathrm{var}\big[g(X)\big] = \frac{1}{N} \left\{ E\big[g^2(X)\big] - \Big(E\big[g(X)\big]\Big)^2 \right\} \\
&= \frac{1}{N} \left\{ \int_{\Omega} g^2(x)\, f(x)\, \mathrm{d}x - \left(\int_{\Omega} g(x)\, f(x)\, \mathrm{d}x \right)^2 \right\}.
\end{aligned} \tag{13.4}$$

In other words: the estimate for the value of the integral is $\widehat{\theta}$, and the statistical error of this estimate depends on the specific form of g, but, as usual, it decreases inversely proportional to the square root of N,

$$\theta \approx \widehat{\theta} \pm \frac{\sigma_{\widehat{\theta}}}{\sqrt{N}}, \qquad \sigma_{\widehat{\theta}} = \sqrt{\widehat{\theta^2} - \widehat{\theta}^2}, \qquad \widehat{\theta} = \frac{1}{N} \sum_{i=1}^{N} g(X_i), \qquad \widehat{\theta^2} = \frac{1}{N} \sum_{i=1}^{N} g^2(X_i).$$

In the special case $X \sim U(a,b)$ and $\Omega = [a,b]$ one has $f(x) = 1/(b-a)$, hence (13.2)–(13.4) can be merged to

$$\int_a^b g(x)\,\mathrm{d}x \approx (b-a)\,\overline{g} \pm \frac{b-a}{\sqrt{N}}\sqrt{\overline{g^2}-\overline{g}^2}, \tag{13.5}$$

where $\overline{g} = \widehat{\theta}$ and $\overline{g^2} = \widehat{\theta^2}$ are the means of the functions g and g^2 on the interval $[a,b]$. An analogous formula can be written for multi-dimensional integration:

$$\int_\Omega g\,\mathrm{d}V \approx V\,\overline{g} \pm \frac{V}{\sqrt{N}}\sqrt{\overline{g^2}-\overline{g}^2}.$$

Example Use the Monte–Carlo method to calculate the definite integral

$$I = \int_0^\pi \frac{2}{\pi}\left(1-\frac{x}{\pi}\right)\mathrm{e}^{-x/3}\sin 3x\,\mathrm{d}x, \tag{13.6}$$

whose integrand is shown by the full curve in Fig. 13.3 (left)! The exact value is $27(41\pi - 3 - 3\,\mathrm{e}^{-\pi/3})/(1681\pi^2) \approx 0.203023214575878$.

The MC integral can be performed in two ways. First the integrand is rewritten by including a probability density f corresponding to the uniform distribution on $[a,b] = [0,\pi]$,

$$\theta = \int_0^\pi g(x) f(x)\,\mathrm{d}x, \qquad g(x) = \frac{2}{\pi}\left(1-\frac{x}{\pi}\right)\mathrm{e}^{-x/3}\sin 3x, \qquad f(x) = \frac{1}{\pi}.$$

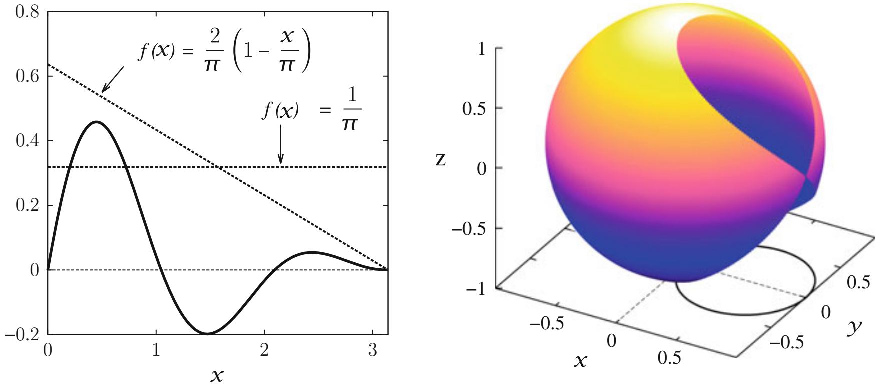


Fig. 13.3 Calculation of definite integrals by the Monte–Carlo method. [Left] The integrand of the integral (13.6) and examples of probability densities used to generate random x_i. [Right] A geometric body whose mass and center of gravity can be computed by the Monte–Carlo method

We draw the values $\{x_i\}_{i=1}^N$ ($N = 10^6$) of a *uniformly distributed* random variable $X \sim U(0, \pi) = \pi U(0, 1)$, and calculate the sums $\widehat{\theta} = \frac{1}{N}\sum_{i=1}^N g(x_i) \approx 0.0649221$ and $\widehat{\theta^2} = \frac{1}{N}\sum_{i=1}^N g^2(x_i) \approx 0.0409225$. (Of course each such draw sequence yields different values.) The estimate for the integral by (13.5) is then

$$I \approx \pi\widehat{\theta} \pm \frac{\pi}{\sqrt{N}}\sqrt{\widehat{\theta^2} - \widehat{\theta}^2} \approx 0.203959 \pm 0.000602. \tag{13.7}$$

On the other hand, the integrand can also be split like this:

$$\theta = \int_0^\pi g(x) f(x)\,\mathrm{d}x, \qquad g(x) = \mathrm{e}^{-x/3}\sin 3x, \qquad f(x) = \frac{2}{\pi}\left(1 - \frac{x}{\pi}\right).$$

In this case x_i must be drawn according to the probability density f corresponding to a triangular distribution on $[0, \pi]$ (fifth row of Table C.1). Taking $N = 10^6$ again we get $\widehat{\theta} = \frac{1}{N}\sum_{i=1}^N g(x_i) \approx 0.203331$ and $\widehat{\theta^2} = \frac{1}{N}\sum_{i=1}^N g^2(x_i) \approx 0.269508$, and thence

$$I \approx \widehat{\theta} \pm \frac{1}{\sqrt{N}}\sqrt{\widehat{\theta^2} - \widehat{\theta}^2} \approx 0.203331 \pm 0.000478.$$

The error of the integral is smaller than that obtained from (13.7) by using the first method. The lesson is that by a different choice of the distribution used to generate the integration variable the variance of the MC estimate can be influenced. This will be the topic of Sect. 13.3. ◁

Example A homogeneous sphere with density ρ_0 and radius $R = 1$ is carved out by a cylinder with radius $R/2$ whose longitudinal symmetry axis is parallel to the z-axis and goes through the point $(x, y) = (R/2, 0)$. The resulting geometrical body is shown in Fig. 13.3 (right). What are its mass and center of gravity?

The mass of a body in three-dimensional space is $m = \int_\Omega \rho(\boldsymbol{r})\,\mathrm{d}V$, where $\mathrm{d}V$ is the volume element of the integration domain Ω. The spherical coordinate system in which $\mathrm{d}V = r^2\,\mathrm{d}r\,\mathrm{d}(\cos\theta)\,\mathrm{d}\phi$ is the most convenient. We also know how to uniformly generate random points in it (formula (C.4)), so for constant density the integrand is nothing but $g(\boldsymbol{r}) = 1$ and $m = \rho_0\int_\Omega \mathrm{d}V$. But integration boundaries are crucial: a cylinder carves out the region defined by

$$\left(x - \frac{R}{2}\right)^2 + y^2 = \left(r\sin\theta\cos\phi - \frac{R}{2}\right)^2 + \big(r\sin\theta\sin\phi\big)^2 \le \frac{R^2}{4}. \tag{13.8}$$

The MC estimate for the mass of the body is therefore simply

$$m = \rho_0\int_\Omega \underbrace{g(\boldsymbol{r})}_{1}\,\mathrm{d}V \approx \frac{\rho_0 V_0}{N}\sum_{i=1}^N \underbrace{g(\boldsymbol{r}_i)}_{1}\,\Delta_i, \tag{13.9}$$

where $V_0 = 4\pi R^3/3$ and where $\Delta_i = 0$, if condition (13.8) is fulfilled or $\Delta_i = 1$ if it is not. The points (r_i, θ_i, ϕ_i) are therefore drawn only to check the validity of (13.8)! With $N = 10^6$ we get, for instance,

$$m \approx 2.98119\, \rho_0 R^3.$$

With some geometric effort the mass can be, in fact, calculated exactly. The body is split to the untouched "left" (L) half and the whittled "right" (R) half which, due to its symmetry, consists of four equal parts. Their masses are

$$m_{\mathrm{L}} = \rho_0 \frac{2\pi R^3}{3}, \qquad m_{\mathrm{R}} = 4\rho_0 \int_0^{\pi/2} \mathrm{d}\phi \int_{\arcsin(\cos\phi)}^{\pi/2} \sin\theta\, \mathrm{d}\theta \int_{R\frac{\cos\phi}{\sin\theta}}^{R} r^2\, \mathrm{d}r = \rho_0 \frac{8R^3}{9},$$

so the total mass of the body is $m = (2\pi/3 + 8/9)\rho_0 R^3 \approx 2.98328\, \rho_0 R^3$. What about the center of gravity, $\boldsymbol{r}^*$? Symmetry clearly dictates $y^* = z^* = 0$, while

$$x^* = \frac{m_{\mathrm{L}} x_{\mathrm{L}}^* + m_{\mathrm{R}} x_{\mathrm{R}}^*}{m_{\mathrm{L}} + m_{\mathrm{R}}}. \tag{13.10}$$

Even this can still be handled analytically:

$$m_{\mathrm{L}} x_{\mathrm{L}}^* = \rho_0 \int_{\pi/2}^{3\pi/2} \int_0^{\pi} \int_0^{R} \underbrace{r\ \sin\theta \cos\theta}_{x} \mathrm{d}V = -\rho_0 \frac{\pi R^4}{4},$$

$$m_{\mathrm{R}} x_{\mathrm{R}}^* = 4\rho_0 \int_0^{\pi/2} \int_{\arcsin(\cos\phi)}^{\pi/2} \int_{R\frac{\cos\phi}{\sin\theta}}^{R} \underbrace{r\ \sin\theta \cos\theta}_{x} \mathrm{d}V = \frac{\rho_0 R^4}{8} \left(\frac{\pi}{2} - \frac{16}{15} \right).$$

By (13.10) it follows that $x^* \approx -0.178774\, R$. How can we calculate x^* by using the MC method? Again we must calculate the sum as in (13.9), where we now set $g(\boldsymbol{r}_i) = x_i = r_i \sin\theta_i \cos\phi_i$. We independently draw the values r_i, θ_i and ϕ_i according to a uniform distribution within the sphere, and finally divide out the total mass; the approximation for the abscissa of the center of gravity is then

$$x^* = \frac{1}{m} \rho_0 \int_{\Omega} r \sin\theta \cos\phi\, \mathrm{d}V \approx \frac{1}{m} \frac{\rho_0 V_0}{N} \sum_{i=1}^{N} r_i \sin\theta_i \cos\phi_i\, \Delta_i. \tag{13.11}$$

With $N = 10^6$ we get $x^* \approx -0.178382\, R$.

We are also interested in the mass and center of gravity of the body if the sphere is inhomogeneous, for example, with the radial dependence of the density $\rho(r) = \rho_0 (r/R)^2$. In this case the recipes (13.9) and (13.11) become

$$m = \rho_0 \int_\Omega \left(\frac{r}{R}\right)^2 \mathrm{d}V \approx \frac{\rho_0 V_0}{N} \sum_{i=1}^{N} \left(\frac{r_i}{R}\right)^2 \Delta_i,$$

$$x^* = \frac{1}{m}\,\rho_0 \int_\Omega \frac{r^3}{R^2} \sin\theta \cos\phi \,\mathrm{d}V \approx \frac{1}{m}\frac{\rho_0 V_0}{N} \sum_{i=1}^{N} \frac{r_i^3}{R^2} \sin\theta_i \cos\phi_i \,\Delta_i.$$

If we wish to deal with this analytically, we must again calculate four integrals for m_L, m_R, $m_\mathrm{L}x^*_\mathrm{L}$ and $m_\mathrm{R}x^*_\mathrm{R}$. This is an increasingly annoying procedure, especially if one imagines a complex carved-out sculpture for which a clear overview of the integration boundaries is lost. On the other hand, the MC method (right side of above equations) only requires us to change a few powers and rerun the program. Which avenue one should pursue depends on the compromise between the desired precision and computing time—yours or computer's. ◁

13.2.1 Advantage of Monte–Carlo Methods over Quadrature Formulas

The statistical error ε of the integral $\widehat{\theta}$ by the Monte–Carlo method decreases with the square root of the sample size: $\theta \approx \widehat{\theta} \pm \varepsilon = \widehat{\theta} \pm \sigma_{\widehat{\theta}}/\sqrt{N}$. One must therefore draw $N \approx \sigma^2_{\widehat{\theta}}/\varepsilon^2$ points in order to determine the value of the integral to a precision of ε. Of course an integral of the type (13.2) could also be computed by using some classical numerical method, say, a quadrature formula

$$\theta \approx \sum_i w_i \, g(x_i) f(x_i).$$

Here w_i are the weights depending on the method and x_i are the quadrature points that suitably fill the integration domain—e.g. the interval $[a, b]$ or a d-dimensional hypercube. The discrete nature of this formula implies an error, too; usually it is estimated as $\varepsilon \le Ch^k$, where h is a measure of the distance between the points of the domain, e.g. $h = (b - a)/N$ on interval $[a, b]$. The error constant C and the power k (quadrature order) depend on the method.

Let T_Q and T_MC denote the times needed to compute the integral by using quadrature and the MC method, respectively. Clearly T_Q grows linearly with the number of points: $T_\mathrm{Q} \propto N \propto (1/h)^d$ where d is the space dimension. From $\varepsilon \le Ch^k$ it follows that $h \ge (\varepsilon/C)^{1/k}$ or $T_\mathrm{Q} \propto (C/\varepsilon)^{d/k} \propto \varepsilon^{-d/k}$. The T_MC is the product of the number of drawn points and time t_1 needed for an individual sample, $T_\mathrm{MC} = t_1 N = t_1 \sigma^2_{\widehat{\theta}}/\varepsilon^2$, thus the ratio of computing times at given ε is

$$\frac{T_\mathrm{MC}}{T_\mathrm{Q}} \propto \frac{\varepsilon^{d/k}}{\varepsilon^2} = \varepsilon^{d/k-2}. \tag{13.12}$$

The ratio $T_{\mathrm{MC}}/T_{\mathrm{Q}}$ decreases with space dimension d and increases with order of quadrature k. Indeed fancy quadrature formulas with high k exist, but the larger the d, the harder it is to find a formula that still ensures $d/k < 2$ and thus $T_{\mathrm{Q}} < T_{\mathrm{MC}}$. Therefore, at large d the MC method is much more efficient than classical quadrature. In practice this applies already at $d \gtrsim 6 - 10$.

13.3 Variance Reduction

Procedures exist which allow us to reduce the variance of MC estimates; for details see [2]. The simplest one is to analytically split the integration domain. Suppose we are seeking the value of the integral $\theta = \int_\Omega g(x)\,\mathrm{d}x$ and we can separate $\Omega = \Omega_1 \cup \Omega_2$ such that $\Omega_1 \cap \Omega_2 = \{\,\}$. The decomposition

$$\theta = \int_\Omega g(x)\,\mathrm{d}x = \int_{\Omega_1} g(x)\,\mathrm{d}x + \int_{\Omega_2} g(x)\,\mathrm{d}x.$$

is useful if the integral can be solved exactly on Ω_1 while the MC method is called for in the remaining domain Ω_2. A separation like this has been done in the Example on page 330: there Ω_1 was the untouched hemisphere that could be handled analytically, while Ω_2 was the carved-out piece where the MC method was applicable. However, one must ensure that there is no statistical correlation between $g(x_1)$ and $g(x_2)$, where $x_1 \in \Omega_1$ and $x_2 \in \Omega_2$.

An obvious simplification is also the splitting the integrand, $g = g_1 + g_2$:

$$\theta = \int_\Omega \big(g_1(x) + g_2(x)\big)\,\mathrm{d}x = \int_\Omega g_1(x)\,\mathrm{d}x + \int_\Omega g_2(x)\,\mathrm{d}x.$$

This seemingly trivial intervention is very effective if the integral of g_1 is relatively easy to compute and g and g_1—in the sense of their “wildness” within the integration domain—are very similar. Then the MC method is only applied to the integral of the “smooth” residual function g_2.

13.3.1 Importance Sampling

The most effective way to reduce the variance of values of integrals by the MC method is *importance sampling*. In the Example on page 329 we realized that the variance can be influenced by the choice of probability density f in the integral (13.2). When the density of the uniform distribution, $f = 1/\pi$, has been used (Fig. 13.3 (left)), the whole interval $[0, \pi]$ has been sampled uniformly, although it is obvious that the points near $x \approx 0.5$ and $x \approx 1.5$ make the dominant contribution to the integral. If, however, sampling with respect to the triangular distribution with density

$f(x) = (2/\pi)(1 - x/\pi)$ has been used, the relevant left part of the interval has been sampled more often that the less important right part, thereby reducing the variance.

Instead of using f, therefore, the values x_i can be drawn according to some other distribution with density p called the *importance function* [2, 3]. With it we compute the integral

$$\theta = \int_\Omega \left[\frac{g(x) f(x)}{p(x)} \right] p(x)\,\mathrm{d}x,$$

where $p(x) \geq 1$, $\int_\Omega p(x)\,\mathrm{d}x = 1$ and $|g(x)f(x)/p(x)| < \infty$. What do we achieve by doing this? The MC estimate based on the sample $\{x_i\}_{i=1}^N$ drawn according to the distribution with density p is then

$$\widehat{\theta} = \frac{1}{N} \sum_{i=1}^{N} \frac{g(x_i) f(x_i)}{p(x_i)}.$$

The variance of this estimator is

$$\mathrm{var}\big[\widehat{\theta}\big] = \frac{1}{N}\,\mathrm{var}\left[\frac{g(X)f(X)}{p(X)}\right] = \frac{1}{N}\left\{ E\left[\frac{g^2(X)f^2(X)}{p^2(X)}\right] - \left(E\left[\frac{g(X)f(X)}{p(X)}\right]\right)^2 \right\},$$

where all expected values are to be taken with respect to the distribution of X with density p—this is crucial! We wish to find p that minimizes this variance. The second term is simply $(E[g(X)f(X)/p(X)])^2 = \left(\int_\Omega g(x) f(x)\,\mathrm{d}x\right)^2 = \theta^2$, so the key to success is hidden in the first term. Jensen's inequality (4.10) dictates its lower bound:

$$E\left[\frac{g^2(X)f^2(X)}{p^2(X)}\right] \geq \left(E\left[\frac{|g(X)f(X)|}{p(X)}\right]\right)^2 = \left(\int_\Omega |g(x)| f(x)\,\mathrm{d}x \right)^2.$$

The bound is reached when

$$p(x) = \frac{|g(x)| f(x)}{\int_\Omega |g(x')| f(x')\,\mathrm{d}x'}.$$

Alas, we do not know the exact value of the integral in the denominator, otherwise we would not be computing it! In practice we therefore seek a function $p(x)$ which is *as similar as possible* to the function $|g(x)| f(x)$, i.e. such $p(x)$ that the ratio $|g(x)| f(x)/p(x)$ is approximately constant throughout the integration domain.

Example (Adapted from [3].) Let us calculate the definite integral

$$\theta = \int_0^1 \cos\left(\frac{\pi x}{2}\right)\,\mathrm{d}x,$$

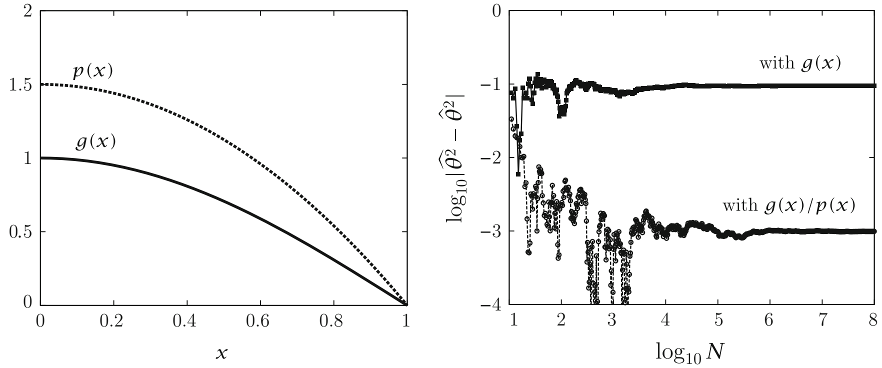


Fig. 13.4 Weighting the integrand g by the importance function p. [Left] Graphs of functions $g(x) = \cos(\pi x/2)$ and $p(x) = \frac{3}{2}(1-x^2)$. [Right] Variance of the plain MC estimate (*upper graph*) and by using the importance function p (*lower graph*)

which is of the form (13.2) with $\Omega = [0, 1]$, $g(x) = \cos(\pi x/2)$ (Fig. 13.4 (left)) and $f(x) = 1$. At first we ignore the importance function and do it the old way: with values $\{x_i\}_{i=1}^N$ of the uniformly distributed variable $X \sim U(0, 1)$ we compute

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N} g(x_i), \qquad \widehat{\theta^2} = \frac{1}{N}\sum_{i=1}^{N} g^2(x_i). \tag{13.13}$$

The variance $\mathrm{var}[\widehat{\theta}]$ of the estimate $\widehat{\theta}$ at large N can even be calculated exactly:

$$\lim_{N\to\infty}\left(\widehat{\theta^2} - \widehat{\theta}^2\right) = \int_0^1 \cos^2\left(\frac{\pi x}{2}\right)\mathrm{d}x - \left[\int_0^1 \cos\left(\frac{\pi x}{2}\right)\mathrm{d}x\right]^2 \approx 0.09472. \tag{13.14}$$

The obtained value $\log_{10} 0.09472 \approx -1.0236$ can be seen in the upper graph of Fig. 13.4 (right), showing the approximation for the variance as a function of N.

Now choose an importance function p which is "as similar as possible" to g, say, $p(x) = \frac{3}{2}\left(1 - x^2\right)$. This function is non-negative and normalized to 1 on [0, 1], so it satisfies the requirements for a probability density. Since we are now computing the integral

$$\theta = \int_0^1 \frac{g(x)}{p(x)}\, p(x)\,\mathrm{d}x,$$

the values x_i in the sums

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N} \frac{g(x_i)}{p(x_i)}, \qquad \widehat{\theta^2} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{g(x_i)}{p(x_i)}\right)^2, \tag{13.15}$$

must be drawn according to the density p! (Random values with such distribution can be generated by using some method of Sect. C.2.) A pleasant surprise is in store:

$$\lim_{N\to\infty}\left(\widehat{\theta^2}-\widehat{\theta}^2\right)\approx 0.000990.$$

By a fortunate choice of p the variance has been reduced by two orders of magnitude compared to the plain estimate (13.14); see the lower graph in Fig. 13.4 (right) which stabilizes at $\log_{10} 0.000990 \approx -3.0044$ for large N. ◁

Example (Adapted from [3].) In the case of singular functions or functions whose certain moments do not exist, scaling the integrand by an importance function is unavoidable. For instance, let us calculate the integral

$$\theta=\int_0^1 \frac{1}{\sqrt{x(1-x)}}\,\mathrm{d}x. \tag{13.16}$$

The integrand $g(x)=1/\sqrt{x(1-x)}$ is singular at $x=0$ and $x=1$ (Fig. 13.5 (left)), hence the plain MC estimate has infinite variance: with increasing N the sums (13.13), where the values x_i are drawn according to the uniform distribution $U(0,1)$, as well as the variance, keep increasing. This divergent behavior is demonstrated by the upper graph of Fig. 13.5 (right).

In such case it is prudent to choose an importance function that has the singularities at the same points and of the same order as g, for example

$$p(x)=\frac{1}{4\sqrt{x}}+\frac{1}{4\sqrt{1-x}}. \tag{13.17}$$

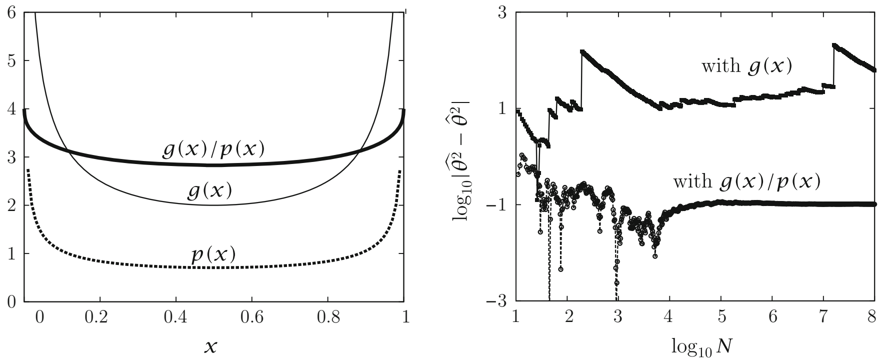


Fig. 13.5 [Left] Choice of importance function p for singular integrands $g(x)$. All singularities of g should be included in p, so that the ratio g/p is regular at the problematic points and approximately constant throughout the domain. [Right] The variance of the MC estimate for Example (13.16) without the importance function and by using the g/p integrand

Random numbers according to this distribution can be drawn by using the tools of Sect. C.2, like the inverse method: it corresponds to a very simple algorithm

1. Draw $\xi_1, \xi_2 \sim U(0, 1)$.
2. If $\xi_2 < \frac{1}{2}$, set $X = \xi_1^2$, otherwise $X = 1 - \xi_1^2$.

(For explanation see also Sect. 3.4.4 in [3].) The weighted integrand $g(x)/p(x)$ is plotted by the thick curve in Fig. 13.5 (left). The sums (13.15), where the values x_i are drawn according to the density (13.17), now yield a finite variance. Its dependence on N is shown by the lower graph in Fig. 13.5 (right). ◁

13.3.2 The Monte–Carlo Method with Quasi-Random Sequences

By a special kind of "drawing" the values $\{x_i\}_{i=1}^N$ the convergence of MC estimates can be accelerated. Instead of the typical $\sim N^{-1/2}$ behavior (see Fig. 13.2) trends like $\sim N^{-2/3}$ or even N^{-1} can be achieved, which, from the viewpoint of (13.12) is an argument in favor of the MC method. The word "drawing" actually implies a deterministic calculation of special *sequences* of d-plets

$$\{\boldsymbol{x}_i\}_{i=1}^N, \qquad \boldsymbol{x}_i = \big(x_{1,i},\, x_{2,i},\, \ldots,\, x_{d,i}\big),$$

which we use to sample the d-dimensional integration region at N points. The essence of the method is precisely the manner of this sampling: it is devised such that the points in d-dimensional space, forming the so-called *quasi-sequence*, "maximally avoid each other". An illustration of such quasi-sequence for $d = 2$, where each value in the pair $(x_1, x_2)_i$ corresponds to a normally distributed variable $X \sim U(0, 1)$, is offered by Fig. 13.6.

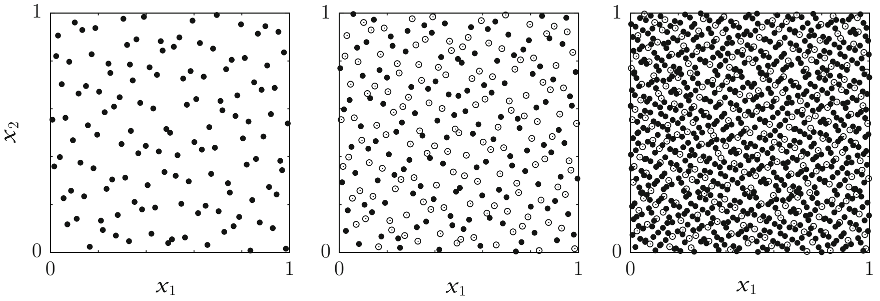


Fig. 13.6 Points of the Sobol quasi-random sequence on the domain [0, 1] × [0, 1]. [Left] The first 128 points. [Center] The first 256 points. [Right] The first 1024 points

The full circles in the left Figure denote the first $N = 128$ points. The center Figure shows $N = 256$ points of the same sequence: the previous and the new 128 points are denoted by empty and full circles, respectively. The right Figure contains $N = 1024$ points, of which the 256 old points are again denoted by empty circles and the 768 new ones with full circles. Apparently space is being filled by an almost regular pattern, yet the straggling of the points is more random and more uniform across the whole space than in drawing the values by *pseudo*random algorithms (see Appendix C.1). Quasi-random sequences therefore truly start to excel only at very large N and high dimensions d.

Several brands of quasi-sequences and methods of their generation exist. Among the most popular is the Sobol sequence [4, 5], which has been used to generate the points in Fig. 13.6. The basic version is available in [6]; improvements for higher dimensions and larger periods are discussed in [7, 8].

Example Let us redo the calculation of the volume of the four-dimensional hypersphere by using the MC method (see (13.1) and Fig. 13.2 (right)), but now we draw the points (x_1, x_2, x_3, x_4) in four-dimensional space as elements of the corresponding $d = 4$ Sobol sequence. The statistical error of the calculated volume estimate as a function of N is shown in Fig. 13.7. ◁

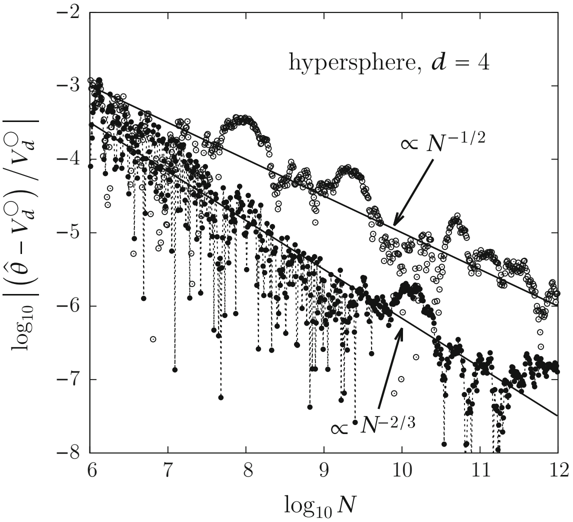


Fig. 13.7 Statistical error of the MC approximation for the volume of a four-dimensional hypersphere. (Compare to Fig. 13.2). Shown is the dependence of the error on the number of points in the Sobol quasi-sequence ($\propto N^{-2/3}$) as compared to the usual generation of pseudo-random numbers ($\propto N^{-1/2}$)

13.4 Markov-Chain Monte Carlo ⋆

Classical Markov chains (see Sect. 12.1) can be used to devise an effective method to generate random numbers according to arbitrary, even very complicated distributions, known as *Markov-chain Monte Carlo* (MCMC) [9]. The essence of the method is that the generated values form the states of a Markov chain whose equilibrium distribution is precisely the required probability distribution.

The key property of the chain that we exploit is *reversibility*. An irreducible Markov chain is reversible if the equilibrium probabilities π_j (see (12.4) and (12.6)) satisfy the requirement of *detailed balance*

$$\pi_i p_{ij} = \pi_j p_{ji}, \qquad \forall i, j \in \Omega. \tag{13.18}$$

Recalling $P(A|B) = P(B|A)P(A)/P(B)$ we see at once that detailed balance also means

$$\begin{aligned} P\bigl(X_{t-1} = i \mid X_t = j\bigr) &= P\bigl(X_t = j \mid X_{t-1} = i\bigr)\frac{P(X_{t-1} = i)}{P(X_t = j)} \\ &= p_{ij}\frac{\pi_i}{\pi_j} = p_{ij}\frac{p_{ji}}{p_{ij}} = p_{ji} = P\bigl(X_t = i \mid X_{t-1} = j\bigr), \end{aligned}$$

in plain words: the probability for a transition between given states forward in time equals the probability for a transition between them backward in time. Assume that the distribution $\boldsymbol{\pi}$ satisfying (13.18) is unique (see Sect. 12.1.1). In the following we use this arsenal to formulate the core procedure of the MCMC method, the so-called Metropolis–Hastings algorithm.

13.4.1 Metropolis–Hastings Algorithm

Let Ω be the state space of the Markov chain possessing the equilibrium distribution $\pi(x)$, where $x \in \Omega$. We shall attempt a quite general description of the MCMC method, where Ω may be discrete or continuous; the value x may even represent some exceedingly complex entity, say, the state of a two-dimensional spin lattice on which (at given temperature) some spins are oriented parallel to the magnetic field and some opposite to it—the formalism remains essentially the same.

If, for example, we wish to use the MCMC method to estimate the value of a one-dimensional integral of the form (13.2) by the sum (13.3), the random numbers must be drawn according to the desired distribution $f(x) = \pi(x)$. The equilibrium distribution is therefore also known as the *target distribution*. Why don't we simply generate the values with the target distribution π by using some method of Sect. C.2, say, the rejection method (Sect. C.2.6)? In one dimension this is sensible, but in multiple dimensions the fraction of rejected points increases to the level of

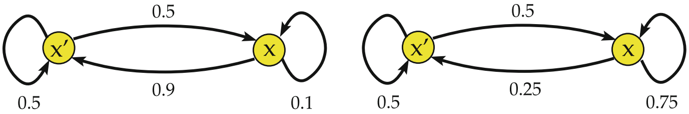


Fig. 13.8 A Markov chain in which only transitions between the states x and x' are allowed. [Left] Chain with equilibrium distribution (13.20). [Right] Chain with equilibrium distribution (13.21)

the method being useless. Similar problems are encountered in importance sampling (Sect. 13.3.1).

The main task is then: draw a sequence of values $S = \{x_0, x_1, \ldots\}$ of the Markov chain such that its equilibrium distribution is precisely $\pi(x)$. For this purpose we first introduce the *candidate distribution*

$$q(x'|x), \qquad x, x' \in S,$$

which is used to place a weight on the transitions between the states x and x'. (In previous notation $q(x'|x)$ is just p_{ij}.) If the present state is x, the state x' is a “candidate” for the next state, with probability $q(x'|x)$. At this point reversibility and detailed balance (13.18) enter. To illustrate the main idea behind the Metropolis–Hastings algorithm, imagine for a moment that we can only jump between the states x and x', as shown in Fig. 13.8, with probabilities $q(x|x) = 0.1$, $q(x'|x) = 0.9$ and $q(x|x') = q(x'|x') = 0.5$, corresponding to the stochastic matrix

$$\mathcal{P} = \begin{pmatrix} 0.1 & 0.9 \\ 0.5 & 0.5 \end{pmatrix}. \tag{13.19}$$

If q satisfies the detailed-balance condition $\pi(x)q(x'|x) = \pi(x')q(x|x')$, we are done, since then $\pi(x)$ is already the equilibrium distribution of the chain. In the case in Fig. 13.8 (left) the condition is fulfilled: the equilibrium distribution satisfying $\boldsymbol{\pi} = \boldsymbol{\pi}\mathcal{P}$ is

$$\boldsymbol{\pi} = \big(\pi(x), \pi(x')\big) = \left(\frac{5}{14}, \frac{9}{14}\right), \tag{13.20}$$

so

$$\pi(x)q(x'|x) = \frac{5}{14}\frac{9}{10} = \pi(x')q(x|x') = \frac{9}{14}\frac{1}{2} = \frac{9}{28}.$$

But what if (13.20) is not our desired (target) distribution and we actually wish to attain the equilibrium distribution

$$\boldsymbol{\pi} = \left(\frac{2}{3}, \frac{1}{3}\right)? \tag{13.21}$$

The matrix (13.19) can not do it, since $\boldsymbol{\pi} \neq \boldsymbol{\pi}\mathcal{P}$. Besides, reversibility is gone:

$$\pi(x)q(x'|x) = \frac{2}{3}\frac{9}{10} > \pi(x')q(x|x') = \frac{1}{3}\frac{1}{2}. \tag{13.22}$$

The inequality reveals that the $x \to x'$ transitions are too frequent with respect to $x' \to x$ for the chain to be in equilibrium. Equilibrium is restored if the left-hand side if multiplied by a suitable factor, $(5/18)\pi(x)q(x'|x) = \pi(x')q(x|x')$. Then instead of $q(x'|x) = 0.9$ the transition probability is $\widetilde{q}(x'|x) = (5/18)0.9 = 0.25$, and one must also fix $\widetilde{q}(x|x) = 1 - \widetilde{q}(x'|x) = 0.75$. The new equilibrated chain with the stochastic matrix

$$\widetilde{\mathcal{P}} = \begin{pmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \end{pmatrix}$$

is shown in Fig. 13.8 (right). For the chain with such matrix the desired distribution (13.21) is indeed stationary, $\boldsymbol{\pi} = \boldsymbol{\pi}\widetilde{\mathcal{P}}$.

This kind of tweaking of non-diagonal transition probabilities is the foundation of the Metropolis–Hastings algorithm [10–12]. The guesswork of finding the correct scaling factor for both matrix elements is replaced by a weight α with which a "good" state or configuration is accepted:

$$\pi(x)\big[q(x'|x)\alpha(x'|x)\big] = \pi(x')\big[q(x|x')\alpha(x|x')\big].$$

If we wish to accept x while the chain tends to move to x', we should be very generous in accepting the $x' \to x$ transitions, so we set $\alpha(x|x') = 1$, while the $x \to x'$ transitions should be stifled with probability $\alpha(x'|x)$ which, by the above equation, is

$$\alpha(x'|x) = \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}.$$

If the balance (13.22) tips in the opposite direction, in favor of the $x' \to x$ transitions, we simply exchange the roles of x and x'. The same reasoning applies to a chain with many states, not just two. For any two states x and y we then define the *acceptance probability* of y with respect to x:

$$\rho(x, y) = \min\left\{\frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}, 1\right\}. \tag{13.23}$$

The Metropolis–Hastings (MH) algorithm that generates a Markov chain X_t with states $x_{t+1}, x_{t+2}, x_{t+3}\ldots$ given the initial state x_t, the desired target distribution π and the chosen candidate distribution q, is therefore exceedingly simple:

1. Draw a value y_t of the random variable $Y \sim q(y|x_t)$.
2. For the next state of the chain take

$$x_{t+1} = \begin{cases} y_t \text{ with probability } \rho\big(x_t, y_t\big), \\ x_t \text{ otherwise.} \end{cases}$$

3. Assign $x_t \leftarrow x_{t+1}$ and go to 1.

Independent Metropolis–Hastings Algorithm

If the candidate distribution does not depend on the present state of the chain, that is, $q(y|x) = g(y)$, the algorithm is even simpler:

1. Draw a value y_t of the random variable $Y \sim g(y)$.
2. For the next state of the chain take

$$x_{t+1} = \begin{cases} y_t \text{ with probability } \min\left\{\dfrac{\pi(y_t)g(x_t)}{\pi(x_t)g(y_t)},\, 1\right\}, \\ x_t \text{ otherwise.} \end{cases}$$

3. Assign $x_t \leftarrow x_{t+1}$ and go to 1.

Both versions of the algorithm generate the equilibrium distribution π even if its normalization constant is unknown, as it cancels in the ratio $\pi(y)/\pi(x)$. Moreover, it is fascinating that it is generated *regardless of the form* of the function q! We must only ensure that π and q have the same definition domains. However, from the perspective of efficiency and precision of the algorithm it does matter what kind of q is chosen: we learn this in the following Example. For additional details see [13].

Example We wish to generate random numbers according to a continuous distribution corresponding to a mixture of two normal densities of the form (3.7),

$$\pi(x) = w_1 f\big(x; \mu_1, \sigma_1^2\big) + w_2 f\big(x; \mu_2, \sigma_2^2\big), \tag{13.24}$$

with weights $w_1 = 0.3$, $w_2 = 0.7$ and parameters $\mu_1 = 0$, $\mu_2 = 10$, $\sigma_1 = \sigma_2 = 2$. The rotated graphs of $\pi(x)$ are shown in the two small rectangles in Fig. 13.9 (right).

Such a density enters, for instance, in a numerical evaluation of the integral

$$\theta = \int_{-\infty}^{\infty} g(x)\pi(x)\,\mathrm{d}x, \tag{13.25}$$

where g is some function. Let the candidate function used to draw the new state y in the MH algorithm also be a normal density with its mean equal to the previous state x:

$$q(y|x) = f\big(y; x, \sigma_q^2\big), \tag{13.26}$$

while σ_q is a free parameter. Choose $\sigma_q = 0.3$, initial state $x_0 = 0$, and run the algorithm for $T = 10{,}000$ steps. The obtained T states of the chain are shown in

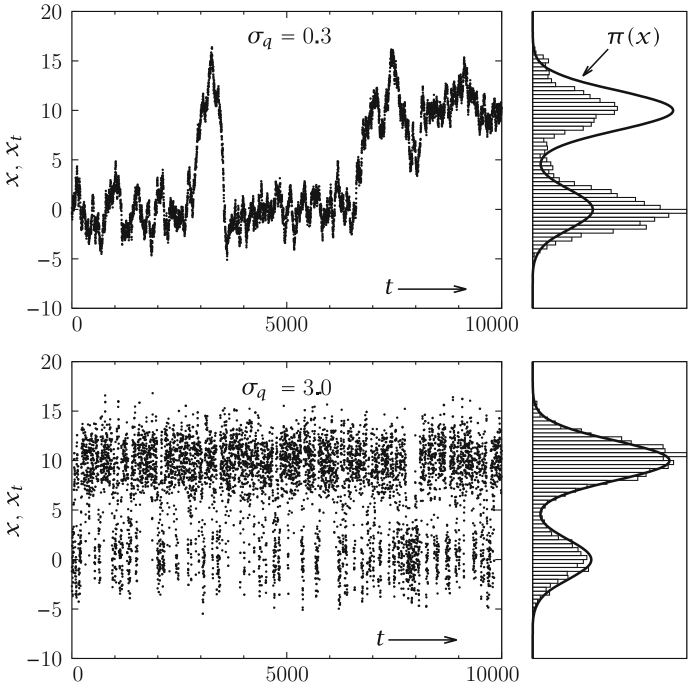


Fig. 13.9 Illustrating the MH algorithm as a generator of sequences of states in a Markov chain. [Top left] The states $\{x_1, x_2, \ldots, x_{10000}\}$ of the chain generated by the algorithm with $\sigma_q = 0.3$ (poor mixing). [Top right] Normalized histogram of states after $T = 10{,}000$ steps compared to the target probability density (13.24). [Bottom] Same as the panels above, but with $\sigma_q = 3.0$ (good mixing)

Fig. 13.9 (top left), while the normalized histogram of these states compared to the target density (13.24) is shown in Fig. 13.9 (top right).

The Figure tells us that the algorithm has spent about 2500 steps in sampling the first region of the target density centered at $x = \mu_1$, switched to the other region around $x = \mu_2$ after approximately 3000 steps, then changed its mind and quickly returned to the first region, sampling it for the next 3000 steps, and spent most of its remaining time in the second region. The histogram of the generated states x_i poorly matches the desired target density, because the algorithm dwelled at rather restricted portions of the definition domain for too long. The culprit is the parameter σ_q being too small, much smaller than σ_1 and σ_2. The density (13.26) used to randomly explore the neighborhood of the old value x in order to come up with the new value y, is too "sharp" and leaves very limited freedom to the acceptance probability (13.23). The algorithm spends too much time in the same place: we say the states are *poorly mixed.*

If σ_q is chosen more prudently, setting e.g. $\sigma_q = 3.0$, which is comparable to $\sigma_1 = \sigma_2 = 2$, we obtain something like Fig. 13.9 (bottom left and right). Now the algorithm is very jittery and keeps on jumping between the two main prominences of

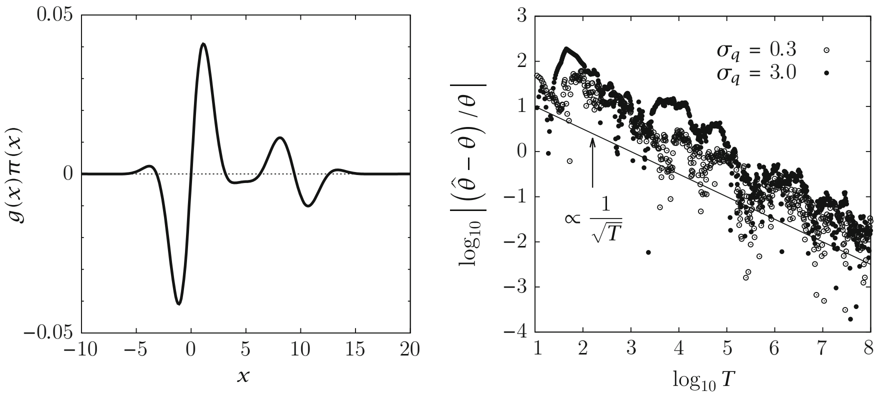


Fig. 13.10 Numerical integration with the MCMC method (MH algorithm). [Left] The graph of the integrand $g(x)\pi(x)$. [Right] Convergence of the estimate (13.28) to the exact value (13.27). As for any other statistical average, the relative error of integration by using the MCMC method and the Metropolis–Hastings algorithm has the typical inverse-square-root dependence on the number of steps

the target distributions; the smearing of the candidate function is just about right that the algorithm can comb through all relevant parts of the domain. We have used the same number of steps, but with a carefully tailored function q the agreement between the generated and target distributions has greatly improved. A simple criterion for a basic tune of the candidate function is the fraction of steps in which the new state is accepted: it should hover around 0.5.

We know how to generate random numbers according to (13.24); now let us calculate some nasty integral of the form (13.25), for example, with the function

$$g(x) = \frac{\sin x}{1 + x^2/10}.$$

The graph of $g(x)\pi(x)$ is in Fig. 13.10 (left), and the exact value of the integral is

$$\theta = \int_{-\infty}^{\infty} g(x)\pi(x)\,\mathrm{d}x \approx 0.0015805060995968 39. \tag{13.27}$$

With random numbers $\{x_1, x_2, \ldots, x_T\}$ generated by the MH algorithm we calculate the estimates of the integral (partial sums)

$$\widehat{\theta} = \frac{1}{T}\sum_{t=1}^{T} g(x_t) \tag{13.28}$$

for various T. The relative error between the estimates $\widehat{\theta}$ at each T and the exact value θ is shown in Fig. 13.10 (right). ◁

With this Example we have barely scratched the surface of Markov-chain Monte Carlo methods. Approaches of the MCMC type truly blossom in multiple dimensions, where the classical methods of generating probability distributions become inefficient or—in the case of more general state spaces Ω—completely useless. Further reading is offered by [14, 15].

References

1. N. Metropolis, S. Ulam, The Monte Carlo method. J. Am. Stat. Assoc. **44**, 335 (1949)
2. J.E. Gentle, *Random Number Generation and Monte Carlo Methods*, 2nd edn. (Springer, Berlin, 2003)
3. M.H. Kalos, P.A. Whitlock, *Monte Carlo Methods*, 2nd edn. (Wiley, Weinheim, 2008)
4. I.M. Sobol', On the distribution of points in a cube and the approximate evaluation of integrals. USSR Comput. Maths. Math. Phys. **7**, 86 (1967)
5. I.A. Antonov, V.M. Saleev, An economic method of computing LP_τ sequences. USSR Comput. Math. Math. Phys. **19**, 252 (1979)
6. W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, 3 edn. (Cambridge University Press, Cambridge, 2007)
7. S.H. Paskov, J.F. Traub, Faster valuation of financial derivatives. J. Portf. Manag. **22**, 113 (1995)
8. J.F. Traub, S.H. Paskov, I.F. Vanderhoof, A. Papageorgiou, *Portfolio Structuring Using Low-discrepancy Deterministic Sequences*, U.S. Patent 6,058,377, http://www.google.com/patents/US6058377
9. S. Chib, in *Handbook of Computational Statistics*, ed. by J.E. Gentle et al. Markov Chain Monte Carlo Technology, (Springer, Berlin, 2012), p. 73
10. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087 (1953)
11. S. Chib, E. Greenberg, Understanding the Metropolis-Hastings algorithm. Am. Stat. **49**, 327 (1995)
12. I. Beichl, F. Sullivan, The Metropolis algorithm. Comput. Sci. Eng. **2**(1), 65 (2000)
13. C.P. Robert, G. Casella, *Introducing Monte Carlo Methods with R* (Springer, Berlin, 2010)
14. W.R. Gilks, S. Richardson, D. Spiegelhalter (eds.), *Markov-Chain Monte Carlo in Practice* (Chapman & Hall / CRC, New York, 1996)
15. S. Brooks, A. Gelman, G. Jones, X.-Li. Meng (eds.), *Handbook of Markov-Chain Monte Carlo* (Chapman & Hall / CRC, New York, 2011)

Chapter 14
Stochastic Population Modeling

Abstract One way to analyze the time evolution of discrete populations is to develop models of birth, death and other mechanisms that influence the size of the population, as well as interactions between two or more populations. Modeling of births and deaths is introduced, followed by a discussion of a combined birth-death model representable in matrix form. The existence of equilibrium states is questioned and the time evolution of population distribution moments is presented.

This Chapter is devoted to population dynamics, i.e. modeling of birth, death and other processes experienced by "individuals" in discrete populations. (Such phenomena can be modeled as discrete-state Markov processes—see Chap. 12—but we discuss them separately here in a slightly simpler form.) "Individuals" may be single cells that die or successfully divide, people getting sick due to an infectious disease to which they succumb or become immune, subatomic particles being born or decaying in cosmic ray showers, or photons and electrons in resonant cavities of multi-level lasers [1–4].

Let $X(t)$ be the size of the population at time t. Births, deaths, emigration, immigration and other mechanism that in any way modify the population size are treated stochastically, so $X(t)$ is a random variable. The probability that X at time t has value n, is denoted by

$$p_n(t) = P\big(X(t) = n\big), \qquad n = 0, 1, 2, \ldots$$

14.1 Modeling Births

Let us begin with the simple case of cells dividing at a constant rate $\lambda > 0$. The probability of one cell dividing in two in the interval $(t, t + \Delta t]$ is $\lambda \Delta t$. The probability that the whole population with $X(t)$ cells at time t increases in size by precisely one cell in the interval $(t, t + \Delta t]$, is therefore $\lambda X(t) \Delta t$. Assume that at time $t + \Delta t$ the population contains n cells. If Δt is small enough, multiple divisions may be neglected, so the population could achieve this state by two ways only: from a state

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6_14

with n cells at time t and no division in $(t, t + \Delta t]$; or from the state with $n - 1$ cells at time t and precisely one division in $(t, t + \Delta t]$,

$$\begin{aligned} p_n(t + \Delta t) &= p_n(t)\big(1 - \lambda\Delta t\big)^n + p_{n-1}(t)\big(1 - \lambda\Delta t\big)^{n-1}\lambda(n-1)\Delta t \\ &= p_n(t)\big(1 - \lambda n\Delta t\big) + p_{n-1}(t)\lambda(n-1)\Delta t + \mathcal{O}(\Delta t). \end{aligned}$$

Assume that the population at time zero contains N cells, $X(0) = N$. Of course, it can only grow, thus one always has $n \geq N$. For very small Δt it holds that $\big(p_n(t + \Delta t) - p_n(t)\big)/\Delta t \approx \mathrm{d}p_n(t)/\mathrm{d}t = \dot{p}_n(t)$. In the $\Delta t \to 0$ limit the difference equation therefore becomes a system of differential equations:

$$\begin{aligned} \dot{p}_N(t) &= -\lambda N p_N(t), \\ \dot{p}_n(t) &= -\lambda n p_n(t) + \lambda(n-1)p_{n-1}(t), \quad n > N. \end{aligned} \tag{14.1}$$

The first equation is simpler, as the probability for the state with N cells can not be nourished by the state with $N - 1$ cells, but can only diminish, therefore $p_{N-1}(t) = 0$. The initial conditions are

$$\begin{aligned} p_N(0) &= 1, \\ p_n(0) &= 0, \quad n > N. \end{aligned} \tag{14.2}$$

The solution of the system (14.1) with initial condition (14.2) for general n is [3]

$$p_n(t) = \binom{n-1}{N-1} \mathrm{e}^{-\lambda N t}\big(1 - \mathrm{e}^{-\lambda t}\big)^{n-N}, \qquad n \geq N,$$

which corresponds to the negative binomial distribution (5.8) with probability $p = \mathrm{e}^{-\lambda t}$ for a "good" outcome. The expected value of the population size at time t is

$$E\big[X(t)|X(0) = N\big] = N\mathrm{e}^{\lambda t},$$

which, of course, implies unhindered growth, and its variance is

$$\mathrm{var}\big[X(t)|X(0) = N\big] = N\mathrm{e}^{\lambda t}\big(\mathrm{e}^{\lambda t} - 1\big).$$

14.2 Modeling Deaths

By analogy to the birth-only model it is easy to establish the time evolution of a population that only experiences deaths. Let us stay with simple cellular division; the probability that a single cell dies in the interval $(t, t + \Delta t]$ is $\mu\Delta t$, where $\mu > 0$ is the mortality rate. The probability that in the whole population with size $X(t)$ at time t a single cell dies in the interval $(t, t + \Delta t]$, is $\mu X(t)\Delta t$. The dynamics of the population is therefore described by the system of differential equations

$$\begin{aligned}\dot{p}_N(t) &= -\mu N p_N(t),\\ \dot{p}_n(t) &= -\mu n p_n(t) + \mu(n+1)p_{n+1}(t), \quad 0 \le n < N.\end{aligned} \tag{14.3}$$

If the size of the population at time zero is N, the initial conditions are

$$\begin{aligned}p_N(0) &= 1,\\ p_n(0) &= 0, \quad 0 \le n < N.\end{aligned} \tag{14.4}$$

The solution of the system (14.3) with initial condition (14.4) for arbitrary n is [3]

$$p_n(t) = \binom{N}{n} \mathrm{e}^{-\mu n t} \left(1 - \mathrm{e}^{-\mu t}\right)^{N-n}, \qquad n = 0, 1, 2, \ldots, N.$$

This is the usual binomial distribution with $p = 1 - q = \mathrm{e}^{-\mu t}$, so the expected value and variance of the variable X at time t are at hand:

$$\begin{aligned}E\big[X(t)|X(0) = N\big] &= Np = N\mathrm{e}^{-\mu t},\\ \mathrm{var}\big[X(t)|X(0) = N\big] &= Npq = N\mathrm{e}^{-\mu t}\big(1 - \mathrm{e}^{-\mu t}\big).\end{aligned} \tag{14.5}$$

Example A dying population can be modeled by a simple computer simulation. The key realization is that death is a Poisson process with a known mortality rate (Poisson parameter) μ. In the whole population with size $X(t)$ at time t, *on average* $\overline{X} = \mu X(t)\Delta t$ cells die during the interval $(t, t + \Delta t]$—while in an actual "experiment" we may record zero, one, two,… deaths. The change of the population size at each time step is then $\Delta X(t) = -\mathcal{P}(\mu X(t)\Delta t)$, where $\mathcal{P}$ denotes a discrete random number, distributed according to the Poisson distribution with mean $\mu X(t)\Delta t$. Poisson generators are available in standard libraries [5]. We begin the simulation with $X(0) = N$ and subtract

$$X(t + \Delta t) = X(t) - \mathcal{P}\big(\mu X(t)\Delta t\big), \tag{14.6}$$

until the population size drops to zero. Five such random death "paths" with $N = 250$, $\mu = 0.5$ and $\Delta t = 0.1$ are shown in Fig. 14.1.

The exponential decay is not surprising: due to the randomness of the process the extinction times are somewhat scattered, but death of the entire population is unavoidable as

$$\lim_{t\to\infty} p_0(t) = \lim_{t\to\infty} \big(1 - \mathrm{e}^{-\mu t}\big)^N = 1.$$

A more relevant question is *when on average* the population dies out: we are interested in the *distribution of extinction times*. We follow many random death "paths" and

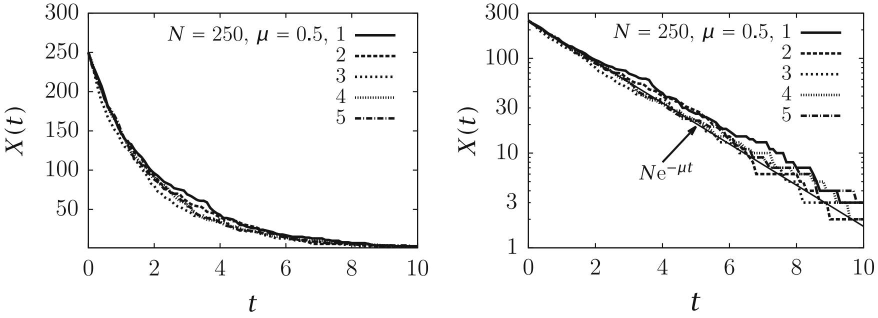


Fig. 14.1 Dying out of a population with Poisson-distributed number of deaths in each time interval. [Left] Depiction on linear scale. [Right] Depiction on logarithmic scale, together with the expected exponential dependence (14.5)

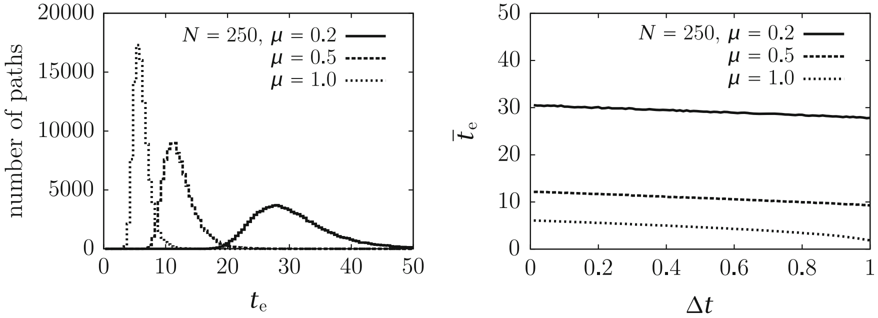


Fig. 14.2 Dying out of populations with different mortality rates. [Left] The distribution of extinction times t_e, generated by following 100,000 different random paths from the initial state with $X(0) = N = 250$ to the final with $X(t_\mathrm{e}) = 0$. [Right] Dependence of the average extinction time $\bar{t}_\mathrm{e}$ on the step length Δt

note the times at which the population size drops to zero. A few such distributions with the same N and Δt as above are shown in Fig. 14.2 (left).

It is also interesting to know whether the calculated extinction time and the distribution of its averages depend on Δt. The time step should not be too large, otherwise one can mow down the entire population in a single Δt. For small enough steps, on the other hand, the results should be approximately independent of Δt: namely, if $X_1, X_2, \ldots, X_k$ are Poisson-distributed random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_k$, their sum $X_1 + X_2 + \cdots + X_k$ is also a Poisson-distributed variable with parameter $\lambda_1 + \lambda_2 + \cdots + \lambda_k$ (convolution; see also Example on page 148). In other words, in drawing random numbers we rely on the approximation

$$\mathcal{P}\big(\lambda X(t)\Delta t\big) \approx \mathcal{P}\big(\lambda X(t)\xi\Delta t\big) + \mathcal{P}\big(\lambda X(t)(1-\xi)\Delta t\big), \qquad 0 \le \xi \le 1.$$

The quality of this approximation can be judged from Fig. 14.2 (right). ◁

14.3 Modeling Births and Deaths

The dynamics of births and deaths can be merged in a unified model that can even be endowed by a more general ansatz for natality and mortality rates in a population with size n. Let us denote them by λ_n and μ_n. So far we have assumed that they are proportional to the population size, i.e. $\lambda_n = n\lambda$ and $\mu_n = n\mu$, but in general they can have a richer functional form which, however, must always satisfy the requirement $\lambda_0 = \mu_0 = 0$: in a population with size $n = 0$ nothing can be born, and such a population can not “die further”.

Birth and death are independent Poisson processes; the probability that a population of size n faces b births and d deaths in the interval $(t, t + \Delta t]$ is therefore the product of individual probabilities,

$$\frac{(\lambda_n \Delta t)^b \mathrm{e}^{-\lambda_n \Delta t}}{b!} \cdot \frac{(\mu_n \Delta t)^d \mathrm{e}^{-\mu_n \Delta t}}{d!}.$$

If Δt is small enough, only the terms with $b = 0, 1$ and $d = 0, 1$ may be considered. Let us use

$$\Delta X(t, t + \Delta t) = X(t + \Delta t) - X(t)$$

to denote the change in population size (increase or decrease) in the time interval $(t, t + \Delta t]$. We discuss a population model with the properties

$$\begin{aligned}
\Delta X = 1 \quad &: \quad P\big(\Delta X(t, t + \Delta t) = 1 \,\big|\, X(t) = n\big) = \lambda_n \Delta t + \mathcal{O}(\Delta t),\\
\Delta X = -1 &: P\big(\Delta X(t, t + \Delta t) = -1 \,\big|\, X(t) = n\big) = \mu_n \Delta t + \mathcal{O}(\Delta t),\\
\Delta X = 0 \quad &: \quad P\big(\Delta X(t, t + \Delta t) = 0 \,\big|\, X(t) = n\big) = 1 - (\lambda_n + \mu_n)\Delta t + \mathcal{O}(\Delta t),\\
\text{otherwise} &: P\big(|\Delta X(t, t + \Delta t)| > 1 \,\big|\, X(t) = n\big) = \mathcal{O}(\Delta t),
\end{aligned}$$

whose dynamics is determined by the system of differential equations

$$p_n(t + \Delta t) = \underbrace{\lambda_{n-1} \Delta t \, p_{n-1}(t)}_{\text{birth}} + \underbrace{\big[1 - (\lambda_n + \mu_n)\Delta t\big] p_n(t)}_{\text{birth or death}} + \underbrace{\mu_{n+1} \Delta t \, p_{n+1}(t)}_{\text{death}} + \mathcal{O}(\Delta t),$$

where $n = 0, 1, 2, \ldots$ The probabilities $p_n(t)$ can be arranged in the vector

$$\boldsymbol{p}(t) = \big(p_0(t), p_1(t), p_2(t), \ldots\big)^{\mathrm{T}}. \tag{14.7}$$

Its dimension depends on the expected population dynamics. In the simple death model with a population of initial size N one needs a $(N + 1)$-dimensional vector to accommodate all possible population sizes between 0 and N. In the simple birth model one needs an infinite-dimensional vector in principle, but in a computer implementation it is whittled down according to the sizes we wish to monitor. The coefficients of the system are stored in the matrix

$$M = \begin{pmatrix} 1-(\lambda_0+\mu_0)\Delta t & \mu_1\Delta t & 0 & 0 & \cdots \\ \lambda_0\Delta t & 1-(\lambda_1+\mu_1)\Delta t & \mu_2\Delta t & 0 & \cdots \\ 0 & \lambda_1\Delta t & 1-(\lambda_2+\mu_2)\Delta t & \mu_3\Delta t & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

so that it can be expressed in matrix form

$$\boldsymbol{p}(t+\Delta t) = M\boldsymbol{p}(t),$$

and precisely the way it has been written it is also solved (given the specific initial condition)—see Example on page 356.

By analogy to the simple birth and death processes, the $\Delta t \to 0$ limit yields the corresponding (tridiagonal) system of linear differential equations

$$\dot{p}_n(t) = \lambda_{n-1}p_{n-1}(t) - (\lambda_n+\mu_n)p_n(t) + \mu_{n+1}p_{n+1}(t), \qquad (14.8)$$

where $n = 0, 1, 2, \ldots$ The system is solved with the initial conditions

$$p_N(0) = 1, \qquad p_n(0) = 0 \quad (n \neq N), \qquad (14.9)$$

or with a more general condition $\boldsymbol{p}(0) = \boldsymbol{p}_0$.

14.3.1 Equilibrium State

Does an equilibrium state exist for the system (14.8), i.e. a stationary (stable) distribution with the property $\dot{\boldsymbol{p}} = \boldsymbol{0}$? This question can be answered by setting all $\dot{p}_n(t)$ to zero and noticing that $\lambda_0 = \mu_0 = 0$:

$$\begin{aligned} &\dot{p}_0 = \mu_1 p_1 = 0 && \Longrightarrow p_1 = 0, \\ &\dot{p}_1 = -(\lambda_1+\mu_1)p_1 + \mu_2 p_2 = 0 && \Longrightarrow p_2 = 0, \\ &\dot{p}_2 = \lambda_1 p_1 - (\lambda_2+\mu_2)p_2 + \mu_3 p_3 = 0 && \Longrightarrow p_3 = 0, \end{aligned}$$

and so on. The only equilibrium state is therefore the state of total extinction, $p_0(t) = 1$, as all $p_i(t)$ must sum to 1 at any t.

14.3.2 General Solution in the Case $\lambda_n = N\lambda$, $\mu_n = N\mu$, $\lambda \neq \mu$

In the case that natality and mortality rates are proportional to the number of individuals in the population and different from each other,

$$\lambda_n = n\lambda, \qquad \mu_n = n\mu, \qquad \lambda, \mu > 0, \qquad \lambda \neq \mu,$$

the solution of the system (14.8) with initial condition (14.9) can be written in closed form, depending on the initial population size N [3]. If the process begins with a single individual, $N = 1$, corresponding to $\boldsymbol{p}(0) = (0, 1, 0, 0, 0, \ldots)^{\mathrm{T}}$, the solution is

$$\begin{aligned} p_0^{(N=1)}(t) &= \mu\rho, \\ p_n^{(N=1)}(t) &= \big(\lambda\rho\big)^n \left[\mu\rho - \frac{\lambda - \mu a}{\big(\mu - \lambda a\big)\lambda\rho}\right], \qquad n \geq 1, \end{aligned}$$

where

$$\rho = \frac{1-a}{\mu - \lambda a}, \qquad a = \mathrm{e}^{(\lambda-\mu)t}.$$

For $N \geq 2$ the general solution is

$$\begin{aligned} p_0(t) &= \big(\mu\rho\big)^N, \\ p_n(t) &= \sum_{i=0}^{i_{\max}} \binom{N}{i} \binom{N+n-i-1}{N-1} \big(\mu\rho\big)^{N-i} \big(\lambda\rho\big)^{n-i} \big[1 - (\lambda+\mu)\rho\big]^i, \end{aligned} \tag{14.10}$$

where $n \geq 1$ and $i_{\max} = \min\{N, n\}$.

14.3.3 General Solution in the Case $\lambda_n = N\lambda$, $\mu_n = N\mu$, $\lambda = \mu$

If natality and mortality rates are equal, $\lambda = \mu$, the corresponding formulas for $p_0(t)$ and $p_n(t)$ are obtained by taking the $\lambda \to \mu$ limit in the above expressions. (The rule of l'Hôpital comes to the rescue.)

14.3.4 Extinction Probability

The probability that a population dies out after time t (the *extinction probability*) is coded in the zeroth element of vector $\boldsymbol{p}$:

$$p_0(t) = \begin{cases} \left(\dfrac{\mu\big(\mathrm{e}^{(\lambda-\mu)t} - 1\big)}{\lambda\,\mathrm{e}^{(\lambda-\mu)t} - \mu}\right)^N & ;\ \lambda \neq \mu, \\[2ex] \left(\dfrac{\lambda t}{1+\lambda t}\right)^N, & ;\ \lambda = \mu. \end{cases}$$

Therefore

$$\lim_{t\to\infty} p_0(t) = \begin{cases} 1 & ;\ \lambda \leq \mu, \\ \left(\dfrac{\mu}{\lambda}\right)^N & ;\ \lambda > \mu. \end{cases}$$

Even if natality and mortality rates are equal, the population certainly dies out!

14.3.5 Moments of the Distribution $P(t)$ in the Case $\lambda_n = n\lambda$, $\mu_n = n\mu$

Apart from the dynamics with initial condition $\boldsymbol{p}(0) = \big(0, 0, 0, \ldots, 1, \ldots, 0, 0, 0\big)^{\mathrm{T}}$ corresponding to an exact initial population size N, we would like to understand the time evolution of a population whose size at time t (possibly $t = 0$) has a more general distribution, e.g. $\big(0, 0, 0, \ldots, 0.1, 0.2, 0.4, 0.2, 0.1, \ldots, 0, 0, 0\big)^{\mathrm{T}}$. For this we need the ith moment of the variable $X(t)$ with distribution $\boldsymbol{p}(t)$,

$$M_i(t) = \sum_{n=0}^{\infty} n^i\, p_n(t).$$

We insist on the form $\lambda_n = n\lambda$, $\mu_n = n\mu$ and calculate the time derivative of the first moment, $\dot{M}_1(t)$. This is done by rewriting (14.8) and considering $p_{-1} = 0$,

$$\begin{aligned} \dot{M}_1 &= \sum\nolimits_{n=0}^{\infty} n\,\dot{p}_n = \sum\nolimits_{n=0}^{\infty} n\big[\lambda(n-1)p_{n-1} - (\lambda+\mu)np_n + \mu(n+1)p_{n+1}\big] \\ &= \big[-(\lambda+\mu) + 2\lambda\big]p_1 + \big[-4(\lambda+\mu) + 2\mu + 6\lambda\big]p_2 + \cdots \\ &= (\lambda-\mu)p_1 + 2(\lambda-\mu)p_2 + 3(\lambda-\mu)p_3 + \cdots \\ &= (\lambda-\mu)\sum\nolimits_{n=0}^{\infty} np_1 = (\lambda-\mu)M_1. \end{aligned} \tag{14.11}$$

We have obtained the equation $\dot{M}_1(t) = (\lambda - \mu)M_1(t)$ with the solution

$$M_1(t) = M_1(0)\,\mathrm{e}^{(\lambda-\mu)t}$$

and the message: if $\lambda > \mu$, the mean of the population distribution (its “center of gravity”) exponentially diverges; if $\lambda < \mu$, it exponentially decreases to zero; if

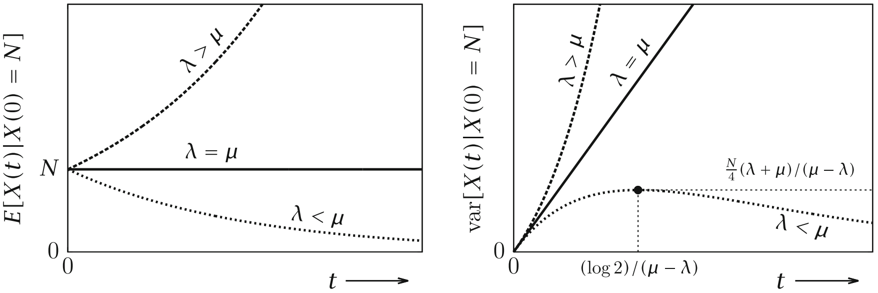


Fig. 14.3 Moments of the distribution $\boldsymbol{p}(t)$ or random variable $X(t)$ with initial condition $X(0) = N$ in the case of dominating births ($\lambda > \mu$), dominating deaths ($\lambda < \mu$) and (quasi) equilibrium ($\lambda = \mu$). [Left] Expected value. [Right] Variance

$\lambda = \mu$, the average does not change—which does not mean that $X(t)$ does not change! With a sharp initial condition $p_N(0) = 1$, which means precisely $M_1(0) = N$, the same realization can be written as

$$E\big[X(t)|X(0) = N\big] = N\,\mathrm{e}^{(\lambda-\mu)t}, \tag{14.12}$$

as shown in Fig. 14.3 (left).

A similar calculation yields the second moment. By analogy to (14.11) we obtain the differential equation $\dot{M}_2(t) = 2(\lambda - \mu)M_2(t) + (\lambda + \mu)M_1(t)$ with the solution

$$M_2(t) = M_2(0)\,\mathrm{e}^{2(\lambda-\mu)t} + M_1(0)\frac{\lambda+\mu}{\lambda-\mu}\,\mathrm{e}^{(\lambda-\mu)t}\left[\mathrm{e}^{(\lambda-\mu)t} - 1\right], \qquad \lambda \neq \mu.$$

Hence the variance is

$$\sigma^2(t) = M_2(t) - \big(M_1(t)\big)^2 = \sigma^2(0)\,\mathrm{e}^{2(\lambda-\mu)t} + M_1(0)\frac{\lambda+\mu}{\lambda-\mu}\,\mathrm{e}^{(\lambda-\mu)t}\left[\mathrm{e}^{(\lambda-\mu)t} - 1\right]$$

where $\sigma^2(0) = M_2(0) - \big(M_1(0)\big)^2$. If the initial size of the population is sharply defined, this expression can be further simplified, since then the initial variance is zero, $\sigma^2(0) = 0$, while the mean is $M_1(0) = N$:

$$\mathrm{var}\big[X(t)|X(0) = N\big] = \begin{cases} N\,\dfrac{\lambda+\mu}{\lambda-\mu}\,\mathrm{e}^{(\lambda-\mu)t}\left[\mathrm{e}^{(\lambda-\mu)t} - 1\right] & ;\ \lambda \neq \mu, \\ 2N\mu t & ;\ \lambda = \mu. \end{cases} \tag{14.13}$$

The time evolution of the variance in three typical dynamical regimes ($\lambda > \mu$, $\lambda = \mu$ and $\lambda < \mu$) is shown in Fig. 14.3 (right).

Example Consider the example of stochastic analysis of a population with natality rate $\lambda = 0.2$ and mortality rate $\mu = 0.4$. We are interested in the time evolution of the size of the population with initial size $X(0) = N = 100$ on the interval $t \in [0, 3]$. All we need is the recipe

$$X(t + \Delta t) = X(t) + \mathcal{P}\big(\lambda X(t)\Delta t\big) - \mathcal{P}\big(\mu X(t)\Delta t\big), \tag{14.14}$$

where $\mathcal{P}$ denotes a random number generated according to the Poisson distribution with the specified mean parameter (compare to (14.6)).

First we calculate the "paths" traced by the population (compare to Fig. 14.1). Figure 14.4 (left) shows 100 such paths. Mortality exceeds natality, thus *on average* exponential decay (14.12) is observed. Yet in spite of $\lambda < \mu$ a few paths even meander beyond $X(t) > N$ at short times. The path "fan" spreads out, as predicted by (14.13), although at time $t \approx (\log 2)/(\mu - \lambda) \approx 3.5$ it should begin to shrink according to

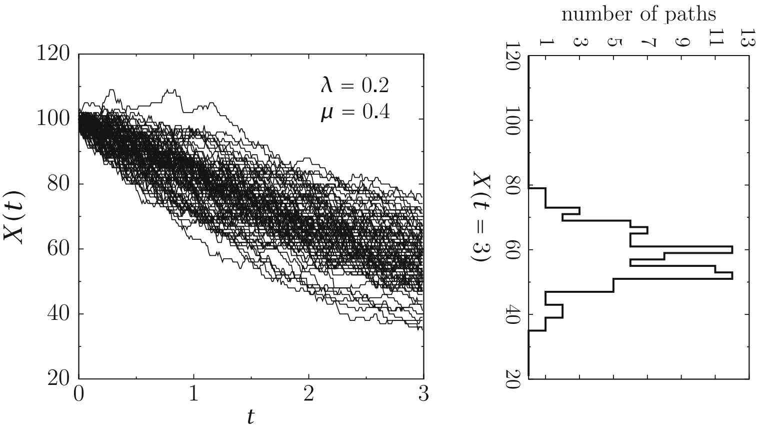


Fig. 14.4 [Left] Simulation of 100 random paths generated according to (14.14) with $X(0) = N = 100$, $\lambda = 0.2$, $\mu = 0.4$ and $\Delta t = 0.01$ until time $t = 3$. Compare to Figs. 14.1 and 1.5 (*left*). [Right] The distribution of final states for all 100 paths

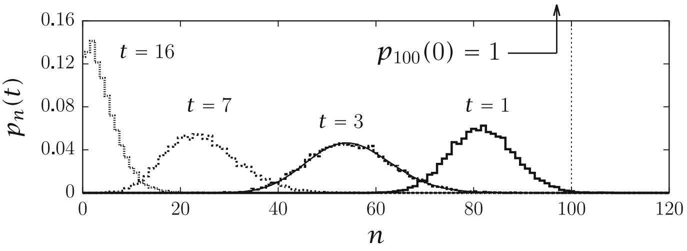


Fig. 14.5 Components of $\boldsymbol{p}(t)$ expressing the state of the population with exact initial size $N = 100$ after time t, i.e. the fraction of paths ending in $X(t) = n$. The *arrow* indicates the initial distribution. For $t = 3$ the analytic solution (14.10) is also plotted

the lower curve in Fig. 14.3 (right). The distribution of the final states $X(t)$ at $t = 3$ for all paths is shown in Fig. 14.4 (right).

The more paths we simulate, the smoother the histogram, which is nothing but a snapshot of the vector $\boldsymbol{p}(t)$ (14.7) at arbitrary time. Let us compute a larger set (10,000 instead of 100) random paths, all started with $X(0) = N = 100$, i.e. $p_{100}(0) = 1$. The components of the vector $\boldsymbol{p}(t)$ at times $t = 1, 3, 7$ and 16 are shown in Fig. 14.5. The population dies out, and of the distribution $\boldsymbol{p}(t)$ in the $t \to \infty$ limit only the $p_0(t) = 1$ component survives. ◁

14.4 Concluding Example: Rabbits and Foxes

The time evolution of a population becomes truly interesting when we consider immigration and emigration, external factors like finite amounts of nutrients or energy, and in particular when we treat several populations and their mutual interactions. So far we have only learned the alphabet: thence the vast expanse of population dynamics opens which is beyond the scope of this chapter and this textbook.

Nevertheless, let us discuss a simple problem of two populations in order to gain at least some insight into the richness of possible states and their inter-dependence. It is the classical conflict of rabbits and foxes. If food is abundant and there are no threats, the rabbits multiply with natality rate λ_1 and die (due to old age or disease) with mortality rate μ_1. The corresponding parameters for the foxes are λ_2 and μ_2. But the crucial ingredient is the interaction between the two: in order for a fox to catch the rabbit, they must meet, so the probability of their meeting is proportional to the product of probabilities that the rabbit (R) and the fox (F) happen to be at the same place at the same time. We can then imagine the random variables $R(t)$ and $F(t)$ to be some sort of time-dependent "concentrations" of rabbits and foxes, and the product $R(t)F(t)$ a kind of measure for the success of the hunt. With this guideline we write the differential equations

$$\begin{aligned}\dot{R}(t) &= \lambda_1 R(t) - \mu_1 R(t) - \gamma R(t)F(t),\\ \dot{F}(t) &= \lambda_2 F(t) - \mu_2 F(t) + \delta R(t)F(t),\end{aligned}$$

where $\gamma > 0$ and $\delta > 0$ are interaction parameters. The last term in the first equation is negative, since the foxes are killing the rabbits. The last term in the second equation is positive, as the fox population is being strengthened.

Assume that $\lambda_1/\mu_1 = \mu_2/\lambda_2 = 5/4$, so in truth only two parameters are genuinely free, since we can write $\lambda_1 = 5\alpha$, $\mu_1 = 4\alpha$, $\lambda_2 = 4\beta$ and $\mu_2 = 5\beta$. Suppose that the rabbit and fox populations are in equilibrium with $R_0 = 200$ rabbits and $F_0 = 50$ foxes. Equilibrium means $\dot{R}(t) = \dot{F}(t) = 0$, thus

$$\begin{aligned}\lambda_1 R_0 - \mu_1 R_0 - \gamma R_0 F_0 &= 0,\\ \lambda_2 F_0 - \mu_2 F_0 + \delta R_0 F_0 &= 0.\end{aligned}$$

The equilibrium condition allows us to compute the interaction parameters,

$$\begin{aligned} 5\alpha R_0 - 4\alpha R_0 &= \gamma R_0 F_0 \Longrightarrow \gamma = \alpha/F_0, \\ 4\beta F_0 - 5\beta F_0 &= -\delta R_0 F_0 \Longrightarrow \delta = \beta/R_0, \end{aligned}$$

so the original system of differential equations can be written as

$$\begin{aligned} \dot{R}(t) &= 5\alpha R(t) - 4\alpha R(t) - (\alpha/F_0)\, R(t)F(t), \\ \dot{F}(t) &= 4\beta F(t) - 5\beta F(t) + (\beta/R_0)\, R(t)F(t). \end{aligned}$$

This system with initial conditions $R(0) = R_0$ and $F(0) = F_0$ can be solved deterministically, i.e. by a suitable program for integration of differential equations. Our question about the state of the populations at a later time t will be given a unique answer. But we can also solve it stochastically, so that in each term we draw a Poisson probability for the increase or decrease of the population with the argument which is a product of the growth or decay parameter, the current population size, and the step length Δt. In short, we repeat the recipe (14.14), except that we now have two interacting populations. We therefore initialize the populations with $R(0)$ and $F(0)$ and enter the loop

$$\begin{aligned} R(t+\Delta t) &= R(t) + \mathcal{P}\big(5\alpha R(t)\Delta t\big) - \mathcal{P}\big(4\alpha R(t)\Delta t\big) - \mathcal{P}\big((\alpha/F_0)R(t)F(t)\Delta t\big), \\ F(t+\Delta t) &= F(t) + \mathcal{P}\big(4\beta F(t)\Delta t\big) - \mathcal{P}\big(5\beta F(t)\Delta t\big) + \mathcal{P}\big((\beta/R_0)R(t)F(t)\Delta t\big), \end{aligned}$$

which is repeated until one of the populations dies out. Two examples of random population "paths" generated in this manner are shown in Fig. 14.6 (top). The two panels at bottom show the corresponding phase portraits.

We see in Fig. 14.6 (top left) that at $0 \lesssim t \lesssim 200$ an approximately constant number of rabbits are available, which benefits the foxes. At $t \approx 200$ the rabbits become a rare commodity, so the fox population dwindles soon thereafter. This is swiftly exploited by the rabbits which happily multiply after $t \approx 300$; this again aids the foxes, and the rabbits are mercilessly devoured up to $t \approx 500$. But this also implies that the food becomes scarce for the foxes as well, so they, too, almost perish.

Figure 14.6 (top right) shows a more interesting case of nearly periodic exchange of predator and prey resurrections: the fox population recovers shortly after the rabbit population culminates. Intervals with negative time derivative of $R(t)$ correspond to intervals with positive derivatives of $F(t)$. As an exercise, repeat the simulation many times and plot the distribution with respect to extinction times of rabbits and foxes as in Fig. 14.2! In the meanwhile, pause to ponder upon miraculous Nature that has managed to sustain such periodicity by using its own Monte–Carlo method for eons!

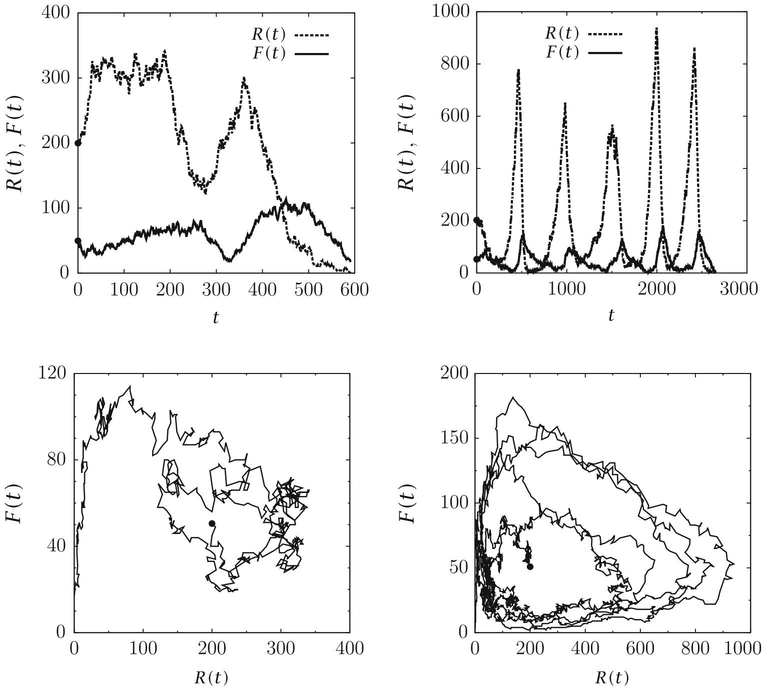


Fig. 14.6 Modeling Poissonian population dynamics of rabbits and foxes with parameters $\alpha = 2$, $\beta = 1$, $R_0 = 200$, $F_0 = 50$ and $\Delta t = 0.01$. [Top left and right] Sample time evolutions of $R(t)$ and $F(t)$. [Bottom left and right] Phase portraits

References

1. J.H. Matis, T.R. Kiffe, *Stochastic Population Models. A Compartmental Perspective*. Lecture Notes in statistics, vol. 145 (Springer, Berlin, 2000)
2. L.J.S. Allen, *Stochastic Population and Epidemic Models* (Springer, Cham, 2015)
3. L.M. Ricciardi, in: *Biomathematics Mathematical Ecology*, eds. by T.G. Hallam, S.A. Levin. Stochastic Population Theory: Birth and Death Processes, vol 17 (Springer, Berlin, 1986) p. 155
4. L.M. Ricciardi, in: *Biomathematics Mathematical Ecology*, eds. by T.G. Hallam, S.A. Levin. Stochastic Population Theory: Birth and Death Processes, vol 17 (Springer, Berlin, 1986) p. 191
5. `gsl_ran_poisson` in GSL library, http://www.gnu.org/software/gsl/, or `poidev` in *Numerical Recipes*, eds. W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling. Numerical Recipes: The Art of Scientific Computing, 3rd edn. (Cambridge University Press, Cambridge, 2007)

Appendix A
Probability as Measure ⋆

Abstract Probability is defined in a mathematically strict manner as a measure. The Dirac delta is defined.

Here we give a mathematically more strict definition of probability and the Dirac delta "function" (phenomenologically introduced in (2.1)) as *measures*.

σ-Algebra

Let X be a non-empty set. A family of subsets $\mathcal{A}$ of X is called a *σ-algebra on X* if it has the following properties:

1. $X \in \mathcal{A}$;
2. for each subset $S \in \mathcal{A}$ also $X \backslash S \in \mathcal{A}$;
3. for each countable family $\{A_i : i \in \mathbb{N}\}$ of elements from $\mathcal{A}$, the union $\bigcup_{i \in \mathbb{N}} A_i$ also belongs to $\mathcal{A}$.

The elements of the family $\mathcal{A}$ are called *measurable sets*, and the set X, furnished with $\mathcal{A}$, is called a *measurable space*. A measurable space is the pair $(X, \mathcal{A})$.

Positive Measure

Examples of positive measures are: length of subset (interval) in $\mathbb{R}$, area of planar geometric shapes, volume of bodies in space. To generalize these special cases to arbitrary measurable spaces, one defines a *positive measure* (or simply *measure*) on a measurable space $(X, \mathcal{A})$ as the function

$$\mu : \mathcal{A} \to [0, \infty],$$

satisfying the conditions

1. $\mu(\{\,\}) = 0$ and
2. $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$

for each countable family of disjoint subsets A_n and $\mathcal{A}$. A positive measure μ on a measurable space $(X, \mathcal{A})$ is called a *finite measure* if $\mu(X) < \infty$.

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6

Probability as a Positive Measure

For random experiments with sample space S we define the event space $\mathcal{E}$, which is the power set of S, i.e. the set of all subsets of S, including the empty set and S itself. The mapping $P : \mathcal{E} \to \mathbb{R}$ is called a *probability measure* on measurable space $(S, \mathcal{E})$ if the following holds true:

1. $P(A) \geq 0$ for each $A \in \mathcal{E}$;
2. $P(S) = 1$;
3. if $A_1, A_2, \ldots$ are mutually exclusive events in $\mathcal{E}$, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

In simple cases, e.g. in throwing a die, $\mathcal{E}$ may indeed be identified with the power set of S, but often we restrict ourselves to a much smaller set: for example, it turns out [1] that there does not exist a probability measure P that would be defined on *all* subsets of the interval $[0, 1]$ and would satisfy the requirement $P(\{x\}) = 0$ for each $x \in [0, 1]$.

Dirac Measure

Let X be an arbitrary non-empty set, $\mathcal{A}$ its power set, and $x \in X$ its arbitrary element. Then the prescription

$$\mu(A; x) = \begin{cases} 1 \; ; \; x \in A, \\ 0 \; ; \; x \in X \backslash A, \end{cases}$$

defines a positive finite measure on measurable space $(X, \mathcal{A})$ called the *Dirac measure at* x and denoted by δ_x. For each function $f : X \to \overline{\mathbb{R}}$ the integration with respect to the Dirac delta represents the evaluation of the function at x,

$$\int_X f \, \mathrm{d}\delta_x = f(x).$$

In the special case $X = \mathbb{R}$ we have

$$\int_A f \, \mathrm{d}\delta_x = f(x)\delta_x(A) = \int_A f(t)\delta(t - x) \, \mathrm{d}t, \qquad A \subseteq X,$$

for each measurable function $f : \mathbb{R} \to \mathbb{R}$, where $\mathrm{d}\delta_x(t) = \delta(t - x) \, \mathrm{d}t$ and δ is the Dirac delta "function".

Reference

1. G. Grimmett, D. Welsh, *Probability. An introduction*, 2nd edn. (Oxford University Press, Oxford, 2014)

Appendix B
Generating and Characteristic Functions ⋆

Abstract Generating and characteristic functions are introduced as transformations of random variables that facilitate the calculation of certain distribution properties, in particular its moments and their convolutions. The problems of inverting probability-generating and characteristic functions, as well as of the existence of generating functions are presented.

Generating and characteristic functions are transformations of probability functions. As such they are not as easy to interpret as the distributions themselves, but in certain cases they offer immense benefits in terms of elegant calculation of distribution properties—for example, their moments—or quantities relating the distribution to each other, in particular their convolutions.

B.1 Probability-Generating Functions

Generating functions are applicable to random variables whose possible values are non-negative integers or their subsets. Such variables are called *non-negative integer random variables*. Let X be such a variable with the probability function

$$f_n = P(X = n). \tag{B.1}$$

Then the function of a real variable

$$G_X(z) = \sum_{n=0}^{\infty} P(X = n) z^n = \sum_{n=0}^{\infty} f_n z^n, \qquad |z| \le 1,$$

is the *[probability]-generating function* of the random variable X, distributed according to (B.1). The coefficients in this power expansion are probabilities with values

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6

between 0 and 1. Since they are bounded, the series is absolutely convergent for any $|z| < 1$, and due to

$$G(1) = \sum_{n=0}^{\infty} f_n = 1$$

the series converges at least on $[-1, 1]$. By comparison to (4.7) we also see that the generating function is equal to the expected value of the random variable z^X,

$$G_X(z) = E\big[z^X\big]. \tag{B.2}$$

The generating function G_X uniquely determines the probability function of X. This can be seen if we take the derivative of the series with respect to z:

$$\frac{\mathrm{d}^r}{\mathrm{d}z^r} G_X(z) = \sum_{n=r}^{\infty} n(n-1)\cdots(n-r+1) z^{n-r} f_n, \qquad r = 1, 2, \ldots$$

Namely, by setting $z = 0$ we obtain

$$f_r = P(X = r) = \frac{1}{r!} \left[\frac{\mathrm{d}^r}{\mathrm{d}z^r} G_X(z) \right] \bigg|_{z=0}, \qquad r = 0, 1, 2, \ldots, \tag{B.3}$$

so indeed by taking consecutive derivatives the complete distribution is determined, as all its components f_r are combed through. Why does this matter? Frequently only the generating function of X is available, while its probability function is not explicitly known. In such cases its components can be calculated by using (B.3): see Sect. B.1.2. Besides, taking the derivatives of the generating function is an easy way to produce the moments of X. For instance, by taking the first and second derivative we get

$$G'_X(z) = \sum_{n=1}^{\infty} n z^{n-1} f_n, \qquad G''_X(z) = \sum_{n=2}^{\infty} n(n-1) z^{n-2} f_n.$$

On the other hand,

$$E[X] = \sum_{n=1}^{\infty} n f_n = \lim_{z \nearrow 1} G'_X(z) = G'_X(1), \tag{B.4}$$

$$E[X(X-1)] = \sum_{n=2}^{\infty} n(n-1) f_n = \lim_{z \nearrow 1} G''_X(z) = G''_X(1),$$

therefore

$$\begin{aligned} E[X^2] &= E[X(X-1)] + E[X] = G''_X(1) + G'_X(1), \\ \mathrm{var}[X] &= E[X^2] - (E[X])^2 = G''_X(1) + G'_X(1)\big[1 - G'_X(1)\big]. \end{aligned} \tag{B.5}$$

Individual moments can be calculated without such detours by using the formula

$$E[X^r] = \left[\left(z \frac{\mathrm{d}}{\mathrm{d}z} \right)^r G_X(z) \right] \bigg|_{z=1} .$$

Example The generating function of the binomial distribution (Definition (5.1)) with the probability function $f_n = P(X = n; N, p)$ is

$$G_X(z) = \sum_{n=0}^{N} f_n z^n = \sum_{n=0}^{N} \binom{N}{n} p^n q^{N-n} z^n = \sum_{n=0}^{N} \binom{N}{n} (pz)^n q^{N-n} = (pz + q)^N .$$

Its first derivative is $G'_X(z) = Np(pz + q)^{N-1}$, and the second derivative is $G''_X(z) = N(N-1)p^2(pz+q)^{N-2}$, thus $G'_X(1) = Np$ and $G''_X(1) = N(N-1)p^2$. From (B.4) and (B.5) it follows that

$$E[X] = Np, \qquad \mathrm{var}[X] = N(N-1)p^2 + Np(1 - Np) = Npq,$$

which are familiar expressions (5.5). ◁

Example The generating function of the Poisson distribution (5.11) is

$$G_X(z) = \sum_{n=0}^{\infty} f_n z^n = \sum_{n=0}^{\infty} \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!} z^n = \mathrm{e}^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda z)^n}{n!} = \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda z} = \mathrm{e}^{\lambda(z-1)} .$$

Differentiation gives $G'_X(z) = \lambda \mathrm{e}^{\lambda(z-1)}$ and $G''_X(z) = \lambda^2 \mathrm{e}^{\lambda(z-1)}$, hence $G'_X(1) = \lambda$ and $G''_X(1) = \lambda^2$. From (B.4) and (B.5) it follows that

$$E[X] = \lambda, \qquad \mathrm{var}[X] = \lambda^2 + \lambda(1 - \lambda) = \lambda,$$

which, again, we know from (5.12). ◁

B.1.1 Generating Functions and Convolution

Let us discuss mutually independent integer random variables X and Y with the probability functions

$$f_n = P(X = n), \qquad g_n = P(Y = n), \qquad n = 0, 1, 2, \ldots$$

Their sum $Z = X + Y$ is also an integer random variable with the corresponding probability function

$$h_n = P(Z = n), \qquad n = 0, 1, 2, \ldots$$

The sum Z has value n when the variables X and Y have values $(X, Y) = (0, n)$ or $(1, n-1)$ or $(2, n-2)$, and so on. Since X and Y are independent, the probabilities of these simultaneous events are $P(X=0)P(Y=n)$ or $P(X=1)P(Y=n-1)$ or $P(X=2)P(Y=n-2)$, and so on. In other words,

$$h_n = P(Z=n) = \sum_{j=0}^{n} P(X=j)P(Y=n-j) = \sum_{j=0}^{n} f_j g_{n-j}, \qquad n = 0, 1, 2, \ldots$$

We are looking at a discrete convolution of the sequences $\{f_n\}$ and $\{g_n\}$, which we denote as

$$\{h_n\} = \{f_n\} * \{g_n\}.$$

This a discrete analogue of the Definition (6.1) or

$$h_Z(z) = \int_{-\infty}^{\infty} f_X(x) g_Y(z-x)\,\mathrm{d}x = \int_{-\infty}^{\infty} f_X(z-y) g_Y(y)\,\mathrm{d}y.$$

Where do generating functions come into play? Let

$$G_X(z) = \sum_{n=0}^{\infty} f_n z^n, \qquad G_Y(z) = \sum_{n=0}^{\infty} g_n z^n$$

be generating functions of X and Y. The generating function of their sum is then

$$G_Z(z) = \sum_{n=0}^{\infty} h_n z^n = \sum_{n=0}^{\infty} \left(\sum_{j=0}^{n} f_j g_{n-j} \right) z^n = \sum_{n=0}^{\infty} \sum_{j=0}^{n} f_j z^j g_{n-j} z^{n-j}.$$

The series on the right is just the product of the series $G_X(z)$ and $G_Y(z)$, so

$$G_Z(z) = G_X(z) G_Y(z). \tag{B.6}$$

The generating function of the sum of independent integer variables is therefore equal to the product of the generating functions of the two terms. An even faster way to this result would be to consider (B.2): if X and Y are independent, the variables $U = z^X$ and $V = z^Y$ are independent, too; since for independent variables U and V one has $E[UV] = E[U]E[V]$, this also means

$$E\big[z^{X+Y}\big] = E\big[z^X z^Y\big] = E\big[z^X\big]E\big[z^Y\big], \tag{B.7}$$

whence (B.6) follows. This should not be read in the opposite direction: having $G_Z(z) = G_X(z)G_Y(z)$ does not necessarily mean that X and Y are independent. But the relation *can* be generalized to several independent variables: if $X_1, X_2, \ldots, X_n$ are mutually independent random variables with generating functions $G_{X_1}(z), G_{X_2}(z), \ldots, G_{X_n}(z)$ and Z is their sum with the generating function $G_Z(z)$, then

$$G_Z(z) = G_{X_1}(z) G_{X_2}(z) \cdots G_{X_n}(z). \tag{B.8}$$

Multiplying generating functions is a much simpler operation than computing convolution sums, so convolution of independent integer random variables is most easily performed by using (B.6) and (B.8).

Example We demonstrate that the convolution of two Poisson distributions is a Poisson distribution. In the Example on p. 148 we have derived this result by a direct calculation of the convolution sum. But if one calls generating functions $G_X(z) = \mathrm{e}^{\lambda(z-1)}$ and $G_Y(z) = \mathrm{e}^{\mu(z-1)}$ to the rescue, the effort is minimal:

$$G_Z(z) = G_X(z) G_Y(z) = \mathrm{e}^{\lambda(z-1)} \mathrm{e}^{\mu(z-1)} = \mathrm{e}^{(\lambda+\mu)(z-1)}.$$

Clearly the variable Z has the generating function of the Poisson distribution with parameter $\lambda + \mu$, so indeed $Z \sim \text{Poisson}(\lambda + \mu)$. ◁

B.1.2 Inverting the Probability-Generating Function

The functional form $G_Z(z) = \mathrm{e}^{(\lambda+\mu)(z-1)}$ in the preceding Example immediately allowed us to conclude that Z is Poissonian, as we already knew the relation between the generating function and its inverse beforehand. The same procedure can be used for more complicated generating functions, as long as they can be split into sums of terms whose inverses are known.

But how do we compute the inverse of an arbitrary generating function? Formula (B.3) can be used for simple explicit functions, but analytic differentiation may be strenuous and is numerically unstable. The solution—in particular when the generating function is only known at discrete points—is offered by the Cauchy integral formula

$$G_X(a) = \frac{1}{2\pi\mathrm{i}} \oint_{\partial D} \frac{G_X(z)}{z - a} \,\mathrm{d}z,$$

where $D = \{z : |z - z_0| \le R\}$ is a subset completely contained in the definition domain of G_X (neighborhood of z_0), ∂D is its boundary and a is any point in the interior of D. For the nth derivative of G_X it holds that

$$G_X^{(n)}(a) = \frac{n!}{2\pi\mathrm{i}} \oint_{\partial D} \frac{G_X(z)}{(z - a)^{n+1}} \,\mathrm{d}z,$$

so the components f_n of the probability distribution of X—use of (B.3) requires derivatives of G_X at $a = 0$—are given by the integral

$$f_n = \frac{1}{2\pi \mathrm{i}} \oint_C \frac{G_X(z)}{z^{n+1}}\, \mathrm{d}z.$$

The closed curve C is a circle in the complex plane. By the substitution $z = R\,\mathrm{e}^{\mathrm{i}\,u}$, where R must be such that G_X is analytic on D, we get

$$f_n = \frac{1}{2\pi R^n} \int_0^{2\pi} G_X\big(R\,\mathrm{e}^{\mathrm{i}\,u}\big)\, \mathrm{e}^{-\mathrm{i}\,nu}\, \mathrm{d}u.$$

The integral can be evaluated by using the trapezoidal rule, resulting in the following approximation for the true distribution f_n $(n = 0, 1, \ldots, N-1)$:

$$\widetilde{f}_n \approx \frac{1}{NR^n} \sum_{m=0}^{N-1} G_X\left(R\,\mathrm{e}^{\mathrm{i}\,2\pi m/N}\right)\, \mathrm{e}^{-\mathrm{i}\,2\pi mn/N}, \qquad \widetilde{f}_{n+N} = \widetilde{f}_n, \tag{B.9}$$

which is the inverse discrete Fourier transformation scaled by R. Due to the discrete nature of the approximation, discretization and aliasing errors are thereby introduced (see, for example, [1], page 166), which can be controlled by the parameter R. For details see [2–4].

Example Let us pretend that we do not know the probability function of the Poisson distribution $f_n(\lambda) = \lambda^n \mathrm{e}^{-\lambda}/n!$, but only its generating function $G_X(z) = \mathrm{e}^{\lambda(z-1)}$. Take $\lambda = 2$, for instance: the exact values f_n up to $n = N = 20$ are shown in Fig. B.1 (left). We compute the approximations for f_n by inverting the generating function via (B.9) with different R, say, $R = 1.0$, $R = 2.0$ and $R = 0.5$. The absolute errors of these reconstructed probability functions are shown in Fig. B.1 (right). Note the absence of the value at $n = N$: due to the periodicity of the Fourier transform one has $f_N = f_0$. ◁

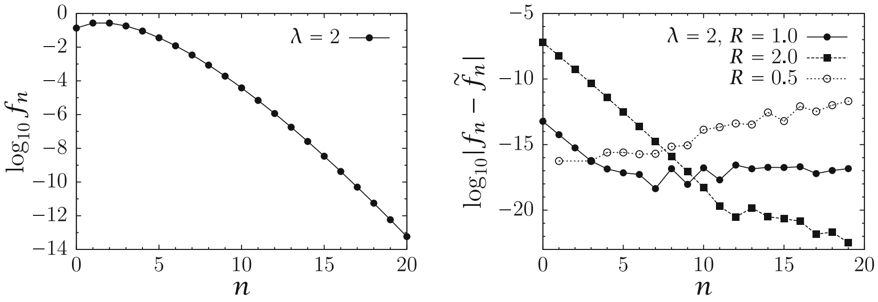


Fig. B.1 [Left] Poisson distribution with parameter $\lambda = 2$ in logarithmic scale. [Right] Difference between the exact values f_n and their approximations, calculated by inverting the probability-generating function by the discrete Fourier transformation (B.9), at several values of the parameter R

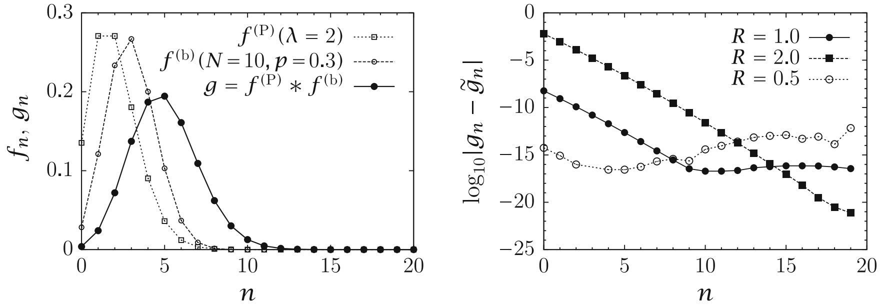


Fig. B.2 [Left] Poisson distribution with parameter $\lambda = 2$, binomial distribution with $N = 10$, $p = 0.3$, and their convolution. [Right] The difference between the exact probabilities g_n and their approximations $\widetilde{g}_n$, calculated by inverting the generating function and by using the discrete Fourier transformation with various values of R

Example It is instructive to compare the convolution of discrete distributions, calculated by the basic formula (6.4), and by multiplying generating functions according to (B.6). Take, for instance, the Poisson distribution

$$P(X = n) = f_n^{(\mathrm{P})} = \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!}, \qquad n = 0, 1, 2, \ldots,$$

where $\lambda = 2$, and the binomial distribution

$$P(Y = n) = f_n^{(\mathrm{b})} = \binom{N}{n} p^n q^{N-n}, \qquad n = 0, 1, 2, \ldots, N,$$

where $N = 10$ and $p = 0.3$. These distributions are shown in Fig. B.2 (left) at its left edge. By the definition of convolution we obtain the distribution

$$P(X + Y = n) = g_n = \left(f^{(\mathrm{P})} * f^{(\mathrm{b})}\right)_n = \sum_{i=0}^{n} f_i^{(\mathrm{P})} f_{n-i}^{(\mathrm{b})}, \tag{B.10}$$

indicated by full circles in the figure. We should expect the same result by multiplying the generating functions of both distributions and computing the inverse Fourier transformation of the product. Thus we compute

$$G_Z(z) = G_{X+Y}(z) = G_X(z) G_Y(z) = \mathrm{e}^{\lambda(z-1)}(pz + q)^N,$$

and then use this function in formula (B.9):

$$\widetilde{g}_n \approx \frac{1}{N_{\mathrm{DFT}} R^n} \sum_{m=0}^{N_{\mathrm{DFT}}-1} G_Z\left(R\,\mathrm{e}^{\mathrm{i}\,2\pi m/N_{\mathrm{DFT}}}\right) \mathrm{e}^{-\mathrm{i}\,2\pi mn/N_{\mathrm{DFT}}}.$$

(Think about it: what N_{DFT} should one take in the above equation and what is the range of n in (B.10), considering that the definition domains of the distributions differ?) We thereby obtain the probabilities $\widetilde{g}_n$ that should be equal to g_n. How well this holds is shown in Fig. B.2 (right). ◁

B.2 Moment-Generating Functions

Probability-generating functions have been defined for random variables with non-negative integer values. The concept can be extended to random variables with arbitrary real values, if $E[z^X]$ (see (B.2)) is replaced by $E\big[\mathrm{e}^{tX}\big]$. If this expected value is finite for t on the interval $[t - T, t + T]$ for some $T > 0$, we may define the *moment-generating function*

$$M_X(t) = E\big[\mathrm{e}^{tX}\big] = \int_{-\infty}^{\infty} \mathrm{e}^{tx} f_X(x)\,\mathrm{d}x, \tag{B.11}$$

which is nothing but the continuous Laplace transform. In the case of a discrete probability distribution of X, which we shall not discuss separately from now on, the corresponding definition is

$$M_X(t) = E\big[\mathrm{e}^{tX}\big] = \sum_i \mathrm{e}^{tx_i} P(X = x_i).$$

The 'moment-generating' attribute is easy to explain if one expands e^{tX} in a power series and exchanges the order of summation and taking the expected values:

$$E\big[\mathrm{e}^{tX}\big] = E\left[\sum_{k=0}^{\infty} t^k \frac{X^k}{k!}\right] = 1 + tE[X] + \frac{t^2}{2!}E\big[X^2\big] + \frac{t^3}{3!}E\big[X^3\big] + \cdots .$$

Namely, individual distribution moments can be obtained by taking consecutive derivatives

$$E\big[X^r\big] = \left[\frac{\mathrm{d}^r M_X(t)}{\mathrm{d}t^r}\right]\bigg|_{t=0}, \qquad r = 1, 2, \ldots, \tag{B.12}$$

thus $E[X] = M_X'(0)$, $E[X^2] = M_X''(0)$, and so on. Compare (B.3) and (B.12)!

Example Let us calculate the moment-generating function of a random variable distributed according to the Cauchy distribution (3.18):

$$M_X(t) = E\big[\mathrm{e}^{tX}\big] = \int_{-\infty}^{\infty} \mathrm{e}^{tx} \frac{1}{\pi}\frac{1}{1+x^2}\mathrm{d}x = \begin{cases} 1 & ;\ t = 0, \\ \infty & ;\ \text{otherwise.} \end{cases}$$

Now we see why Definition (B.11) had to be formulated so carefully: the expected value $E\big[\mathrm{e}^{tX}\big]$ for arbitrary t may not even exist! This obstacle will be circumvented in Sect. B.3. ◁

Example What about the moment-generating function of a random variable distributed according to the standardized normal distribution (3.10)? By elementary integration[1] we immediately obtain

$$M_X(t) = E\big[\mathrm{e}^{tX}\big] = \int_{-\infty}^{\infty} \mathrm{e}^{tx} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x = \mathrm{e}^{t^2/2}. \tag{B.13}$$

The main dish follows: we expand the exponential in a power series

$$M_X(t) = \mathrm{e}^{t^2/2} = \sum_{k=0}^{\infty} \frac{(t^2/2)^k}{k!} = \sum_{k=0}^{\infty} \frac{(2k)!}{k!2^k} \frac{t^{2k}}{(2k)!} = \sum_{k=0}^{\infty} E\big[X^k\big] \frac{t^k}{k!}$$

and compare the terms with equal powers of t on both sides of the last equality. This gives us the odd moments

$$E\big[X^k\big] = 0, \qquad k \text{ odd},$$

while the even ones are

$$E\big[X^{2k}\big] = \frac{(2k)!}{k!2^k} = 1 \cdot 3 \cdot 5 \cdot (2k-1),$$

thus $E\big[X^2\big] = 1$, $E\big[X^4\big] = 3$, $E\big[X^6\big] = 15$, $E\big[X^8\big] = 105$, and so on. The first two values should already be familiar from Sect. 4.7, while the others were derived elegantly, with minimal effort. ◁

Let X and Y be random variables with moment-generating functions $M_X(t)$ and $M_Y(t)$. If X and Y are mutually independent, the same reasoning that brought us to (B.6) implies also

$$M_{X+Y}(t) = M_X(t) M_Y(t). \tag{B.14}$$

For random variables X and Y related through $Y = aX + b$, it holds that

$$M_Y(t) = \mathrm{e}^{bt} M_X(at). \tag{B.15}$$

So the obvious generalization of (B.14) to a sum of several variables is at hand: if $X_1, X_2, \ldots, X_n$ are mutually independent random variables and $Y = c_1 X_1 +$

[1] Gaussian integrals with linear terms in the exponent can be handled by using the formulas

$$\int_{-\infty}^{\infty} \mathrm{e}^{-ax^2/2+bx}\,\mathrm{d}x = \sqrt{\frac{2\pi}{a}}\mathrm{e}^{b^2/2a}, \qquad \int_{-\infty}^{\infty} \mathrm{e}^{-ax^2/2+\mathrm{i}\,bx}\,\mathrm{d}x = \sqrt{\frac{2\pi}{a}}\mathrm{e}^{-b^2/2a}\,.$$

$c_2X_2 + \cdots + c_nX_n$ is their linear combination with real coefficients c_i, the moment-generating function of Y is equal to the product of moment-generating functions of individual variables X_i:

$$M_Y(t) = E\bigl[\mathrm{e}^{tY}\bigr] = \prod_{i=1}^{n} E\bigl[\mathrm{e}^{tX_i}\bigr] = M_{X_1}(c_1t)M_{X_2}(c_2t)\cdots M_{X_n}(c_nt). \tag{B.16}$$

Just as in (B.7) these recipes may not be read in reverse: $M_{X+Y}(t) = M_X(t)M_Y(t)$ does not necessarily mean that X and Y are independent.

Example The convolution problem from the Example on p. 148 can also be solved by generating functions. The moment-generating functions of X and Y are

$$M_X(t) = \sum_{n=-\infty}^{\infty} f_n \mathrm{e}^{tx_n}, \qquad M_Y(t) = \sum_{n=-\infty}^{\infty} g_n \mathrm{e}^{ty_n},$$

that is,

$$\begin{aligned} M_X(t) &= 0.15\,\mathrm{e}^{-3t} + 0.25\,\mathrm{e}^{-t} + 0.1\,\mathrm{e}^{2t} + 0.3\,\mathrm{e}^{6t} + 0.2\,\mathrm{e}^{8t}, \\ M_Y(t) &= 0.2\,\mathrm{e}^{-2t} + 0.1\,\mathrm{e}^{t} + 0.3\,\mathrm{e}^{5t} + 0.4\,\mathrm{e}^{8t}. \end{aligned}$$

Since X and Y are mutually independent, the moment-generating function $M_Z(t)$ of their sum $Z = X + Y$ is the product of the individual moment-generating functions $M_X(t)$ and $M_Y(t)$:

$$\begin{aligned} M_Z(t) = M_X(t)M_Y(t) &= \sum_n h_n \mathrm{e}^{tz_n} \\ &= 0.03\,\mathrm{e}^{-5t} + 0.05\,\mathrm{e}^{-3t} + 0.015\,\mathrm{e}^{-2t} + 0.045\,\mathrm{e}^{0t} + 0.045\,\mathrm{e}^{2t} \\ &\quad + 0.01\,\mathrm{e}^{3t} + 0.135\,\mathrm{e}^{4t} + 0.06\,\mathrm{e}^{5t} + 0.04\,\mathrm{e}^{6t} + 0.16\,\mathrm{e}^{7t} + 0.02\,\mathrm{e}^{9t} \\ &\quad + 0.04\,\mathrm{e}^{10t} + 0.09\,\mathrm{e}^{11t} + 0.06\,\mathrm{e}^{13t} + 0.12\,\mathrm{e}^{14t} + 0.08\,\mathrm{e}^{16t}. \end{aligned}$$

All we need, then, is to read off the coefficient in front of e^{4t}, which is $h_4 = P(Z = 4) = 0.135$, and analogously for any other h_n. ◁

B.3 Characteristic Functions

Let X be a real (discrete or continuous) random variable and t a non-random real variable. The quantity

$$\phi_X(t) = E\bigl[\mathrm{e}^{\mathrm{i}tX}\bigr], \quad t \in \mathbb{R}, \tag{B.17}$$

is called the *characteristic function* of the random variable X [5, 6]. In contrast to the moment-generating function a characteristic function exists regardless of the distribution of X, and its definition domain is the whole real axis. Any characteristic function satisfies

$$|\phi_X(t)| \le 1, \qquad \phi_X(0) = 1.$$

Besides, one has $\phi_X(t) = M_X(\mathrm{i}\,t)$ and $\phi_X(-\mathrm{i}\,t) = M_X(t)$, if M_X exists. If the distribution of X is discrete, with probability function $f_n = P(X = x_n)$, where $n = 0, 1, 2, \ldots$, it has the characteristic function

$$\phi_X(t) = \sum_{n=0}^{\infty} f_n \mathrm{e}^{\mathrm{i}\,t x_n}. \tag{B.18}$$

If the distribution is continuous, with probability density f_X, it corresponds to

$$\phi_X(t) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\,tx} f_X(x)\,\mathrm{d}x, \tag{B.19}$$

which is the usual Fourier transformation of f_X.

Example The Poisson distribution with the probability function

$$f_n = \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!}, \qquad n = 0, 1, 2, \ldots$$

has the characteristic function

$$\phi_X(t) = \sum_{n=0}^{\infty} f_n \mathrm{e}^{\mathrm{i}\,tn} = \sum_{n=0}^{\infty} \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!} \mathrm{e}^{\mathrm{i}\,tn} = \mathrm{e}^{-\lambda} \sum_{n=0}^{\infty} \frac{\left(\lambda \mathrm{e}^{\mathrm{i}\,t}\right)^n}{n!} = \mathrm{e}^{-\lambda}\mathrm{e}^{\lambda \mathrm{e}^{\mathrm{i}\,t}} = \mathrm{e}^{\lambda(\mathrm{e}^{\mathrm{i}\,t}-1)}.$$

Calculate also the corresponding moment-generating function! ◁

Example The standard normal distribution (3.10) has the characteristic function

$$\phi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\,tx} \mathrm{e}^{-x^2/2}\,\mathrm{d}x = \mathrm{e}^{-t^2/2},$$

while the equivalent for the non-standardized normal distribution (3.7) is

$$\phi_X(t) = \mathrm{e}^{\mathrm{i}\,\mu t - \sigma^2 t^2/2}, \tag{B.20}$$

where we have used the formula in (see footnote 1). ◁

The following important properties of characteristic functions are given without proof. If a and b are constants and $Y = aX + b$, it holds that

$$\phi_Y(t) = \mathrm{e}^{\mathrm{i}\,bt}\phi_X(at), \tag{B.21}$$

which is also seen from (B.15). If random variables $X_1, X_2, \ldots, X_n$ are mutually independent and $Y = c_1X_1 + c_2X_2 + \cdots + c_nX_n$ is their linear combination, then

$$\phi_Y(t) = \phi_{X_1}(c_1t)\phi_{X_2}(c_2t)\cdots\phi_{X_n}(c_nt). \tag{B.22}$$

This theorem also can not be reversed: having $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$ does not necessarily mean that X and Y are independent.

As the moment-generating functions, the characteristic functions, too, can be used to derive the statistical moments $E[X^n]$, $n = 0, 1, 2, \ldots$ Namely, if ϕ_X is at least p-times continuously differentiable at the origin, it holds that

$$E\big[X^n\big] = \frac{1}{\mathrm{i}^n}\left[\frac{\mathrm{d}^n\phi_X}{\mathrm{d}t^n}\right]\bigg|_{t=0}, \qquad n = 1, 2, \ldots, p.$$

There is a one-to-one correspondence between the characteristic function and the probability distribution: any two random variables X and Y have the same probability distribution precisely when $\phi_X = \phi_Y$, therefore

$$\phi_X = \phi_Y \quad \Leftrightarrow \quad X \sim Y.$$

Example Let us calculate the characteristic function of the binomial distribution

$$f_n = P(X = n) = \binom{N}{n} p^n q^{N-n}, \qquad n = 0, 1, 2, \ldots, N.$$

Imagine a Bernoulli (binomial) sequence of trials. To the jth trial in this sequence we assign a random variable Y_j with value 1 for a "good" event A (probability p), or value 0 for the complementary event $\overline{A}$ (probability $q = 1 - p$). Since the trials in the sequence are mutually independent, the same applies to the random variables Y_j. The variable X takes the value n if there were n occurrences of A in N trials: in this case precisely n variables Y_j have value 1, while the others are zero, hence $X = Y_1 + Y_2 + \cdots + Y_N$. For an individual Y_j we then use (B.18) to calculate

$$\phi_{Y_j}(t) = \sum_{k=0}^{1} P(Y_j = y_k)\mathrm{e}^{\mathrm{i}\,ty_k} = \underbrace{P(Y_j = 0)}_{q}\,\mathrm{e}^{\mathrm{i}\,t0} + \underbrace{P(Y_j = 1)}_{p}\,\mathrm{e}^{\mathrm{i}\,t1} = p\,\mathrm{e}^{\mathrm{i}\,t} + q.$$

By (B.22) we then obtain the characteristic function of the binomial distribution

$$\phi_X(t) = \big(\phi_{Y_j}(t)\big)^N = \big(p\mathrm{e}^{\mathrm{i}\,t} + q\big)^N, \tag{B.23}$$

which we shall use in the following. ◁

B.3.1 Proof of Laplace's Limit Theorem

Characteristic functions allow us to show why the *discrete* binomial distribution at large N can be approximated by the *continuous* normal distribution, as claimed in Sect. 5.4. One starts with a sequence of binomially distributed random variables $\{X_N\}$ $(N = 1, 2, 3, \ldots)$ with the probability functions

$$P(X_N = n) = \binom{N}{n} p^n q^{N-n}, \qquad n = 0, 1, 2, \ldots, N, \qquad N = 1, 2, 3, \ldots$$

By (B.23) each such distribution possesses the characteristic function

$$\phi(t; N) = \left(p\mathrm{e}^{\mathrm{i}\,t} + q\right)^N.$$

We introduce standardized random variables

$$Y_N = \frac{X_N - E[X_N]}{\sqrt{\mathrm{var}[X_N]}} = \frac{X_N - Np}{\sqrt{Npq}}$$

and denote the characteristic function of each of them by $\widetilde{\phi}(t; N)$. By (B.21) we get

$$\widetilde{\phi}(t; N) = \left(p\,\mathrm{e}^{\mathrm{i}\,qt/\sqrt{Npq}} + q\,\mathrm{e}^{-\mathrm{i}\,pt/\sqrt{Npq}}\right)^N. \tag{B.24}$$

The terms in the brackets can be expanded in a power series:

$$p\,\mathrm{e}^{\mathrm{i}\,qt/\sqrt{Npq}} \approx p + \mathrm{i}\,t\sqrt{\frac{pq}{N}} - \frac{qt^2}{2N} + \mathcal{O}\left(t^2/N\right),$$
$$q\,\mathrm{e}^{-\mathrm{i}\,pt/\sqrt{Npq}} \approx q - \mathrm{i}\,t\sqrt{\frac{pq}{N}} - \frac{pt^2}{2N} + \mathcal{O}\left(t^2/N\right).$$

Here for each t one has $\lim_{N\to\infty} N\,\mathcal{O}\left(t^2/N\right) = 0$. When this is inserted in (B.24), we get

$$\widetilde{\phi}(t; N) = \left(1 - \frac{t^2}{2N} + \mathcal{O}\left(\frac{t^2}{N}\right)\right)^N \sim \mathrm{e}^{-t^2/2}, \quad \text{when } N \to \infty. \tag{B.25}$$

The limit of the sequence of characteristic functions $\widetilde{\phi}(t; N)$ is thus a continuous function, which is just the characteristic function of the standardized normal distribution. The aid to the final result is the theorem (given without proof): "*If the sequence of characteristic functions $\{\phi_n(t)\}$ at any real t converges to the function $\phi(t)$ and if ϕ is continuous on an arbitrary small interval $(-T, T)$, the sequence $\{F_n(x)\}$ of corresponding distribution functions converges to the distribution function $F(x)$, whose characteristic function is precisely $\phi(t)$.*" This means that for any x one has

$$\lim_{N\to\infty} P(Y_N \le x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-u^2/2}\,\mathrm{d}u,$$

so, at large N (and any x) also

$$P(X_N \le x) \approx \frac{1}{\sqrt{2\pi N p q}} \int_{-\infty}^{x} \mathrm{e}^{-(u-Np)^2/(2Npq)}\,\mathrm{d}u.$$

Put in a more practical form: if the experiment outcome A has a probability p ($0 < p < 1$, $q = 1 - p$) of occurring and X is its frequency in N trials of this experiment, then for arbitrary real numbers a and b ($a < b$) it holds that

$$\lim_{N\to\infty} P\left(a \le \frac{X - Np}{\sqrt{Npq}} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_{a}^{b} \mathrm{e}^{-x^2/2}\,\mathrm{d}x.$$

Now we understand why, at large N, the binomial distribution could be approximated by the normal distribution with the same mean and variance as possessed by the given binomial distribution. This realization is known as the *Laplace's limit theorem* (in its integral form).

By the same token the general central limit theorem can be derived that applies to any probability distribution, as long as its first and second moments exist. The tool is always the same: we power-expand the characteristic function and analyze its behaviour in the $N \to \infty$ limit, which always has the form (B.25).

B.3.2 Inverting the Characteristic Function and Uniqueness of the Density

The characteristic function—as well as its closest relative, the moment-generating function—uniquely determine the probability distribution. In other words, the probability distribution and the characteristic functions offer equivalent description of statistical properties of a random variable. Both worlds are linked by the Fourier transformation: the inverse of (B.18) is

$$f_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_X(t)\,\mathrm{e}^{-\mathrm{i}tn}\,\mathrm{d}t \qquad \text{(discrete case)},$$

while the inverse of (B.19) is

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_X(t)\,\mathrm{e}^{-\mathrm{i}tx}\,\mathrm{d}t \qquad \text{(continuous case)}.$$

But one must realize that the distribution is *not* necessarily uniquely determined if all its moments are known. A well-known case [7] are the probability densities

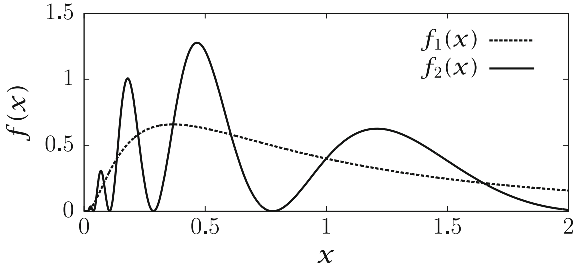


Fig. B.3 Example of different probability densities with identical moments. The function f_1 is the probability density of the *log-normal distribution:* is Y is a normally distributed continuous random variable and $X = \mathrm{e}^Y$, then X is log-normally distributed

$$\begin{aligned} f_1(x) &= \frac{1}{\sqrt{2\pi x^2}}\,\mathrm{e}^{-(\log^2 x)/2}, \qquad & x \ge 0, \\ f_2(x) &= f_1(x)\big[1 + \sin(2\pi \log x)\big], \qquad & x \ge 0, \end{aligned} \tag{B.26}$$

which have very different functional dependencies (see Fig. B.3) yet identical moments, namely

$$E\big[X\big] = \sqrt{\mathrm{e}}, \quad E\big[X^2\big] = \mathrm{e}^2, \quad E\big[X^3\big] = \mathrm{e}^{9/2}, \ldots, E[X^n] = \mathrm{e}^{n^2/2}.$$

This is the so-called *indeterminate moment problem*, briefly outlined below.

If the variables X and Y have identical moments, their characteristic functions $\phi_X(t)$ and $\phi_Y(t)$ have identical expansions near the origin of the real axis. But having equal expansions does not say much about the equality of ϕ_X and ϕ_Y, as the expansion may not converge at all—the terms are calculable in principle, but they can not be summed: in the case just mentioned the convergence of the Taylor series of the characteristic function corresponding to the log-normal density (B.26),

$$\phi_X(t) = \sum_{n=0}^{\infty} a_n (\mathrm{i}\,t)^n, \qquad a_n = \frac{1}{n!}\,E\big[X^n\big],$$

is zero:

$$\rho = \left(\limsup_{n\to\infty} |a_n|^{1/n}\right)^{-1} = 0.$$

However, in specific cases the convergence *is* guaranteed (Theorem 9.6.2 in [8]): if one can find $\rho > 0$ such that near the origin, $|t| < \rho$, the expected value of $\mathrm{e}^{t|X|}$ is finite, i. e. $E[\mathrm{e}^{t|X|}] < \infty$, then $\phi_X(t)$ is absolutely convergent for $|t| < \rho$. Then one may conclude $E[\mathrm{e}^{\mathrm{i}tX}] = E[\mathrm{e}^{\mathrm{i}tY}] \Leftrightarrow X \sim Y$ or

$$\phi_X(t) = \phi_Y(t), \quad |t| < \rho \quad \Leftrightarrow \quad X \sim Y,$$

and under these conditions the equality of the moments of X and Y implies the equality of their probability distributions.

References

1. S. Širca, M. Horvat, *Computational Methods for Physicists* (Springer, Berlin, 2012)
2. J.K. Cavers, On the fast Fourier transform inversion of probability generating functions. J. Inst. Math. Appl. **22**, 275 (1978)
3. J. Abate, W. Whitt, Numerical inversion of probability generating functions. Oper. Res. Lett. **12**, 245 (1992)
4. J. Abate, W. Whitt, The Fourier-series method for inverting transforms of probability distributions. Queueing Syst. **10**, 5 (1992)
5. S. Kullback, An application of characteristic functions to the distribution problem of statistics. Ann. Math. Stat. **5**, 263 (1934)
6. E. Lukacs, Applications of characteristic functions in statistics. Indian J. Stat. A **25**, 175 (1963)
7. C. Berg, Indeterminate moment problem and the theory of entire functions. J. Comput. Appl. Math. **65**, 27 (1995).
8. T. Kawata, *Fourier Analysis in Probability Theory* (Academic Press, New York, 1972)

Appendix C
Random Number Generators

Abstract Methods of generating almost random numbers by means of computer algorithms are presented, starting from integer-based linear and non-linear congruential generators of uniformly distributed random numbers. They are followed by a discussion of methods to draw random numbers from arbitrary continuous distribution, and a brief mention of the ways to generate truly random numbers.

Statistical methods and numerical procedures often require us to use random samples or some kind of "source" of numbers that are as random as possible, that is, *pseudo-random*. The computer namely can not do anything "by chance", so in order to generate pseudo-random numbers we rely on *deterministic* processes of computing particular sequences that are only *seemingly* random [1]. Generating pseudo-random numbers—labeled simply 'random' in the following—is called *drawing*.

C.1 Uniformly Distributed Pseudo-Random Numbers

In order to generate uniformly distributed pseudo-random numbers one uses *uniform generators*. They are supposed to deliver uniform numbers $X \sim U(0, 1)$, distributed according to (3.1).

The sequences $\{x_i\}$ generated by a good uniform generator are expected to be *uncorrelated:* this means that the vectors of sub-sequences $(x_i, x_{i+1}, \dots, x_{i+k})$ are as weakly correlated as possible, for each k separately. One also wishes for the sequence to possess a *long period:* it should not repeat itself too quickly. Besides, one would like the sequence $\{x_i\}$ to be *uniform and unbiased*, meaning that the same number of generated points fall in the same volume of space. An important request is a good uniformity of the distribution of the points $(x_i, x_{i+1}, \dots, x_{i+k-1})$ in a k-dimensional hypercube, with k as large as possible: this is known as the *serial uniformity of the sequence.*

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6

Most uniform generators are devised in integer arithmetic. Such generators return numbers with equal probabilities on the interval $[0, m-1]$, where $m = 2^{32}$ or 2^{64}. Uniform generators are standard components of general libraries and tools, e.g. `rand()` in MATLAB and C/C++, `gsl_rng_rand` in GSL or `Random[]` in MATHEMATICA. Random integers $x_i \in \mathbb{Z}_m$ generated by an integer generator can be converted to uniformly distributed real numbers $\xi_i \sim U(0,1)$ by using the transformations

$$\begin{aligned}
\xi_i &= x_i/m && \text{approximately uniform in } [0,1), \\
\xi_i &= x_i/(m-1) && [0,1], \\
\xi_i &= (x_i+1)/m && (0,1], \\
\xi_i &= (x_i+1/2)/m && (0,1).
\end{aligned}$$

If one uses floating-point arithmetic (precision 2^{-n}, mantissa length n), the numbers generated in this way have $b = \log_2 m$ random most significant bits, which is often not enough, and certainly less than n. An approximation of a real number ξ on the interval $[0,1)$ with all bits random is obtained by independently drawing integers $\{x_i \in \mathbb{Z}_m\}_{i=1}^{h}$ and using the formula $\xi = x_1 m^{-1} + x_2 m^{-2} + \cdots + x_h m^{-h}$, where $(h-1)b < n < (h+1)b$.

C.1.1 Linear Congruential Generators

Classical random generators are based on the relation of congruence.[2] Congruential generators of numbers $x_i \in \mathbb{Z}_m = [0, m-1]$, where $i \in \mathbb{N}_0 = \{0, 1, \ldots\}$, are defined by the *transition function* ϕ and the relation

$$x_{i+1} \equiv \phi(x_i, x_{i-1}, \ldots, x_{i-k+1}) \mod m,$$

where k is the generator *order*. Thus ϕ is restricted to $\mathbb{Z}_m$ by the congruence relation modulo m. The initial state of the generator $\{x_0, x_1, \ldots, x_{k-1}\}$ is a unique function of the number called the *seed* by which the sequence is completely determined: a generator initialized by the same seed always delivers the same sequence of numbers. If ϕ is a linear function of parameters, one refers to *linear* generators, otherwise they are *non-linear*.

The simplest *linear congruential generator* (LCG) has the form

$$x_{i+1} \equiv (a x_i + c) \mod m, \tag{C.1}$$

[2] One has $\{x \equiv x \mod m\}$; the congruence relation is commutative, $\{x \equiv y \mod m\} \Leftrightarrow \{y \equiv x \mod m\}$, and transitive, $\{x \equiv y \mod m\} \wedge \{y \equiv z \mod m\} \Rightarrow \{x \equiv z \mod m\}$.

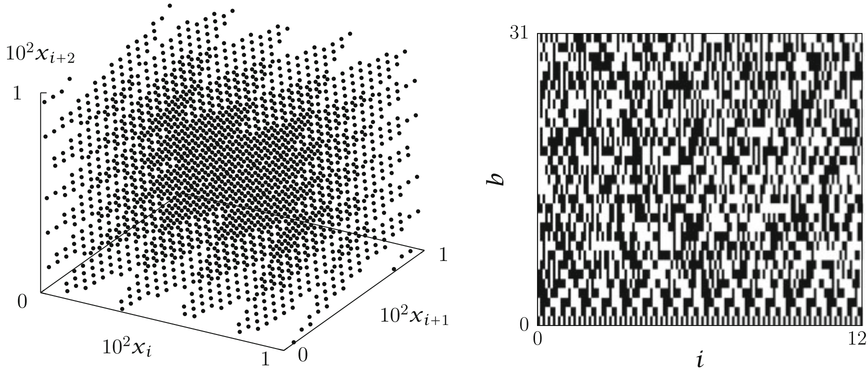


Fig. C.1 [Left] Zoom-in of the phase space $[0,1]^3$ of points $2^{-31}(x_i, x_{i+1}, x_{i+2})$ of the sequence x_i obtained by the standard random generator from the `glibc` library with $x_0 = 12345$. [Right] The bits b of the random numbers x_i (*black* = 1, *white* = 0)

where a is the *multiplier* and c is the *carry* parameter, while x_0 is the seed. Since x_i are determined by x_{i-1} and can take only m different values, the period of a LCG is at most m for $c \neq 0$ and at most $m - 1$ for $c = 0$.

Example Take a LCG with $m = 31$, $a = 3$, $c = 0$ and $x_0 = 9$. Run the recurrence (C.1) a hundred times: we get $\{x_1, x_2, x_3, \ldots, x_{30}\} = \{27, 19, 26, \ldots, 9\}$, then again $\{x_{31}, x_{32}, x_{33}, \ldots, x_{60}\} = \{27, 19, 26, \ldots, 9\}$, and so on. The period of the generator is therefore only 30, but this is not the only problem. If subsequent pairs $(x_i, x_{i+1}) = (9, 27), (27, 19), (19, 26), \ldots$ are plotted on a graph—do it!—we realize that all points lie on straight lines with slope 3. That certainly does not appear to be random!

Is one better off by increasing m and a, and changing c? Take, for instance, $m = 2^{32}$, $a = 1103515245$ and $c = 123454$: this corresponds to the default generator in the 32-bit `glibc` library. Figure C.1 (left) shows the distribution of subsequent triplets (x_i, x_{i+1}, x_{i+2}). Obviously the points are arranged in planes and this deficiency of LCG persists at larger k as well: in general the points $m^{-1}(x_i, \ldots, x_{i+k-1})$ do not fill the entire k-dimensional hypercube, but rather lie on at most $(mk!)^{1/k}$ hyperplanes. Besides, the least significant bits are less random that the rest (Fig. C.1 (right)). A good generator ought to produce points on many hyperplanes and make all their bits random. In applications where such deficiencies are irrelevant, LCG-type generators are nevertheless put to good use, as they are supported by all programming languages, simple and fast. ◁

Further representatives of the LCG family are the generators of the Add-with-Carry (AWC), Subtract-with-Borrow (SWB) and Multiply-with-Carry (MWC) type:

$$\begin{aligned}
\text{AWC}: x_i &\equiv (x_{i-r} + x_{i-k} + c_{i-1}) &&\bmod m, \; c_i = \lfloor (x_{i-r} + x_{i-k} + c_{i-1})/m \rfloor, \\
\text{SWB}: x_i &\equiv (x_{i-r} - x_{i-k} - c_{i-1}) &&\bmod m, \; c_i = \lfloor (x_{i-r} - x_{i-k} - c_{i-1})/m \rfloor, \\
\text{MWC}: x_i &\equiv (a x_{i-r} + c_{i-1}) &&\bmod m, \; c_i = \lfloor (a x_{i-r} + c_{i-1})/m \rfloor.
\end{aligned}$$

The SWB algorithm is the basis of the RANLUX generator from the GSL library. *Multiple recursive generators* (MRG) are also in wide-spread use:

$$x_i \equiv (a_1 x_{i-1} + \cdots + a_k x_{i-k} + c_i) \mod m, \tag{C.2}$$

where $a_k \in \mathbb{Z}_m$ are constants. The MR generators usually exhibit much larger periods than simple LC generators. If m is a prime, the maximal period may be as high as $m^k - 1$. An example of such a generator of the fifth order is

$$x_i \equiv \big(107374182\, x_{i-1} + 104480\, x_{i-5}\big) \mod \big(2^{31} - 1\big).$$

C.1.2 Non-linear Congruential Generators

In general, non-linear generators are more random than linear ones, but they are also slower. Their main representatives are the *inversive congruential generators* (ICG) defined by the recurrence

$$x_i \equiv (a\overline{x}_{i-1} + b) \mod m,$$

where $1 \equiv (\overline{x}x) \mod m$, and the *explicit inversive congruential generators* (EICG) based on the relation

$$x_i \equiv \overline{a(i + i_0) + b} \mod m.$$

For prime modules m the generators of IC and EIC types generate points that avoid accumulation in planes, a behavior so typical of the LC generators, yet modular inversion is a numerically intensive procedure, while the filling of space tends to be slightly less uniform.

C.1.3 Generators Based on Bit Shifts

A completely different approach to generating random numbers is offered by *feedback shift register* generators. If the numbers x_i are written as n-plets of bits, the relation (C.2) can be written as

$$b_i \equiv (a_p b_{i-p} + a_{p-1} b_{i-p+1} + \cdots + a_1 b_{i-1}) \mod 2, \tag{C.3}$$

where all variables can take only values 0 or 1. It turns out that the recurrence (C.3) can be performed by shifting bits: an example is shown in Fig. C.2.

The evicted bit is then combined by the pattern of bits on its right by using a variety of logical operations. The recurrence (C.3) often has the form $b_i \equiv (b_{i-p} + b_{i-p+q}) \mod 2$ or $b_i \equiv b_{i-p} \oplus b_{i-p+q}$, where $\oplus$ is the exclusive "or" (adding 0 and 1 modulo

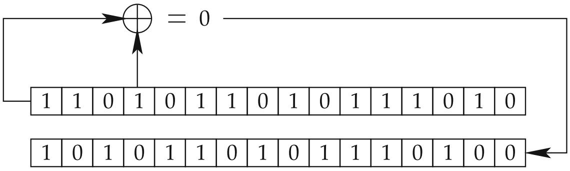


Fig. C.2 Example of bit shifts in a FSR-type generator. The pushed-out bit 1 at the extreme left and the bit 1 deeper in the register are combined by an exclusive "or" (XOR). The result 0 replaces the missing bit at the extreme right

2). For n-tuples x_i this means $x_i = x_{i-p} \oplus x_{i-p+q}$, where the operation $\oplus$ is performed bit-wise. This game can be continued: if x_i are interpreted as n-dimensional vectors, they can be multiplied by $n \times n$ matrices:

$$x_i \equiv x_{i-p} \oplus A x_{i-p+q} .$$

Where is all this heading? The matrix A can be used to *twist* the bit n-tuples prior to being logically combined, thereby increasing the randomness of the generated x_i. Such "kneading" of bit samples is at the heart of the *Mersenne twister* generator [2] (algorithm MT19937), which we recommend for serious applications. It is implemented in 32-bit integer arithmetic, has been theoretically well explored and is accessible in standard packages and libraries. Its period is $2^{19937} - 1$ and is serially uniform for dimensions $k \in [1, 623]$. Its weakness is a somewhat lower randomness of subsequent bits between consecutive generated numbers.

C.1.4 Some Hints for Use of Random Generators

Any random number generator, no matter how sophisticated, has some deficiency, which is usually very specific. If we, as non-specialists, need a generator to be invoked many times in our code, we may consider the following guidelines.

Only choose a generator devised and tested by experts. The code should be as terse as possible and based on integer arithmetic in favor of greater speed. Use generators with long periods and high serial uniformity in as many dimensions as possible. If the generator is accessible in source code, incorporate it into the program, as modern compilers can link the code segments in the form of `inline` functions. Before using a generator, study its statistical properties and ascertain whether its deficiencies may jeopardize the correctness of the result. Perform each calculation by using different generators and different seeds.

C.2 Drawing from Arbitrary Continuous Distributions

A generator of random numbers with arbitrary distribution is obtained by transforming the numbers returned by a uniform generator. An exhaustive overview of the area is offered by the classical monograph [3]; here we present some cases of transformations to continuous distributions most commonly encountered in physics. For transformation to discrete distributions see Sect. C.2 in [4].

C.2.1 Uniform Distribution Within a Circle or Sphere

How do we draw points that are homogeneously distributed within a circle? Homogeneity means: the ratio of a tiny probability $\mathrm{d}P$ that the drawn point falls in a tiny surface element, to its area, $\mathrm{d}S = r\,\mathrm{d}r\,\mathrm{d}\phi$, is equal to the ratio of probability 1 that the point falls in the whole circle, to its area, πR^2:

$$\frac{\mathrm{d}P}{\mathrm{d}S} = \frac{\mathrm{d}P}{r\,\mathrm{d}r\,\mathrm{d}\phi} = \frac{1}{\pi R^2} \quad \overset{R=1}{\Longrightarrow} \quad \frac{\mathrm{d}P}{\mathrm{d}\bigl(r^2\bigr)\mathrm{d}\bigl(\phi/2\pi\bigr)} = 1.$$

Therefore we must draw *uniformly* in r^2 from 0 to 1 (not r from 0 to 1!) and in ϕ from 0 to 2π. We need two random numbers $U_1, U_2 \sim U[0, 1)$ and compute

$$\bigl(r_i, \phi_i\bigr) = \bigl(R\sqrt{U_1},\, 2\pi U_2\bigr).$$

In the three-dimensional case the circle area $S = \pi R^2$ needs to be replaced by the sphere volume $V = 4\pi R^3/3$, and the area element $\mathrm{d}S$ by the volume element $\mathrm{d}V = r^2\,\mathrm{d}r\,\mathrm{d}(\cos\theta)\,\mathrm{d}\phi$. Thus

$$\frac{\mathrm{d}P}{\mathrm{d}V} = \frac{\mathrm{d}P}{r^2\,\mathrm{d}r\,\mathrm{d}(\cos\theta)\,\mathrm{d}\phi} = \frac{3}{4\pi R^3} \quad \overset{R=1}{\Longrightarrow} \quad \frac{\mathrm{d}P}{\mathrm{d}\bigl(r^3\bigr)\mathrm{d}\bigl(\frac{1}{2}\cos\theta\bigr)\mathrm{d}\bigl(\phi/2\pi\bigr)} = 1.$$

Hence we must draw uniformly in r^3 from 0 to 1, uniformly in $\cos\theta$ from -1 to 1 (not θ from 0 to π!) and uniformly in ϕ from 0 to 2π. The three numbers $U_1, U_2, U_3 \sim U[0, 1)$ drawn according to these distributions define the point

$$\bigl(r_i,\, \theta_i,\, \phi_i\bigr) = \bigl(R\sqrt[3]{U_1},\, \arccos(2U_2 - 1),\, 2\pi U_3\bigr). \tag{C.4}$$

C.2.2 Uniform Distribution with Respect to Directions in $\mathbb{R}^3$ and $\mathbb{R}^d$

A uniform distribution over *directions* in space (usually $\mathbb{R}^3$) is called isotropic. Isotropy means that the ratio between the number of points $\mathrm{d}N$ on the small surface $\mathrm{d}S$ on the unit sphere to an infinitesimal solid angle $\mathrm{d}\Omega$, is equal to the ratio of the number of points N on the whole surface to the full solid angle $\Omega = 4\pi$. A frequent beginner's mistake is to uniformly draw the angles θ and ϕ according to $U(0,\pi)$ and $U(0,2\pi)$, respectively, and compute $(x, y, z) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. But this generates points that prefer to accumulate near the poles, as shown in Fig. C.3 (left). The correct way to draw is by recipe (C.4), where the radial coordinate is simply ignored. This results in a homogeneous surface distribution, as shown in Fig. C.3 (right).

The points $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$, uniformly distributed over the $(d-1)$-dimensional sphere $\boldsymbol{S}_{d-1} \in \mathbb{R}^d$, can be generated by independently drawing the components of the vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_d)^{\mathrm{T}}$ with probability density $N(0,1)$ and normalizing it: $x_i = y_i/\|\boldsymbol{y}\|_2$, where $\|\boldsymbol{y}\|_2 = \left(\sum_{i=1}^d y_i^2\right)^{1/2}$.

C.2.3 Uniform Distribution Over a Hyperplane

The points $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^{\mathrm{T}}$, $x_i > 0$, uniformly distributed over a hyperplane defined by the equation $\sum_{i=1}^d a_i x_i = b$ ($a_i > 0$, $b > 0$), are generated by independently drawing d components of the vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_d)^{\mathrm{T}}$ with exponential density $f(y) = \exp(-y)$ (see Table C.1) and calculating [5]

$$S = \sum_{i=1}^{d} a_i y_i, \qquad x_i = \frac{b}{S}\, a_i y_i.$$

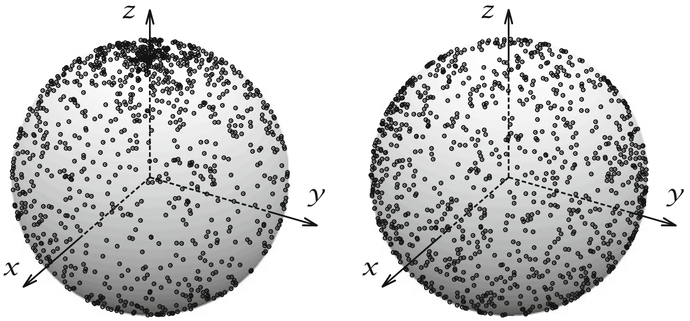


Fig. C.3 Generating an isotropic distribution in $\mathbb{R}^3$. [Left] Incorrect drawing by using $\theta_i = \pi\xi$, $\xi \sim U[0,1)$. [Right] Correct drawing by using $\theta_i = \arccos(2\xi - 1)$

C.2.4 Transformation (Inverse) Method

Our knowledge of variable transformations from Sects. 2.7 and 2.10 can be used to generate random numbers according to an arbitrary continuous distribution. We know how uniform numbers $Y \sim U(0,1)$ can be generated; but as for arbitrary probability densities f_X and f_Y one has $|f_X(x)\,\mathrm{d}x| = |f_Y(y)\,\mathrm{d}y|$, this means that

$$f_X(x) = \frac{\mathrm{d}y}{\mathrm{d}x},$$

since $f_Y(y) = 1$. The solution of this equation is $y = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = F_X(x)$, where F_X is the distribution function of X. In other words,

$$x = F_X^{-1}(y), \qquad Y \sim U(0,1),$$

where F_X^{-1} is the *inverse* function of F_X (not its reciprocal value). Clearly we have obtained a tool to generate random variables distributed according to F_X (see Fig. C.4 (left)).

The transformation method is useful if the inverse F_X^{-1} is relatively easy to compute. The collection of such functions is quickly exhausted; some common examples are listed in Table C.1.

Example Let us construct a generator of dipole electro-magnetic radiation! The distribution of radiated power with respect to the solid angle is $\mathrm{d}P/\mathrm{d}\Omega \propto \sin^2\theta$,

$$f_\Theta(\theta) = \frac{\mathrm{d}P}{\mathrm{d}\theta} = \frac{3}{4}\sin^3\theta, \qquad 0 \le \theta \le \pi,$$

where the normalization constant has been determined by $C\int_0^\pi \sin^3\theta\,\mathrm{d}\theta = 1$. (The radiation is uniform in ϕ.) The corresponding distribution function is

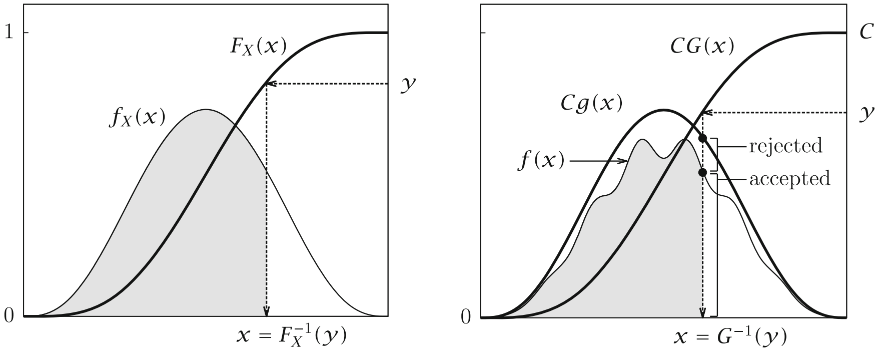


Fig. C.4 Generating random numbers according to arbitrary continuous distributions. [Left] Transformation (inverse of distribution function) method. [Right] Rejection method

Table C.1 Generating random numbers according to chosen probability distributions by the transformation method

Distribution	$f_X(x)$	$F_X(x)$	$X = F_X^{-1}(U)$
Exponential $(x \geq 0)$	$\lambda e^{-\lambda x}$	$1 - e^{-\lambda x}$	$-\frac{1}{\lambda} \log U$
Normal $(-\infty < x < \infty)$	$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	$\frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$	$\sqrt{2}\text{erf}^{-1}(2U - 1)$
Cauchy $(-\infty < x < \infty)$	$\frac{a}{\pi(a^2 + x^2)}$	$\frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x}{a}\right)$	$a \tan \pi U$
Pareto $(0 < b \leq x)$	$\frac{ab^a}{x^{a+1}}$	$1 - \left(\frac{b}{x}\right)^a$	$\frac{b}{U^{1/a}}$
Triangular on $[0, a]$ $(0 \leq x \leq a)$	$\frac{2}{a}\left(1 - \frac{x}{a}\right)$	$\frac{2}{a}\left(x - \frac{x^2}{2a}\right)$	$a\left(1 - \sqrt{U}\right)$
Rayleigh $(x \geq 0)$	$\frac{x}{\sigma} e^{-x^2/(2\sigma^2)}$	$1 - e^{-x^2/(2\sigma^2)}$	$\sigma\sqrt{-\log U}$

Note that drawing Y by the uniform distribution $U(0, 1)$ is equivalent to drawing by $1 - U(0, 1)$. For the normal distribution see also Sect. C.2.5

$$F_\Theta(\theta) = \int_0^\theta f_\Theta(\theta')\, d\theta' = \frac{3}{4}\left[\frac{\cos^3\theta}{3} - \cos\theta + \frac{2}{3}\right].$$

The desired distribution in θ is obtained by drawing the values x according to $U(0, 1)$ and calculating $\theta = F_\Theta^{-1}(x)$. The inverse of F_Θ is annoying but can be done. By substituting $t = \cos\theta$ the problem amounts to solving the cubic equation $t^3 - 3t + 2 = 4x$, for which explicit formulas exist. Alternatively, one can seek the solution of the equation $\mathcal{F}(\theta) = F_\Theta(\theta) - x = 0$. ◁

C.2.5 Normally Distributed Random Numbers

If U_1 and U_2 are independent random variables, distributed as $U_1 \sim U(0, 1]$ and $U_2 \sim [0, 1)$, their Box-Muller transformation [6]

$$X_1 = \sqrt{-2\log U_1}\, \cos(2\pi U_2), \qquad X_2 = \sqrt{-2\log U_1}\, \sin(2\pi U_2),$$

yields independent random variables X_1 and X_2, distributed according to the standard normal distribution $N(0, 1)$. The variables U_1 and U_2 define the length $R = \sqrt{-2\log U_1}$ and the directional angle $\theta = 2\pi U_2$ of a planar vector $(X_1, X_2)^T$. The numerically intensive calculation of trigonometric functions can be avoided by using Marsaglia's implementation (see [7], Chap. 7, Algorithm P):

```
repeat
    Independently draw $u_1$ by $U(0,1]$ and $u_2$ by $U[0,1)$;
    $\boldsymbol{v} = 2(u_1, u_2)^{\mathrm{T}} - (1,1)^{\mathrm{T}}$;
    $s = |\boldsymbol{v}|^2$;
until $(s \geq 1 \vee s \neq 0)$;
$(x_1, x_2)^{\mathrm{T}} = \sqrt{-(2/s)\log s}\,\boldsymbol{v}$;
```

The drawn vector $\boldsymbol{v}$ on average uniformly covers the unit circle, while approximately $1-\pi/4 \approx 21.5\%$ generated points are rejected, so that for one pair (x_1, x_2) one needs to draw $2/(\pi/4) \approx 2.54$ uniform numbers.

Values of the random vector $\boldsymbol{X} \in \mathbb{R}^d$, distributed according to the multivariate probability density (4.23) with mean $\boldsymbol{\mu}$ and correlation matrix Σ are generated by independently drawing d components of the vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_d)^{\mathrm{T}}$ by the standardized normal distribution $N(0,1)$ and computing

$$\boldsymbol{x} = L\boldsymbol{y} + \boldsymbol{\mu},$$

where L is the lower-triangular $d \times d$ matrix from the Cholesky decomposition of the correlation matrix, $\Sigma = LL^{\mathrm{T}}$.

C.2.6 Rejection Method

Suppose we wish to draw random numbers according to some complicated density f, while some very efficient way is at hand to generate the numbers according to another, simpler density g. We first try to find $C > 1$ such that f is bounded by Cg from above as tightly as possible (Fig. C.4 (right)), that is, to ensure $f(x) < Cg(x)$ for all x with C as close to 1 as possible. Then the random numbers Y distributed according to f can be generated by the procedure:

1. Generate the value x of random variable X according to density g.
2. Generate the value u of random variable U according to $U(0,1)$.
3. If $u \leq f(x)/(Cg(x))$, assign $y = x$ (x is "accepted"), otherwise return to step 1 (x is "rejected").

Does this recipe really do what it is supposed to do? Let us define the event $B = \big\{U \leq f(X)/\big(Cg(X)\big)\big\}$. From the given recipe and the Figure it is clear that

$$P\big(B \,|\, X = x\big) = P\left(U \leq \frac{f(X)}{Cg(X)}\,\middle|\, X = x\right) = \frac{f(x)}{Cg(x)},$$

hence

$$P(B) = \int_{-\infty}^{\infty} P\big(B \,|\, X = x\big)\, g(x)\, \mathrm{d}x = \int_{-\infty}^{\infty} \frac{f(x)}{Cg(x)}\, g(x)\, \mathrm{d}x = \frac{1}{C}\int_{-\infty}^{\infty} f(x)\, \mathrm{d}x = \frac{1}{C}.$$

Now define the event $A = \{X \le x\}$. We must prove that the conditional distribution function for X, given condition B, is indeed F, that is, we must check

$$P(A|B) = P\left(X \le x \,\middle|\, U \le \frac{f(X)}{Cg(X)}\right) \overset{?}{=} F(x).$$

For this purpose we first calculate $P(B|A)$, where we exploit the definition of conditional probability (1.10) in the form $P(B|A) = P(AB)/P(A)$,

$$\begin{aligned} P(B|A) &= P\left(U \le \frac{f(X)}{Cg(X)} \,\middle|\, X \le x\right) = \frac{P\big(U \le f(X)/\big(Cg(X)\big) \cap X \le x\big)}{P(X \le x)} \\ &= \int_{-\infty}^{x} \frac{P\big(U \le f(X)/\big(Cg(X)\big)\big|\, X = z \le x\big)}{P(X \le x)}\, g(z)\,\mathrm{d}z \\ &= \frac{1}{G(x)} \int_{-\infty}^{x} \frac{f(z)}{Cg(z)}\, g(z)\,\mathrm{d}z = \frac{1}{CG(x)} \int_{-\infty}^{x} f(z)\,\mathrm{d}z = \frac{F(x)}{CG(x)}, \end{aligned}$$

and then invoke the product formula (1.10) for the final result

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{F(x)}{CG(x)} \frac{G(x)}{1/C} = F(x).$$

Example For the Cauchy distribution with probability density (3.18) the distribution function and its inverse are easy to calculate:

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x, \qquad F_X^{-1}(t) = \tan\left[\pi\left(t - \frac{1}{2}\right)\right]. \tag{C.5}$$

To generate the values of a Cauchy-distributed variable X one could therefore resort to the transformation method by using in (C.5) a random variable U, uniformly distributed over $[-1/2, 1/2]$—or, due to symmetry, over $[0, 1]$—and calculating $X = \tan \pi U$ (third row of Table C.1). But since computing the tangent is slow, it is better to seek the values of X as the ratios between the projections of the points within the circle onto x and y axes. These points are uniformly distributed with respect to the angles. We use the algorithm

```
repeat
  | Draw $u_1$ according to $U(-1, 1)$ and $u_2$ according to $U(0, 1)$.
until ( $u_1^2 + u_2^2 > 1 \vee u_2 = 0$ );
$x = u_1/u_2$;
```

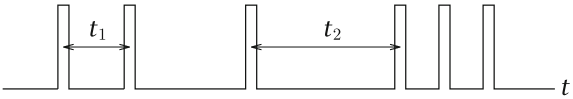
Note that the fraction of rejected points is $1 - \pi/4$ and that the accepted points (u_1, u_2) lie in the upper half of the unit circle. (Check this!) ◁

C.3 Generating *Truly* Random Numbers

If we wish to cast off the burden of the ‘pseudo’ attribute in our discussion and generate truly random numbers, we must also reach for a genuinely random process. An example of such process is the radioactive decay of atomic nuclei, which is exploited by the `HotBits` generator of random bit sequences [8]. The laboratory maintains a sample of radioactive cesium, decaying to an excited state of barium, electron and anti-neutrino with a decay time of 30.17 years:

$$^{137}\mathrm{Cs} \longrightarrow {}^{137}\mathrm{Ba}^* + \mathrm{e}^- + \overline{\nu}_\mathrm{e}.$$

The decay instant is defined by the detected electron. The time of the decay of any nucleus in the source is completely random, so the time difference between subsequent decays is also completely random. The apparatus measures the time differences between two *pairs* of decays, t_1 and t_2, as shown in the figure.



If $t_1 = t_2$ (within instrumental resolution), the measurement is discarded. If $t_1 < t_2$, the value 0 is recorded, and if $t_1 > t_2$, the value 1 is recorded. The sense of comparing t_1 to t_2 is reversed with each subsequent pair in order to avoid systematic errors in the apparatus or in the measurement that could bias one outcome against the other. The final result is a random bit sequence like

```
 1111011100100001101110100010110001001100110110011100111100000001
0100001010011111111001011101111001101001101110000100010110001111 ...
```

The speed of generation depends on the activity of the radioactive source.

Example Imagine a descent along a binary tree (Fig. C.5) where each branch point represents a random step to the left ($n_i = 1$) with probability p or to the right ($n_i = 0$) with probability $1-p$. (The left-right decision can be made, for example, by “asking” the radioactive source discussed above.) The values n_i corresponding to the traversed branches are arranged in a k-digit binary number $B_k = \left(n_{k-1}n_{k-2}\ldots n_1 n_0\right)_2$ and suitably normalized,

$$X_k = N_k B_k = N_k \sum_{i=0}^{k-1} 2^i n_i, \qquad N_k = \left(2^k - 1\right)^{-1},$$

so that we ultimately end up with $0 \le X_k < 1$. What is the expected value of X_k in the decimal system (base 10)? The individual digits n_i take the values 0 or 1 with probabilities $P_i = p\delta_{i,1} + (1-p)\delta_{i,0}$. Obviously $E[n_i] = p$, hence

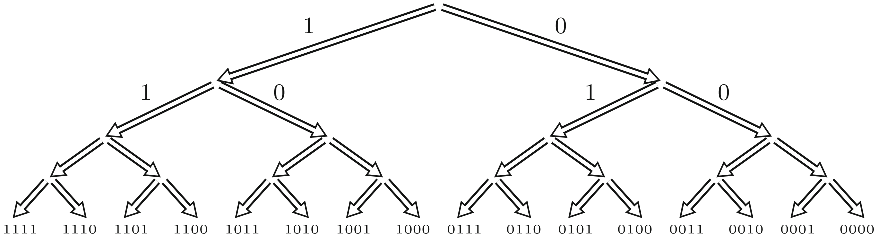


Fig. C.5 Binary tree used to generate a random k-digit binary number

$$E\big[X_k\big] = E\left[N_k \sum_{i=0}^{k-1} 2^i n_i\right] = N_k E[n_i] \sum_{i=0}^{k-1} 2^i = N_k p \left(2^k - 1\right) = p.$$

The variance of X_k is

$$\begin{aligned}\mathrm{var}[X_k] &= E\big[X_k^2\big] - E\big[X_k\big]^2 = N_k^2 \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} 2^{i+j} \left(\underbrace{E[n_i n_j]}_{p\delta_{i,j}} - \underbrace{E[n_i]E[n_j]}_{p^2\delta_{i,j}} \right) \\ &= N_k^2 p(1-p) \sum_{i=0}^{k-1} 4^i = N_k^2 p(1-p)\,\frac{4^k-1}{3} = \frac{p(1-p)}{3}\,\frac{2^k+1}{2^k-1}.\end{aligned}$$

We have thus devised a generator of *truly random* numbers, distributed according to $U[0,1)$. In particular, for $p = 1/2$ one indeed has $E[X_k] = 1/2$, while $\lim_{k\to\infty} \mathrm{var}\big[X_k\big] = 1/12$, as expected of a uniform distribution. ◁

References

1. P. L'Ecuyer, Uniform random number generators, in *Non-uniform random variate generation*, *International Encyclopedia of Statistical Science*, ed. by M. Lovric (Springer, Berlin, 2011)
2. M. Matsumoto, T. Nishimura, Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans. Model. Comput. Simul. **8**, 3 (1998)
3. L. Devroye, *Non-uniform Random Variate Generation* (Springer, Berlin, 1986)
4. S. Širca, M. Horvat, *Computational Methods for Physicists* (Springer, Berlin, 2012)
5. M. Horvat, The ensemble of random Markov matrices. J. Stat. Mech. **2009**, P07005 (2009)
6. G.E.P. Box, M.E. Muller, A note on the generation of random normal deviates. Ann. Math. Stat. **29**, 610 (1958)
7. D. Knuth, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd edn. (Addison-Wesley Professional, Reading 1998)
8. J. Walker, Hotbits; see http://www.fourmilab.ch/hotbits

Appendix D
Tables of Distribution Quantiles

Abstract Definite integrals of the normal distribution are given in tabular form, along with the most frequently used quantiles of the χ^2, t and F distributions.

Definite integrals of some distributions have awkward analytic expressions, so one may prefer to read them off from tables. Table D.1 lists the integrals of the standardized normal distribution (Fig. D.1 (top left)), Table D.2 contains the values of the erf function, and Table D.3 has the quantiles χ^2_p of the χ^2 distribution with ν degrees of freedom (Fig. D.1 (top right)). Table D.4 lists the quantiles t_p of the Student's t distribution with ν degrees of freedom (Fig. D.1 (bottom left)). Tables D.5 and D.6 contain the 95. percentile ($F_{0.95}$) and 99. percentile ($F_{0.99}$), respectively, of the F distribution with ν_1 and ν_2 degrees of freedom in the numerator and denominator, respectively (Fig. D.1 (bottom right)).

Note that the integral of the standardized normal distribution (Table D.1) and the value of the erf function (Table D.2) are related by

$$\frac{1}{\sqrt{2\pi}}\int_0^z \mathrm{e}^{-t^2/2}\mathrm{d}t = \frac{1}{2}\,\mathrm{erf}\left(\frac{z}{\sqrt{2}}\right).$$

The distribution function of the standardized normal distribution is

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^z \mathrm{e}^{-t^2/2}\,\mathrm{d}t = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_0^z \mathrm{e}^{-t^2/2}\,\mathrm{d}t = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right].$$

D.1 Calculating Quantiles with Mathematica

Arbitrary quantiles not given in the following tables can be calculated by interpolation or by resorting to a general tool like Mathematica [1]. For example, to obtain the 90. percentile of the χ^2 distribution with $\nu = 5$ degrees of freedom, the

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6

Table D.1 Integral of the standardized normal distribution (3.10) and (3.12) from 0 to z in steps of 0.01

z	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

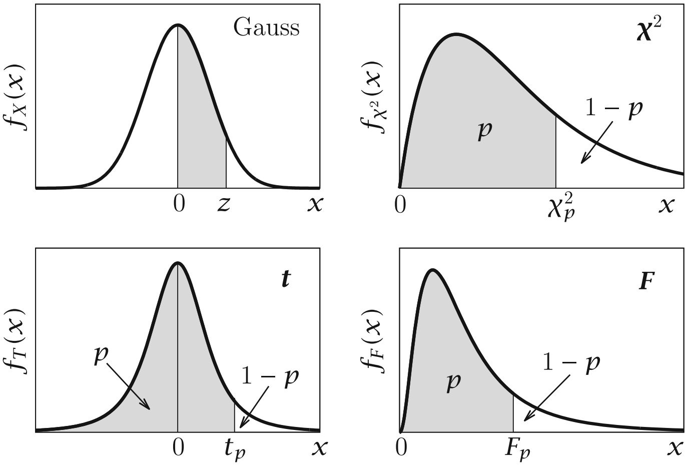


Fig. D.1 [Top left] Definite integral of the standard normal distribution (3.10) and (3.12) from 0 to z. [Top right] Definition of the pth quantile of the χ^2 distribution (3.21). [Bottom left] Definition of the pth quantile of the t distribution (3.22). [Bottom right] Definition of the pth quantile of the F distribution (3.23)

0.995th quantile of the Student's t distribution with $\nu = 1$ degree of freedom and the 95. percentile of the F distribution for $\nu_1 = \nu_2 = 10$ we issue the commands

```
Quantile[ChiSquareDistribution[5], 0.90],
Quantile[StudentTDistribution[1], 0.995],
Quantile[FRatioDistribution[10,10], 0.95],
```

which give (in the same order as above)

```
9.23636,
63.6567,
2.97824.
```

(Compare these values to entries in the corresponding Tables.) Definite integrals of all mentioned distributions can be obtained by commands of the form

```
NIntegrate[PDF[FRatioDistribution[7,9], x], {x, 0, 3.293}],
NIntegrate[PDF[FRatioDistribution[7,9], x], {x, 0, 3.70}],
NIntegrate[PDF[FRatioDistribution[7,9], x], {x, 0, 5.613}],
NIntegrate[PDF[FRatioDistribution[9,7], x], {x, 0, 1./3.70}].
```

Here we have only demonstrated a sample calculation of integrating the density of the F distribution with parameters required by the Example on p. 187: the four command lines listed above yield the values

Table D.2 Values of the erf function (3.8) from 0 to z in steps of 0.01

z	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0113	0.0226	0.0338	0.0451	0.0564	0.0676	0.0789	0.0901	0.1013
0.1	0.1125	0.1236	0.1348	0.1459	0.1569	0.1680	0.1790	0.1900	0.2009	0.2118
0.2	0.2227	0.2335	0.2443	0.2550	0.2657	0.2763	0.2869	0.2974	0.3079	0.3183
0.3	0.3286	0.3389	0.3491	0.3593	0.3694	0.3794	0.3893	0.3992	0.4090	0.4187
0.4	0.4284	0.4380	0.4475	0.4569	0.4662	0.4755	0.4847	0.4937	0.5027	0.5117
0.5	0.5205	0.5292	0.5379	0.5465	0.5549	0.5633	0.5716	0.5798	0.5879	0.5959
0.6	0.6039	0.6117	0.6194	0.6270	0.6346	0.6420	0.6494	0.6566	0.6638	0.6708
0.7	0.6778	0.6847	0.6914	0.6981	0.7047	0.7112	0.7175	0.7238	0.7300	0.7361
0.8	0.7421	0.7480	0.7538	0.7595	0.7651	0.7707	0.7761	0.7814	0.7867	0.7918
0.9	0.7969	0.8019	0.8068	0.8116	0.8163	0.8209	0.8254	0.8299	0.8342	0.8385
1.0	0.8427	0.8468	0.8508	0.8548	0.8586	0.8624	0.8661	0.8698	0.8733	0.8768
1.1	0.8802	0.8835	0.8868	0.8900	0.8931	0.8961	0.8991	0.9020	0.9048	0.9076
1.2	0.9103	0.9130	0.9155	0.9181	0.9205	0.9229	0.9252	0.9275	0.9297	0.9319
1.3	0.9340	0.9361	0.9381	0.9400	0.9419	0.9438	0.9456	0.9473	0.9490	0.9507
1.4	0.9523	0.9539	0.9554	0.9569	0.9583	0.9597	0.9611	0.9624	0.9637	0.9649
1.5	0.9661	0.9673	0.9684	0.9695	0.9706	0.9716	0.9726	0.9736	0.9745	0.9755
1.6	0.9763	0.9772	0.9780	0.9788	0.9796	0.9804	0.9811	0.9818	0.9825	0.9832
1.7	0.9838	0.9844	0.9850	0.9856	0.9861	0.9867	0.9872	0.9877	0.9882	0.9886
1.8	0.9891	0.9895	0.9899	0.9903	0.9907	0.9911	0.9915	0.9918	0.9922	0.9925
1.9	0.9928	0.9931	0.9934	0.9937	0.9939	0.9942	0.9944	0.9947	0.9949	0.9951
2.0	0.9953	0.9955	0.9957	0.9959	0.9961	0.9963	0.9964	0.9966	0.9967	0.9969
2.1	0.9970	0.9972	0.9973	0.9974	0.9975	0.9976	0.9977	0.9979	0.9980	0.9980
2.2	0.9981	0.9982	0.9983	0.9984	0.9985	0.9985	0.9986	0.9987	0.9987	0.9988
2.3	0.9989	0.9989	0.9990	0.9990	0.9991	0.9991	0.9992	0.9992	0.9992	0.9993
2.4	0.9993	0.9993	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995	0.9995	0.9996
2.5	0.9996	0.9996	0.9996	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
2.6	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9999
2.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
2.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	1.000	1.000	1.000
2.9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

```
0.95001,
0.96385,
0.99001,
0.03615 = 1-0.96385.
```

The calculation for other distributions proceeds along the same lines.

Table D.3 Quantiles χ_p^2 of the χ^2 distribution (3.21) with ν degrees of freedom for some typical (most commonly used) values of p from 0.005 to 0.999

ν	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.1}$	$\chi^2_{.25}$	$\chi^2_{.5}$	$\chi^2_{.75}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$	$\chi^2_{.999}$
1	0.000	0.0002	0.0010	0.0039	0.0158	0.102	0.455	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	0.010	0.0201	0.0506	0.103	0.211	0.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	0.072	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6	43.8
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0	45.3
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8	48.3

(continued)

Table D.3 (continued)

ν	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.1}$	$\chi^2_{.25}$	$\chi^2_{.5}$	$\chi^2_{.75}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$	$\chi^2_{.999}$
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6	51.2
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3	58.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7	59.7
40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8	73.4
50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.5	71.4	76.2	79.5	86.7
60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.4	79.1	83.3	88.4	92.0	99.6
70	43.3	45.4	48.8	51.7	55.3	61.7	69.3	77.6	85.5	90.5	95.0	100	104	112
80	51.2	53.5	57.2	60.4	64.3	71.1	79.3	88.1	96.6	102	107	112	116	125
90	59.2	61.8	65.6	69.1	73.3	80.6	89.3	98.6	108	113	118	124	128	137
100	67.3	70.1	74.2	77.9	82.4	90.1	99.3	109	118	124	130	136	140	149

Table D.4 Quantiles t_p of the Student's t distribution (3.22) with ν degrees of freedom for some typical (most commonly used) values of p from 0.55 to 0.999

ν	$t_{0.55}$	$t_{0.60}$	$t_{0.70}$	$t_{0.75}$	$t_{0.80}$	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$	$t_{0.999}$
1	0.158	0.325	0.727	1.000	1.376	3.08	6.31	12.7	31.8	63.7	3183
2	0.142	0.289	0.617	0.816	1.061	1.89	2.92	4.30	6.96	9.92	70.7
3	0.137	0.277	0.584	0.765	0.978	1.64	2.35	3.18	4.54	5.84	22.2
4	0.134	0.271	0.569	0.741	0.941	1.53	2.13	2.78	3.75	4.60	13.0
5	0.132	0.267	0.559	0.727	0.920	1.48	2.02	2.57	3.36	4.03	9.68
6	0.131	0.265	0.553	0.718	0.906	1.44	1.94	2.45	3.14	3.71	8.02
7	0.130	0.263	0.549	0.711	0.896	1.41	1.89	2.36	3.00	3.50	7.06
8	0.130	0.262	0.546	0.706	0.889	1.40	1.86	2.31	2.90	3.36	6.44
9	0.129	0.261	0.543	0.703	0.883	1.38	1.83	2.26	2.82	3.25	6.01
10	0.129	0.260	0.542	0.700	0.879	1.37	1.81	2.23	2.76	3.17	5.69
11	0.129	0.260	0.540	0.697	0.876	1.36	1.80	2.20	2.72	3.11	5.45
12	0.128	0.259	0.539	0.695	0.873	1.36	1.78	2.18	2.68	3.05	5.26
13	0.128	0.259	0.538	0.694	0.870	1.35	1.77	2.16	2.65	3.01	5.11
14	0.128	0.258	0.537	0.692	0.868	1.35	1.76	2.14	2.62	2.98	4.99
15	0.128	0.258	0.536	0.691	0.866	1.34	1.75	2.13	2.60	2.95	4.88
16	0.128	0.258	0.535	0.690	0.865	1.34	1.75	2.12	2.58	2.92	4.79
17	0.128	0.257	0.534	0.689	0.863	1.33	1.74	2.11	2.57	2.90	4.71
18	0.127	0.257	0.534	0.688	0.862	1.33	1.73	2.10	2.55	2.88	4.65
19	0.127	0.257	0.533	0.688	0.861	1.33	1.73	2.09	2.54	2.86	4.59
20	0.127	0.257	0.533	0.687	0.860	1.33	1.72	2.09	2.53	2.85	4.54
21	0.127	0.257	0.532	0.686	0.859	1.32	1.72	2.08	2.52	2.83	4.49

(continued)

Table D.4 (continued)

ν	$t_{0.55}$	$t_{0.60}$	$t_{0.70}$	$t_{0.75}$	$t_{0.80}$	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$	$t_{0.999}$
22	0.127	0.256	0.532	0.686	0.858	1.32	1.72	2.07	2.51	2.82	4.45
23	0.127	0.256	0.532	0.685	0.858	1.32	1.71	2.07	2.50	2.81	4.42
24	0.127	0.256	0.531	0.685	0.857	1.32	1.71	2.06	2.49	2.80	4.38
25	0.127	0.256	0.531	0.684	0.856	1.32	1.71	2.06	2.49	2.79	4.35
26	0.127	0.256	0.531	0.684	0.856	1.31	1.71	2.06	2.48	2.78	4.32
27	0.127	0.256	0.531	0.684	0.855	1.31	1.70	2.05	2.47	2.77	4.30
28	0.127	0.256	0.530	0.683	0.855	1.31	1.70	2.05	2.47	2.76	4.28
29	0.127	0.256	0.530	0.683	0.854	1.31	1.70	2.05	2.46	2.76	4.25
30	0.127	0.256	0.530	0.683	0.854	1.31	1.70	2.04	2.46	2.75	4.23
40	0.126	0.255	0.529	0.681	0.851	1.30	1.68	2.02	2.42	2.70	4.09
60	0.126	0.254	0.527	0.679	0.848	1.30	1.67	2.00	2.39	2.66	3.96
120	0.126	0.254	0.526	0.677	0.845	1.29	1.66	1.98	2.36	2.62	3.84
∞	0.126	0.253	0.524	0.674	0.842	1.28	1.64	1.96	2.21	2.58	3.72

Table D.5 95. percentiles ($F_{0.95}$) of the F distribution (3.23); ν_1 degrees of freedom in the numerator and ν_2 in the denominator

	$\nu_1 = 1$	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
$\nu_2 = 1$	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69

(continued)

Table D.5 (continued)

	$\nu_1 = 1$	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.41	2.11	1.92	1.79	1.70	1.62	1.56	1.52	1.44	1.37	1.28	1.52	1.46	1.39	1.32	1.22	1.00

Table D.6 99. percentiles ($F_{0.99}$) of the F distribution (3.23); ν_1 degrees of freedom in the numerator and ν_2 in the denominator

	$\nu_1 = 1$	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
$\nu_2 = 1$	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.4	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.6	14.4	14.2	14.0	13.9	13.8	13.8	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.8	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.3	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13

(continued)

Table D.6 (continued)

	$\nu_1 = 1$	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Reference

1. S. Wolfram, Wolfram MATHEMATICA. http://www.wolfram.com

Index

S. Širca, *Probability for Physicists*, Graduate Texts in Physics,
DOI 10.1007/978-3-319-31611-6