F.M. Dekking   C. Kraaikamp
H.P. Lopuhaä   L.E. Meester

A Modern Introduction to Probability and Statistics

Understanding Why and How

With 120 Figures

Springer

Frederik Michel Dekking
Cornelis Kraaikamp
Hendrik Paul Lopuhaä
Ludolf Erwin Meester
Delft Institute of Applied Mathematics
Delft University of Technology
Mekelweg 4
2628 CD Delft
The Netherlands

Preface

Probability and statistics are fascinating subjects on the interface between mathematics and applied sciences that help us understand and solve practical problems. We believe that you, by learning how stochastic methods come about and why they work, will be able to understand the meaning of statistical statements as well as judge the quality of their content, when facing such problems on your own. Our philosophy is one of *how* and *why*: instead of just presenting stochastic methods as cookbook recipes, we prefer to explain the principles behind them.

In this book you will find the basics of probability theory and statistics. In addition, there are several topics that go somewhat beyond the basics but that ought to be present in an introductory course: simulation, the Poisson process, the law of large numbers, and the central limit theorem. Computers have brought many changes in statistics. In particular, the bootstrap has earned its place. It provides the possibility to derive confidence intervals and perform tests of hypotheses where traditional (normal approximation or large sample) methods are inappropriate. It is a modern useful tool one should learn about, we believe.

Examples and datasets in this book are mostly from real-life situations, at least that is what we looked for in illustrations of the material. Anybody who has inspected datasets with the purpose of using them as elementary examples knows that this is hard: on the one hand, you do not want to boldly state assumptions that are clearly not satisfied; on the other hand, long explanations concerning side issues distract from the main points. We hope that we found a good middle way.

A first course in calculus is needed as a prerequisite for this book. In addition to high-school algebra, some infinite series are used (exponential, geometric). Integration and differentiation are the most important skills, mainly concerning one variable (the exceptions, two dimensional integrals, are encountered in Chapters 9–11). Although the mathematics is kept to a minimum, we strived

to be mathematically correct throughout the book. With respect to probability and statistics the book is self-contained.

The book is aimed at undergraduate engineering students, and students from more business-oriented studies (who may gloss over some of the more mathematically oriented parts). At our own university we also use it for students in applied mathematics (where we put a little more emphasis on the math and add topics like combinatorics, conditional expectations, and generating functions). It is designed for a one-semester course: on average two hours in class per chapter, the first for a lecture, the second doing exercises. The material is also well-suited for self-study, as we know from experience.

We have divided attention about evenly between probability and statistics. The very first chapter is a sampler with differently flavored introductory examples, ranging from scientific success stories to a controversial puzzle. Topics that follow are elementary probability theory, simulation, joint distributions, the law of large numbers, the central limit theorem, statistical modeling (informal: why and how we can draw inference from data), data analysis, the bootstrap, estimation, simple linear regression, confidence intervals, and hypothesis testing. Instead of a few chapters with a long list of discrete and continuous distributions, with an enumeration of the important attributes of each, we introduce a few distributions when presenting the concepts and the others where they arise (more) naturally. A list of distributions and their characteristics is found in Appendix A.

With the exception of the first one, chapters in this book consist of three main parts. First, about four sections discussing new material, interspersed with a handful of so-called Quick exercises. Working these—two-or-three-minute—exercises should help to master the material and provide a break from reading to do something more active. On about two dozen occasions you will find indented paragraphs labeled *Remark*, where we felt the need to discuss more mathematical details or background material. These remarks can be skipped without loss of continuity; in most cases they require a bit more mathematical maturity. Whenever persons are introduced in examples we have determined their sex by looking at the chapter number and applying the rule "He is odd, she is even." Solutions to the quick exercises are found in the second to last section of each chapter.

The last section of each chapter is devoted to exercises, on average thirteen per chapter. For about half of the exercises, answers are given in Appendix C, and for half of these, full solutions in Appendix D. Exercises with both a short answer and a full solution are marked with ⊞ and those with only a short answer are marked with ⊡ (when more appropriate, for example, in "Show that . . ." exercises, the short answer provides a hint to the key step). Typically, the section starts with some easy exercises and the order of the material in the chapter is more or less respected. More challenging exercises are found at the end.

Much of the material in this book would benefit from illustration with a computer using statistical software. A complete course should also involve computer exercises. Topics like simulation, the law of large numbers, the central limit theorem, and the bootstrap loudly call for this kind of experience. For this purpose, all the datasets discussed in the book are available at http://www.springeronline.com/1-85233-896-2. The same Web site also provides access, for instructors, to a complete set of solutions to the exercises; go to the Springer online catalog or contact textbooks@springer-sbm.com to apply for your password.

Delft, The Netherlands   F. M. Dekking
January 2005   C. Kraaikamp
                                                           H. P. Lopuhaä
                                                            L. E. Meester

Contents

1

Why probability and statistics?

Is everything on this planet determined by randomness? This question is open to philosophical debate. What is certain is that every day thousands and thousands of engineers, scientists, business persons, manufacturers, and others are using tools from probability and statistics.

The theory and practice of probability and statistics were developed during the last century and are still actively being refined and extended. In this book we will introduce the basic notions and ideas, and in this first chapter we present a diverse collection of examples where randomness plays a role.

1.1 Biometry: iris recognition

Biometry is the art of identifying a person on the basis of his or her personal biological characteristics, such as fingerprints or voice. From recent research it appears that with the human iris one can beat all existing automatic human identification systems. Iris recognition technology is based on the visible qualities of the iris. It converts these—via a video camera—into an "iris code" consisting of just 2048 bits. This is done in such a way that the code is hardly sensitive to the size of the iris or the size of the pupil. However, at different times and different places the iris code of the same person will not be exactly the same. Thus one has to allow for a certain percentage of mismatching bits when identifying a person. In fact, the system allows about 34% mismatches! How can this lead to a reliable identification system? The miracle is that different persons have very different irides. In particular, over a large collection of different irides the code bits take the values 0 and 1 about half of the time. But that is certainly not sufficient: if one bit would determine the other 2047, then we could only distinguish two persons. In other words, single bits may be random, but the correlation between bits is also crucial (we will discuss correlation at length in Chapter 10). John Daugman who has developed the iris recognition technology made comparisons between 222 743 pairs of iris

codes and concluded that of the 2048 bits 266 may be considered as uncorrelated ([6]). He then argues that we may consider an iris code as the result of 266 coin tosses with a fair coin. This implies that if we compare two such codes from different persons, then there is an astronomically small probability that these two differ in less than 34% of the bits—almost all pairs will differ in about 50% of the bits. This is illustrated in Figure 1.1, which originates from [6], and was kindly provided by John Daugman. The iris code data consist of numbers between 0 and 1, each a Hamming distance (the fraction of mismatches) between two iris codes. The data have been summarized in two histograms, that is, two graphs that show the number of counts of Hamming distances falling in a certain interval. We will encounter histograms and other summaries of data in Chapter 15. One sees from the figure that for codes from the same iris (left side) the mismatch fraction is only about 0.09, while for different irides (right side) it is about 0.46.
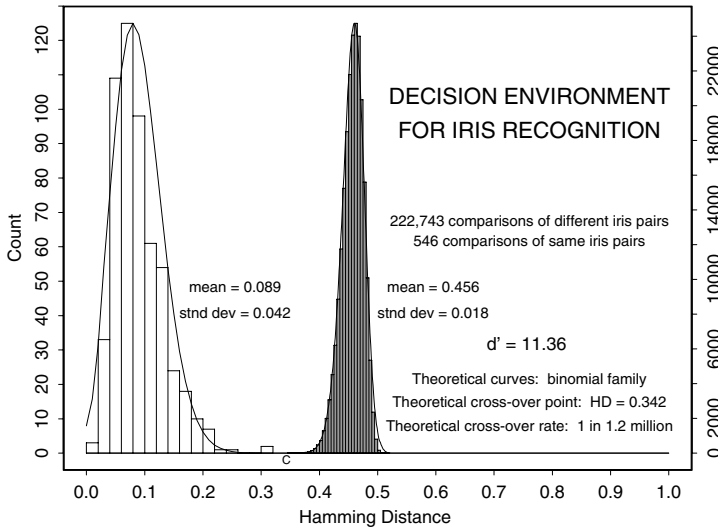
Fig. 1.1. Comparison of same and different iris pairs.

You may still wonder how it is possible that irides distinguish people so well. What about twins, for instance? The surprising thing is that although the color of eyes is hereditary, many features of iris patterns seem to be produced by so-called epigenetic events. This means that during embryo development the iris structure develops randomly. In particular, the iris patterns of (monozygotic) twins are as discrepant as those of two arbitrary individuals.

For this reason, as early as in the 1930s, eye specialists proposed that iris patterns might be used for identification purposes.

1.2 Killer football

A couple of years ago the prestigious *British Medical Journal* published a paper with the title "Cardiovascular mortality in Dutch men during 1996 European football championship: longitudinal population study" ([41]). The authors claim to have shown that the effect of a single football match is detectable in national mortality data. They consider the mortality from infarctions (heart attacks) and strokes, and the "explanation" of the increase is a combination of heavy alcohol consumption and stress caused by watching the football match on June 22 between the Netherlands and France (lost by the Dutch team!). The authors mainly support their claim with a figure like Figure 1.2, which shows the number of deaths from the causes mentioned (for men over 45), during the period June 17 to June 27, 1996. The middle horizontal line marks the average number of deaths on these days, and the upper and lower horizontal lines mark what the authors call the 95% confidence interval. The construction of such an interval is usually performed with standard statistical techniques, which you will learn in Chapter 23. The interpretation of such an interval is rather tricky. That the bar on June 22 sticks out off the confidence interval should support the "killer claim."
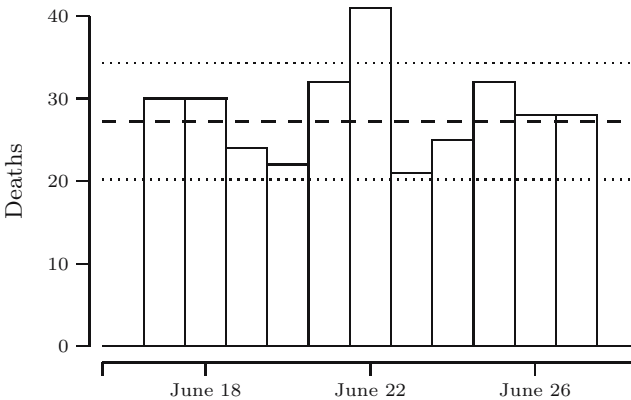
Fig. 1.2. Number of deaths from infarction or stroke in (part of) June 1996.

It is rather surprising that such a conclusion is based on a *single* football match, and one could wonder why no probability model is proposed in the paper. In fact, as we shall see in Chapter 12, it would not be a bad idea to model the time points at which deaths occur as a so-called Poisson process.

Once we have done this, we can compute how often a pattern like the one in the figure might occur—without paying attention to football matches and other high-risk national events. To do this we need the mean number of deaths per day. This number can be obtained from the data by an estimation procedure (the subject of Chapters 19 to 23). We use the sample mean, which is equal to $(10 \cdot 27.2 + 41)/11 = 313/11 = 28.45$. (Here we have to make a computation like this because we only use the data in the paper: 27.2 is the average over the 5 days preceding and following the match, and 41 is the number of deaths on the day of the match.) Now let p_{high} be the probability that there are 41 or more deaths on a day, and let p_{usual} be the probability that there are between 21 and 34 deaths on a day—here 21 and 34 are the lowest and the highest number that fall in the interval in Figure 1.2. From the formula of the Poisson distribution given in Chapter 12 one can compute that $p_{\text{high}} = 0.008$ and $p_{\text{usual}} = 0.820$. Since events on different days are independent according to the Poisson process model, the probability p of a pattern as in the figure is

$$p = p_{\text{usual}}^5 \cdot p_{\text{high}} \cdot p_{\text{usual}}^5 = 0.0011.$$

From this it can be shown by (a generalization of) the law of large numbers (which we will study in Chapter 13) that such a pattern would appear about once every $1/0.0011 = 899$ days. So it is not overwhelmingly exceptional to find such a pattern, and the fact that there was an important football match on the day in the middle of the pattern might just have been a coincidence.

1.3 Cars and goats: the Monty Hall dilemma

On Sunday September 9, 1990, the following question appeared in the "Ask Marilyn" column in *Parade*, a Sunday supplement to many newspapers across the United States:

> Suppose you're on a game show, and you're given the choice of three doors; behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?—Craig F. Whitaker, Columbia, Md.

Marilyn's answer—one should switch—caused an avalanche of reactions, in total an estimated 10 000. Some of these reactions were not so flattering ("You are the goat"), quite a lot were by professional mathematicians ("You blew it, and blew it big," "You are utterly incorrect .... How many irate mathematicians are needed to change your mind?"). Perhaps some of the reactions were so strong, because Marilyn vos Savant, the author of the column, is in the *Guinness Book of Records* for having one of the highest IQs in the world.

The switching question was inspired by Monty Hall's "Let's Make a Deal" game show, which ran with small interruptions for 23 years on various U.S. television networks.

Although it is not explicitly stated in the question, the game show host will *always* open a door with a goat after you make your initial choice. Many people would argue that in this situation it does not matter whether one would change or not: one door has a car behind it, the other a goat, so the odds to get the car are fifty-fifty. To see why they are wrong, consider the following argument. In the original situation two of the three doors have a goat behind them, so with probability 2/3 your initial choice was wrong, and with probability 1/3 it was right. Now the host opens a door with a goat (note that he can always do this). In case your initial choice was *wrong* the host has only one option to show a door with a goat, and switching leads you to the door with the car. In case your initial choice was *right* the host has two goats to choose from, so switching will lead you to a goat. We see that switching is the best strategy, doubling our chances to win. To stress this argument, consider the following generalization of the problem: suppose there are 10 000 doors, behind one is a car and behind the rest, goats. After you make your choice, the host will open 9998 doors with goats, and offers you the option to switch. To change or not to change, that's the question! Still not convinced? Use your Internet browser to find one of the zillion sites where one can run a simulation of the Monty Hall problem (more about simulation in Chapter 6).

In fact, there are quite a lot of variations on the problem. For example, the situation that there are four doors: you select a door, the host always opens a door with a goat, and offers you to select another door. After you have made up your mind he opens a door with a goat, and again offers you to switch. After you have decided, he opens the door you selected. What is now the best strategy? In this situation switching only at the last possible moment yields a probability of 3/4 to bring the car home. Using the law of total probability from Section 3.3 you will find that this is indeed the best possible strategy.

1.4 The space shuttle *Challenger*

On January 28, 1986, the space shuttle *Challenger* exploded about one minute after it had taken off from the launch pad at Kennedy Space Center in Florida. The seven astronauts on board were killed and the spacecraft was destroyed. The cause of the disaster was explosion of the main fuel tank, caused by flames of hot gas erupting from one of the so-called solid rocket boosters.

These solid rocket boosters had been cause for concern since the early years of the shuttle. They are manufactured in segments, which are joined at a later stage, resulting in a number of joints that are sealed to protect against leakage. This is done with so-called O-rings, which in turn are protected by a layer of putty. When the rocket motor ignites, high pressure and high temperature

build up within. In time these may burn away the putty and subsequently erode the O-rings, eventually causing hot flames to erupt on the outside. In a nutshell, this is what actually happened to the *Challenger*.

After the explosion, an investigative commission determined the causes of the disaster, and a report was issued with many findings and recommendations ([24]). On the evening of January 27, a decision to launch the next day had been made, notwithstanding the fact that an extremely low temperature of 31°F had been predicted, well below the operating limit of 40°F set by Morton Thiokol, the manufacturer of the solid rocket boosters. Apparently, a "management decision" was made to overrule the engineers' recommendation not to launch. The inquiry faulted both NASA and Morton Thiokol management for giving in to the pressure to launch, ignoring warnings about problems with the seals.

The *Challenger* launch was the 24th of the space shuttle program, and we shall look at the data on the number of failed O-rings, available from previous launches (see [5] for more details). Each rocket has three O-rings, and two rocket boosters are used per launch, so in total six O-rings are used each time. Because low temperatures are known to adversely affect the O-rings, we also look at the corresponding launch temperature. In Figure 1.3 the dots show the number of failed O-rings per mission (there are 23 dots—one time the boosters could not be recovered from the ocean; temperatures are rounded to the nearest degree Fahrenheit; in case of two or more equal data points these are shifted slightly.). If you ignore the dots representing zero failures, which all occurred at high temperatures, a temperature effect is not apparent.
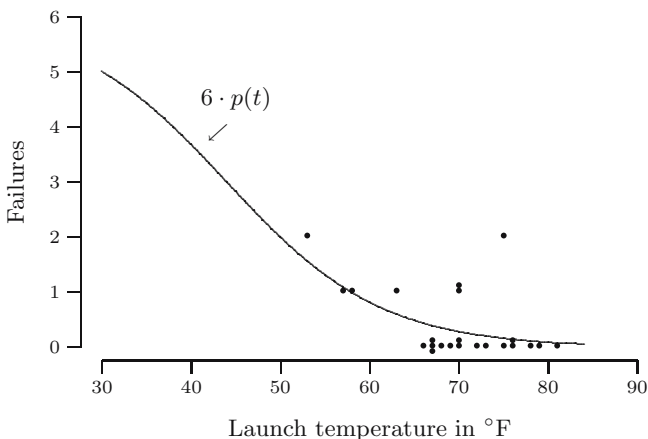
Fig. 1.3. Space shuttle failure data of pre-*Challenger* missions and fitted model of expected number of failures per mission function.

In a model to describe these data, the probability $p(t)$ that an individual O-ring fails should depend on the launch temperature t. Per mission, the number of failed O-rings follows a so-called binomial distribution: six O-rings, and each may fail with probability $p(t)$; more about this distribution and the circumstances under which it arises can be found in Chapter 4. A *logistic* model was used in [5] to describe the dependence on t:

$$p(t) = \frac{e^{a+b\cdot t}}{1 + e^{a+b\cdot t}}.$$

A high value of $a + b \cdot t$ corresponds to a high value of $p(t)$, a low value to low $p(t)$. Values of a and b were determined from the data, according to the following principle: choose a and b so that the probability that we get data as in Figure 1.3 is as high as possible. This is an example of the use of the method of maximum likelihood, which we shall discuss in Chapter 21. This results in $a = 5.085$ and $b = -0.1156$, which indeed leads to lower probabilities at higher temperatures, and to $p(31) = 0.8178$. We can also compute the (estimated) expected number of failures, $6 \cdot p(t)$, as a function of the launch temperature t; this is the plotted line in the figure.

Combining the estimates with estimated probabilities of other events that should happen for a *complete* failure of the field-joint, the estimated probability of such a failure is 0.023. With six field-joints, the probability of at least one complete failure is then $1 - (1 - 0.023)^6 = 0.13$!

1.5 Statistics versus intelligence agencies

During World War II, information about Germany's war potential was essential to the Allied forces in order to schedule the time of invasions and to carry out the allied strategic bombing program. Methods for estimating German production used during the early phases of the war proved to be inadequate. In order to obtain more reliable estimates of German war production, experts from the Economic Warfare Division of the American Embassy and the British Ministry of Economic Warfare started to analyze markings and serial numbers obtained from captured German equipment.

Each piece of enemy equipment was labeled with markings, which included all or some portion of the following information: (a) the name and location of the marker; (b) the date of manufacture; (c) a serial number; and (d) miscellaneous markings such as trademarks, mold numbers, casting numbers, etc. The purpose of these markings was to maintain an effective check on production standards and to perform spare parts control. However, these same markings offered Allied intelligence a wealth of information about German industry.

The first products to be analyzed were tires taken from German aircraft shot over Britain and from supply dumps of aircraft and motor vehicle tires captured in North Africa. The marking on each tire contained the maker's name,

a serial number, and a two-letter code for the date of manufacture. The first step in analyzing the tire markings involved breaking the two-letter date code. It was conjectured that one letter represented the month and the other the year of manufacture, and that there should be 12 letter variations for the month code and 3 to 6 for the year code. This, indeed, turned out to be true. The following table presents examples of the 12 letter variations used by four different manufacturers.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Dunlop	T	I	E	B	R	A	P	O	L	N	U	D
Fulda	F	U	L	D	A	M	U	N	S	T	E	R
Phoenix	F	O	N	I	X	H	A	M	B	U	R	G
Sempirit	A	B	C	D	E	F	G	H	I	J	K	L

For instance, the Dunlop code was Dunlop Arbeit spelled backwards. Next, the year code was broken and the numbering system was solved so that for each manufacturer individually the serial numbers could be dated. Moreover, for each month, the serial numbers could be recoded to numbers running from 1 to some unknown largest number N, and the observed (recoded) serial numbers could be seen as a subset of this. The objective was to estimate N for each month and each manufacturer separately by means of the observed (recoded) serial numbers. In Chapter 20 we discuss two different methods of estimation, and we show that the method based on only the maximum observed (recoded) serial number is much better than the method based on the average observed (recoded) serial numbers.

With a sample of about 1400 tires from five producers, individual monthly output figures were obtained for almost all months over a period from 1939 to mid-1943. The following table compares the accuracy of estimates of the average monthly production of all manufacturers of the first quarter of 1943 with the statistics of the Speer Ministry that became available after the war. The accuracy of the estimates can be appreciated even more if we compare them with the figures obtained by Allied intelligence agencies. They estimated, using other methods, the production between 900 000 and 1 200 000 per month!

Type of tire	Estimated production	Actual production
Truck and passenger car	147 000	159 000
Aircraft	28 500	26 400
	———	———
Total	175 500	186 100

1.6 The speed of light

In 1983 the definition of the meter (the SI unit of one meter) was changed to: *The meter is the length of the path traveled by light in vacuum during a time interval of* 1/299 792 458 *of a second.* This implicitly defines the speed of light as 299 792 458 meters per second. It was done because one thought that the speed of light was so accurately known that it made more sense to define the meter in terms of the speed of light rather than vice versa, a remarkable end to a long story of scientific discovery. For a long time most scientists believed that the speed of light was infinite. Early experiments devised to demonstrate the finiteness of the speed of light failed because the speed is so extraordinarily high. In the 18th century this debate was settled, and work started on determination of the speed, using astronomical observations, but a century later scientists turned to earth-based experiments. Albert Michelson refined experimental arrangements from two previous experiments and conducted a series of measurements in June and early July of 1879, at the U.S. Naval Academy in Annapolis. In this section we give a very short summary of his work. It is extracted from an article in *Statistical Science* ([18]).

The principle of speed measurement is easy, of course: measure a distance and the time it takes to travel that distance, the speed equals distance divided by time. For an accurate determination, both the distance and the time need to be measured accurately, and with the speed of light this is a problem: either we should use a very large distance and the accuracy of the distance measurement is a problem, or we have a very short time interval, which is also very difficult to measure accurately.

In Michelson's time it was known that the speed of light was about 300 000 km/s, and he embarked on his study with the goal of an improved value of the speed of light. His experimental setup is depicted schematically in Figure 1.4. Light emitted from a light source is aimed, through a slit in a fixed plate, at a rotating mirror; we call its distance from the plate the radius. At one particular angle, this rotating mirror reflects the beam in the direction of a distant (fixed) flat mirror. On its way the light first passes through a focusing lens. This second mirror is positioned in such a way that it reflects the beam back in the direction of the rotating mirror. In the time it takes the light to travel back and forth between the two mirrors, the rotating mirror has moved by an angle α, resulting in a reflection on the plate that is displaced with respect to the source beam that passed through the slit. The radius and the displacement determine the angle α because

$$\tan 2\alpha = \frac{\text{displacement}}{\text{radius}}$$

and combined with the number of revolutions per seconds (rps) of the mirror, this determines the elapsed time:

$$\text{time} = \frac{\alpha/2\pi}{\text{rps}}.$$
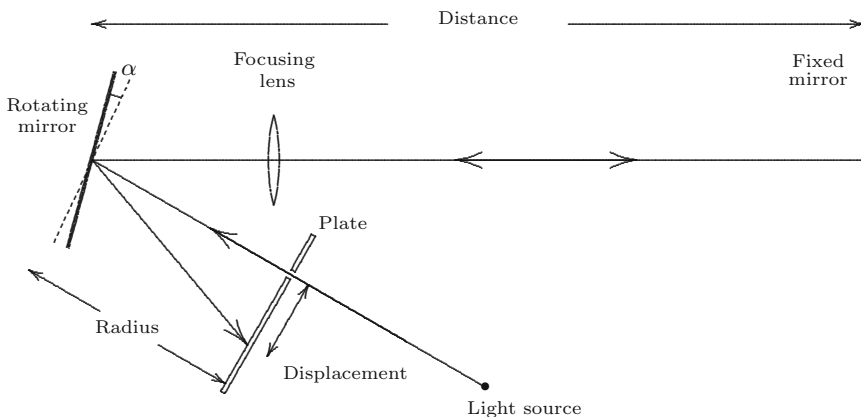
Fig. 1.4. Michelson's experiment.

During this time the light traveled twice the distance between the mirrors, so the speed of light in air now follows:

$$c_{\text{air}} = \frac{2 \cdot \text{distance}}{\text{time}}.$$

All in all, it looks simple: just measure the four quantities—distance, radius, displacement and the revolutions per second—and do the calculations. This is much harder than it looks, and problems in the form of inaccuracies are lurking everywhere. An error in any of these quantities translates directly into some error in the final result.

Michelson did the utmost to reduce errors. For example, the distance between the mirrors was about 2000 feet, and to measure it he used a steel measuring tape. Its nominal length was 100 feet, but he carefully checked this using a copy of the official "standard yard." He found that the tape was in fact 100.006 feet. This way he eliminated a (small) systematic error.

Now imagine using the tape to measure a distance of 2000 feet: you have to use the tape 20 times, each time marking the next 100 feet. Do it again, and you probably find a slightly different answer, no matter how hard you try to be very precise in every step of the measuring procedure. This kind of variation is inevitable: sometimes we end up with a value that is a bit too high, other times it is too low, but on average we're doing okay—assuming that we have eliminated sources of systematic error, as in the measuring tape. Michelson measured the distance five times, which resulted in values between 1984.93 and 1985.17 feet (after correcting for the temperature-dependent stretch), and he used the average as the "true distance."

In many phases of the measuring process Michelson attempted to identify and determine systematic errors and subsequently applied corrections. He

also systematically repeated measuring steps and averaged the results to re-
duce variability. His final dataset consists of 100 separate measurements (see
Table 17.1), but each is in fact summarized and averaged from repeated mea-
surements on several variables. The final result he reported was that the speed
of light in vacuum (this involved a conversion) was $299\,944 \pm 51$ km/s, where
the 51 is an indication of the uncertainty in the answer. In retrospect, we must
conclude that, in spite of Michelson's admirable meticulousness, some source
of error must have slipped his attention, as his result is off by about 150 km/s.
With current methods we would derive from his data a so-called 95% confi-
dence interval: $299\,944 \pm 15.5$ km/s, suggesting that Michelson's uncertainty
analysis was a little conservative. The methods used to construct confidence
intervals are the topic of Chapters 23 and 24.

2

Outcomes, events, and probability

The world around us is full of phenomena we perceive as random or unpredictable. We aim to model these phenomena as *outcomes* of some experiment, where you should think of *experiment* in a very general sense. The outcomes are elements of a *sample space* Ω, and subsets of Ω are called *events.* The events will be assigned a *probability*, a number between 0 and 1 that expresses how likely the event is to occur.

2.1 Sample spaces

Sample spaces are simply sets whose elements describe the outcomes of the experiment in which we are interested.

We start with the most basic experiment: the tossing of a coin. Assuming that we will never see the coin land on its rim, there are two possible outcomes: heads and tails. We therefore take as the sample space associated with this experiment the set $\Omega = \{H, T\}$.

In another experiment we ask the next person we meet on the street in which month her birthday falls. An obvious choice for the sample space is

$$\Omega = \{\text{Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec}\}.$$

In a third experiment we load a scale model for a bridge up to the point where the structure collapses. The outcome is the load at which this occurs. In reality, one can only measure with finite accuracy, e.g., to five decimals, and a sample space with just those numbers would strictly be adequate. However, in principle, the load itself could be any positive number and therefore $\Omega = (0, \infty)$ is the right choice. Even though in reality there may also be an upper limit to what loads are conceivable, it is not necessary or practical to try to limit the outcomes correspondingly.

In a fourth experiment, we find on our doormat three envelopes, sent to us by three different persons, and we look in which order the envelopes lie on top of each other. Coding them 1, 2, and 3, the sample space would be

$$\Omega = \{123, 132, 213, 231, 312, 321\}.$$

QUICK EXERCISE 2.1 If we received mail from four different persons, how many elements would the corresponding sample space have?

In general one might consider the order in which n different objects can be placed. This is called a *permutation* of the n objects. As we have seen, there are 6 possible permutations of 3 objects, and $4 \cdot 6 = 24$ of 4 objects. What happens is that if we add the nth object, then this can be placed in any of n positions in any of the permutations of $n - 1$ objects. Therefore there are

$$n \cdot (n - 1) \cdot \cdots \cdot 3 \cdot 2 \cdot 1 = n!$$

possible permutations of n objects. Here $n!$ is the standard notation for this product and is pronounced "n factorial." It is convenient to define $0! = 1$.

2.2 Events

Subsets of the sample space are called *events*. We say that an event A *occurs* if the outcome of the experiment is an element of the set A. For example, in the birthday experiment we can ask for the outcomes that correspond to a long month, i.e., a month with 31 days. This is the event

$$L = \{\text{Jan, Mar, May, Jul, Aug, Oct, Dec}\}.$$

Events may be combined according to the usual set operations.

For example if R is the event that corresponds to the months that have the letter r in their (full) name (so $R = \{\text{Jan, Feb, Mar, Apr, Sep, Oct, Nov, Dec}\}$), then the long months that contain the letter r are

$$L \cap R = \{\text{Jan, Mar, Oct, Dec}\}.$$

The set $L \cap R$ is called the *intersection* of L and R and occurs if both L and R occur. Similarly, we have the *union* $A \cup B$ of two sets A and B, which occurs if at least one of the events A and B occurs. Another common operation is taking complements. The event $A^c = \{\omega \in \Omega : \omega \notin A\}$ is called the *complement* of A; it occurs if and only if A does *not* occur. The complement of Ω is denoted $\emptyset$, the empty set, which represents the impossible event. Figure 2.1 illustrates these three set operations.
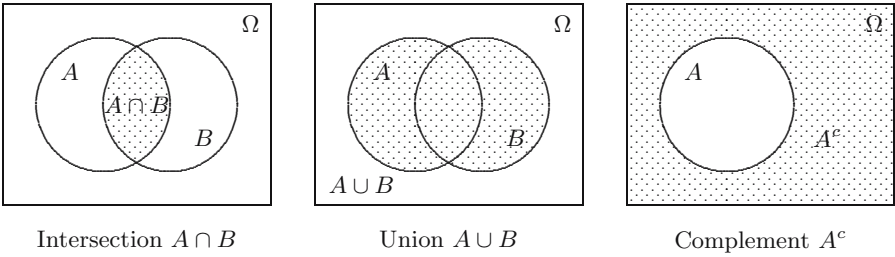
Fig. 2.1. Diagrams of intersection, union, and complement.

We call events A and B *disjoint* or *mutually exclusive* if A and B have no outcomes in common; in set terminology: $A \cap B = \emptyset$. For example, the event L "the birthday falls in a long month" and the event {Feb} are disjoint.

Finally, we say that event A *implies* event B if the outcomes of A also lie in B. In set notation: $A \subset B$; see Figure 2.2.

Some people like to use double negations:

"It is certainly not true that neither John nor Mary is to blame."

This is equivalent to: "John or Mary is to blame, or both." The following useful rules formalize this mental operation to a manipulation with events.

DeMorgan's laws. For any two events A and B we have

$$(A \cup B)^c = A^c \cap B^c \text{ and } (A \cap B)^c = A^c \cup B^c.$$

Quick exercise 2.2 Let J be the event "John is to blame" and M the event "Mary is to blame." Express the two statements above in terms of the events J, J^c, M, and M^c, and check the equivalence of the statements by means of DeMorgan's laws.
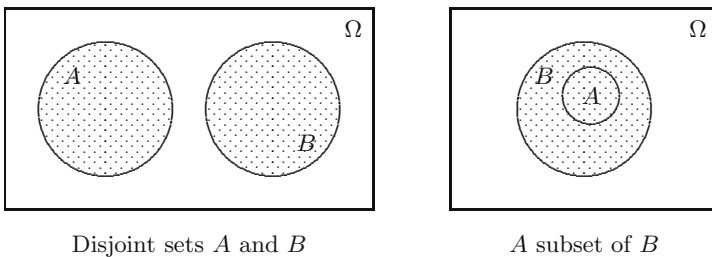
Fig. 2.2. Minimal and maximal intersection of two sets.

2.3 Probability

We want to express how likely it is that an event occurs. To do this we will assign a probability to each event. The assignment of probabilities to events is in general not an easy task, and some of the coming chapters will be dedicated directly or indirectly to this problem. Since *each* event has to be assigned a probability, we speak of a probability *function*. It has to satisfy two basic properties.

> DEFINITION.   A *probability function* P on a finite sample space Ω assigns to each event A in Ω a number $P(A)$ in [0,1] such that
> (i) $P(\Omega) = 1$, and
> (ii) $P(A \cup B) = P(A) + P(B)$ if A and B are disjoint.
> The number $P(A)$ is called the probability that A occurs.

Property (i) expresses that the outcome of the experiment is always an element of the sample space, and property (ii) is the additivity property of a probability function. It implies additivity of the probability function over more than two sets; e.g., if A, B, and C are disjoint events, then the two events $A \cup B$ and C are also disjoint, so

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

We will now look at some examples. When we want to decide whether Peter or Paul has to wash the dishes, we might toss a coin. The fact that we consider this a fair way to decide translates into the opinion that heads and tails are equally likely to occur as the outcome of the coin-tossing experiment. So we put

$$P(\{H\}) = P(\{T\}) = \frac{1}{2}.$$

Formally we have to write $\{H\}$ for the set consisting of the single element H, because a probability function is defined on *events*, not on outcomes. From now on we shall drop these brackets.

Now it might happen, for example due to an asymmetric distribution of the mass over the coin, that the coin is not completely fair. For example, it might be the case that

$$P(H) = 0.4999 \text{ and } P(T) = 0.5001.$$

More generally we can consider experiments with two possible outcomes, say "failure" and "success", which have probabilities $1 - p$ and p to occur, where p is a number between 0 and 1. For example, when our experiment consists of buying a ticket in a lottery with 10 000 tickets and only one prize, where "success" stands for winning the prize, then $p = 10^{-4}$.

How should we assign probabilities in the second experiment, where we ask for the month in which the next person we meet has his or her birthday? In analogy with what we have just done, we put

$$P(\text{Jan}) = P(\text{Feb}) = \cdots = P(\text{Dec}) = \frac{1}{12}.$$

Some of you might object to this and propose that we put, for example,

$$P(\text{Jan}) = \frac{31}{365} \quad \text{and} \quad P(\text{Apr}) = \frac{30}{365},$$

because we have long months and short months. But then the very precise among us might remark that this does not yet take care of leap years.

QUICK EXERCISE 2.3 If you would take care of the leap years, assuming that one in every four years is a leap year (which again is an approximation to reality!), how would you assign a probability to each month?

In the third experiment (the buckling load of a bridge), where the outcomes are real numbers, it is impossible to assign a positive probability to each outcome (there are just too many outcomes!). We shall come back to this problem in Chapter 5, restricting ourselves in this chapter to finite and countably infinite[1] sample spaces.

In the fourth experiment it makes sense to assign equal probabilities to all six outcomes:

$$P(123) = P(132) = P(213) = P(231) = P(312) = P(321) = \frac{1}{6}.$$

Until now we have only assigned probabilities to the individual outcomes of the experiments. To assign probabilities to events we use the additivity property. For instance, to find the probability $P(T)$ of the event T that in the three envelopes experiment envelope 2 is on top we note that

$$P(T) = P(213) + P(231) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

In general, additivity of P implies that the probability of an event is obtained by summing the probabilities of the outcomes belonging to the event.

QUICK EXERCISE 2.4 Compute $P(L)$ and $P(R)$ in the birthday experiment.

Finally we mention a rule that permits us to compute probabilities of events A and B that are *not* disjoint. Note that we can write $A = (A \cap B) \cup (A \cap B^c)$, which is a disjoint union; hence

$$P(A) = P(A \cap B) + P(A \cap B^c).$$

If we split $A \cup B$ in the same way with B and B^c, we obtain the events $(A \cup B) \cap B$, which is simply B and $(A \cup B) \cap B^c$, which is nothing but $A \cap B^c$.

---

[1] This means: although infinite, we can still count them one by one; $\Omega = \{\omega_1, \omega_2, \dots\}$. The interval [0,1] of real numbers is an example of an uncountable sample space.

Thus
$$P(A \cup B) = P(B) + P(A \cap B^c).$$
Eliminating $P(A \cap B^c)$ from these two equations we obtain the following rule.

THE PROBABILITY OF A UNION. For any two events A and B we have
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

From the additivity property we can also find a way to compute probabilities of complements of events: from $A \cup A^c = \Omega$, we deduce that
$$P(A^c) = 1 - P(A).$$

2.4 Products of sample spaces

Basic to statistics is that one usually does not consider *one* experiment, but that the same experiment is performed several times. For example, suppose we throw a coin two times. What is the sample space associated with this new experiment? It is clear that it should be the set
$$\Omega = \{H, T\} \times \{H, T\} = \{(H, H), (H, T), (T, H), (T, T)\}.$$
If in the original experiment we had a fair coin, i.e., $P(H) = P(T)$, then in this new experiment all 4 outcomes again have equal probabilities:
$$P((H, H)) = P((H, T)) = P((T, H)) = P((T, T)) = \frac{1}{4}.$$
Somewhat more generally, if we consider two experiments with sample spaces Ω_1 and Ω_2 then the combined experiment has as its sample space the set
$$\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}.$$
If Ω_1 has r elements and Ω_2 has s elements, then $\Omega_1 \times \Omega_2$ has rs elements. Now suppose that in the first, the second, and the combined experiment all outcomes are equally likely to occur. Then the outcomes in the first experiment have probability $1/r$ to occur, those of the second experiment $1/s$, and those of the combined experiment probability $1/rs$. Motivated by the fact that $1/rs = (1/r) \times (1/s)$, we will assign probability $p_i p_j$ to the outcome (ω_i, ω_j) in the combined experiment, in the case that ω_i has probability p_i and ω_j has probability p_j to occur. One should realize that this is by no means the only way to assign probabilities to the outcomes of a combined experiment. The preceding choice corresponds to the situation where the two experiments do not influence each other in any way. What we mean by this influence will be explained in more detail in the next chapter.

QUICK EXERCISE 2.5 Consider the sample space $\{a_1, a_2, a_3, a_4, a_5, a_6\}$ of some experiment, where outcome a_i has probability p_i for $i = 1, \ldots, 6$. We perform this experiment twice in such a way that the associated probabilities are

$$P((a_i, a_i)) = p_i, \quad \text{and} \quad P((a_i, a_j)) = 0 \text{ if } i \neq j, \quad \text{for } i, j = 1, \ldots, 6.$$

Check that P is a probability function on the sample space $\Omega = \{a_1, \ldots, a_6\} \times \{a_1, \ldots, a_6\}$ of the combined experiment. What is the relationship between the first experiment and the second experiment that is determined by this probability function?

We started this section with the experiment of throwing a coin twice. If we want to learn more about the randomness associated with a particular experiment, then we should repeat it more often, say n times. For example, if we perform an experiment with outcomes 1 (success) and 0 (failure) five times, and we consider the event A "exactly one experiment was a success," then this event is given by the set

$$A = \{(0, 0, 0, 0, 1), (0, 0, 0, 1, 0), (0, 0, 1, 0, 0), (0, 1, 0, 0, 0), (1, 0, 0, 0, 0)\}$$

in $\Omega = \{0, 1\} \times \{0, 1\} \times \{0, 1\} \times \{0, 1\} \times \{0, 1\}$. Moreover, if success has probability p and failure probability $1 - p$, then

$$P(A) = 5 \cdot (1 - p)^4 \cdot p,$$

since there are five outcomes in the event A, each having probability $(1-p)^4 \cdot p$.

QUICK EXERCISE 2.6 What is the probability of the event B "exactly two experiments were successful"?

In general, when we perform an experiment n times, then the corresponding sample space is

$$\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n,$$

where Ω_i for $i = 1, \ldots, n$ is a copy of the sample space of the original experiment. Moreover, we assign probabilities to the outcomes $(\omega_1, \ldots, \omega_n)$ in the standard way described earlier, i.e.,

$$P((\omega_1, \omega_2, \ldots, \omega_n)) = p_1 \cdot p_2 \cdots \cdot p_n,$$

if each ω_i has probability p_i.

2.5 An infinite sample space

We end this chapter with an example of an experiment with infinitely many outcomes. We toss a coin repeatedly until the first head turns up. The outcome

of the experiment is the number of tosses it takes to have this first occurrence of a head. Our sample space is the space of all positive natural numbers

$$\Omega = \{1, 2, 3, \ldots\}.$$

What is the probability function P for this experiment?

Suppose the coin has probability p of falling on heads and probability $1-p$ to fall on tails, where $0 < p < 1$. We determine the probability $P(n)$ for each n. Clearly $P(1) = p$, the probability that we have a head right away. The event $\{2\}$ corresponds to the outcome (T, H) in $\{H, T\} \times \{H, T\}$, so we should have

$$P(2) = (1 - p)p.$$

Similarly, the event $\{n\}$ corresponds to the outcome $(T, T, \ldots, T, T, H)$ in the space $\{H, T\} \times \cdots \times \{H, T\}$. Hence we should have, in general,

$$P(n) = (1 - p)^{n-1} p, \qquad n = 1, 2, 3, \ldots.$$

Does this define a probability function on $\Omega = \{1, 2, 3, \ldots\}$? Then we should at least have $P(\Omega) = 1$. It is not directly clear how to calculate $P(\Omega)$: since the sample space is no longer finite we have to amend the definition of a probability function.

> DEFINITION. A *probability function* on an infinite (or finite) sample space Ω assigns to each event A in Ω a number $P(A)$ in $[0, 1]$ such that
> (i) $P(\Omega) = 1$, and
> (ii) $P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$
>   if $A_1, A_2, A_3, \ldots$ are disjoint events.

Note that this new additivity property is an extension of the previous one because if we choose $A_3 = A_4 = \cdots = \emptyset$, then

$$P(A_1 \cup A_2) = P(A_1 \cup A_2 \cup \emptyset \cup \emptyset \cup \cdots)$$
$$= P(A_1) + P(A_2) + 0 + 0 + \cdots = P(A_1) + P(A_2).$$

Now we can compute the probability of Ω:

$$P(\Omega) = P(1) + P(2) + \cdots + P(n) + \cdots$$
$$= p + (1-p)p + \cdots (1-p)^{n-1} p + \cdots$$
$$= p[1 + (1-p) + \cdots (1-p)^{n-1} + \cdots].$$

The sum $1 + (1-p) + \cdots + (1-p)^{n-1} + \cdots$ is an example of a *geometric series*. It is well known that when $|1-p| < 1$,

$$1 + (1-p) + \cdots + (1-p)^{n-1} + \cdots = \frac{1}{1 - (1-p)} = \frac{1}{p}.$$

Therefore we do indeed have $P(\Omega) = p \cdot \dfrac{1}{p} = 1.$

QUICK EXERCISE 2.7 Suppose an experiment in a laboratory is repeated every day of the week until it is successful, the probability of success being p. The first experiment is started on a Monday. What is the probability that the series ends on the next Sunday?

2.6 Solutions to the quick exercises

2.1 The sample space is $\Omega = \{1234, 1243, 1324, 1342, \ldots, 4321\}$. The best way to count its elements is by noting that for *each* of the 6 outcomes of the three-envelope experiment we can put a fourth envelope in any of 4 positions. Hence Ω has $4 \cdot 6 = 24$ elements.

2.2 The statement "It is certainly not true that neither John nor Mary is to blame" corresponds to the event $(J^c \cap M^c)^c$. The statement "John or Mary is to blame, or both" corresponds to the event $J \cup M$. Equivalence now follows from DeMorgan's laws.

2.3 In four years we have $365 \times 3 + 366 = 1461$ days. Hence long months each have a probability $4 \times 31/1461 = 124/1461$, and short months a probability $120/1461$ to occur. Moreover, $\{\text{Feb}\}$ has probability $113/1461$.

2.4 Since there are 7 long months and 8 months with an "r" in their name, we have $P(L) = 7/12$ and $P(R) = 8/12$.

2.5 Checking that P is a probability function Ω amounts to verifying that $0 \le P((a_i, a_j)) \le 1$ for all i and j and noting that

$$P(\Omega) = \sum_{i,j=1}^{6} P((a_i, a_j)) = \sum_{i=1}^{6} P((a_i, a_i)) = \sum_{i=1}^{6} p_i = 1.$$

The two experiments are *totally* coupled: one has outcome a_i if and only if the other has outcome a_i.

2.6 Now there are 10 outcomes in B (for example $(0,1,0,1,0)$), each having probability $(1-p)^3 p^2$. Hence $P(B) = 10(1-p)^3 p^2$.

2.7 This happens if and only if the experiment fails on Monday,..., Saturday, and is a success on Sunday. This has probability $p(1-p)^6$ to happen.

2.7 Exercises

2.1 ⊡ Let A and B be two events in a sample space for which $P(A) = 2/3$, $P(B) = 1/6$, and $P(A \cap B) = 1/9$. What is $P(A \cup B)$?

2.2 Let E and F be two events for which one knows that the probability that at least one of them occurs is $3/4$. What is the probability that neither E nor F occurs? *Hint:* use one of DeMorgan's laws: $E^c \cap F^c = (E \cup F)^c$.

2.3 Let C and D be two events for which one knows that $P(C) = 0.3, P(D) = 0.4$, and $P(C \cap D) = 0.2$. What is $P(C^c \cap D)$?

2.4 ⊡ We consider events A, B, and C, which can occur in some experiment. Is it true that the probability that *only* A occurs (and not B or C) is equal to $P(A \cup B \cup C) - P(B) - P(C) + P(B \cap C)$?

2.5 The event $A \cap B^c$ that A occurs but not B is sometimes denoted as $A \setminus B$. Here $\setminus$ is the set-theoretic minus sign. Show that $P(A \setminus B) = P(A) - P(B)$ if B implies A, i.e., if $B \subset A$.

2.6 When $P(A) = 1/3$, $P(B) = 1/2$, and $P(A \cup B) = 3/4$, what is

a. $P(A \cap B)$?
b. $P(A^c \cup B^c)$?

2.7 ⊡ Let A and B be two events. Suppose that $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \cap B) = 0.1$. Find the probability that A or B occurs, but not both.

2.8 ⊞ Suppose the events D_1 and D_2 represent disasters, which are rare: $P(D_1) \leq 10^{-6}$ *and* $P(D_2) \leq 10^{-6}$. What can you say about the probability that at least one of the disasters occurs? What about the probability that they *both* occur?

2.9 We toss a coin three times. For this experiment we choose the sample space

$$\Omega = \{HHH, THH, HTH, HHT, TTH, THT, HTT, TTT\}$$

where T stands for tails and H for heads.

a. Write down the set of outcomes corresponding to each of the following events:

A : "we throw tails exactly two times."
B : "we throw tails at least two times."
C : "tails did not appear *before* a head appeared."
D : "the first throw results in tails."

b. Write down the set of outcomes corresponding to each of the following events: A^c, $A \cup (C \cap D)$, and $A \cap D^c$.

2.10 In some sample space we consider two events A and B. Let C be the event that A or B occurs, but not both. Express C in terms of A and B, using only the basic operations "union," "intersection," and "complement."

2.11 ⊡ An experiment has only two outcomes. The first has probability p to occur, the second probability p^2. What is p?

2.12 ⊞ In the UEFA Euro 2004 playoffs draw 10 national football teams were matched in pairs. A lot of people complained that "the draw was not fair," because each strong team had been matched with a weak team (this is commercially the most interesting). It was claimed that such a matching is extremely unlikely. We will compute the probability of this "dream draw" in this exercise. In the spirit of the three-envelope example of Section 2.1 we put the names of the 5 strong teams in envelopes labeled $1, 2, 3, 4$, and 5 and of the 5 weak teams in envelopes labeled $6, 7, 8, 9$, and 10. We shuffle the 10 envelopes and then match the envelope on top with the next envelope, the third envelope with the fourth envelope, and so on. One particular way a "dream draw" occurs is when the five envelopes labeled $1, 2, 3, 4, 5$ are in the odd numbered positions (in any order!) and the others are in the even numbered positions. This way corresponds to the situation where the first match of each strong team is a home match. Since for each pair there are two possibilities for the home match, the total number of possibilities for the "dream draw" is $2^5 = 32$ times as large.

a. An outcome of this experiment is a sequence like $4, 9, 3, 7, 5, 10, 1, 8, 2, 6$ of labels of envelopes. What is the probability of an outcome?

b. How many outcomes are there in the event "the five envelopes labeled $1, 2, 3, 4, 5$ are in the odd positions—in any order, and the envelopes labeled $6, 7, 8, 9, 10$ are in the even positions—in any order"?

c. What is the probability of a "dream draw"?

2.13 In some experiment first an arbitrary choice is made out of four possibilities, and then an arbitrary choice is made out of the remaining three possibilities. One way to describe this is with a product of two sample spaces $\{a, b, c, d\}$:

$$\Omega = \{a, b, c, d\} \times \{a, b, c, d\}.$$

a. Make a 4×4 table in which you write the probabilities of the outcomes.

b. Describe the event "c is one of the chosen possibilities" and determine its probability.

2.14 ⊞ Consider the Monty Hall "experiment" described in Section 1.3. The door behind which the car is parked we label a, the other two b and c. As the sample space we choose a product space

$$\Omega = \{a, b, c\} \times \{a, b, c\}.$$

Here the first entry gives the choice of the candidate, and the second entry the choice of the quizmaster.

a. Make a 3×3 table in which you write the probabilities of the outcomes. *N.B.* You should realize that the candidate does *not know* that the car is in a, but the quizmaster will never open the door labeled a because he *knows* that the car is there. You may assume that the quizmaster makes an arbitrary choice between the doors labeled b and c, when the candidate chooses door a.

b. Consider the situation of a "no switching" candidate who will stick to his or her choice. What is the event "the candidate wins the car," and what is its probability?

c. Consider the situation of a "switching" candidate who will not stick to her choice. What is now the event "the candidate wins the car," and what is its probability?

2.15 The rule $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ from Section 2.3 is often useful to compute the probability of the union of two events. What would be the corresponding rule for three events A, B, and C? It should start with

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

Hint: you could use the sum rule suitably, or you could make a diagram as in Figure 2.1.

2.16 ⊞ Three events E, F, and G cannot occur simultaneously. Further it is known that $P(E \cap F) = P(F \cap G) = P(E \cap G) = 1/3$. Can you determine $P(E)$?
Hint: if you try to use the formula of Exercise 2.15 then it seems that you do not have enough information; make a diagram instead.

2.17 A post office has two counters where customers can buy stamps, etc. If you are interested in the number of customers in the two queues that will form for the counters, what would you take as sample space?

2.18 In a laboratory, two experiments are repeated every day of the week in different rooms until at least one is successful, the probability of success being p for each experiment. Supposing that the experiments in different rooms and on different days are performed independently of each other, what is the probability that the laboratory scores its first successful experiment on day n?

2.19 ⊡ We repeatedly toss a coin. A head has probability p, and a tail probability $1 - p$ to occur, where $0 < p < 1$. The outcome of the experiment we are interested in is the number of tosses it takes until a head occurs for the *second* time.

a. What would you choose as the sample space?
b. What is the probability that it takes 5 tosses?

3

Conditional probability and independence

Knowing that an event has occurred sometimes forces us to reassess the probability of another event; the new probability is the *conditional* probability. If the conditional probability equals what the probability was before, the events involved are called *independent*. Often, conditional probabilities and independence are needed if we want to compute probabilities, and in many other situations they simplify the work.

3.1 Conditional probability

In the previous chapter we encountered the events L, "born in a long month," and R, "born in a month with the letter r." Their probabilities are easy to compute: since $L = \{$Jan, Mar, May, Jul, Aug, Oct, Dec$\}$ and $R = \{$Jan, Feb, Mar, Apr, Sep, Oct, Nov, Dec$\}$, one finds

$$\mathrm{P}(L) = \frac{7}{12} \quad \text{and} \quad \mathrm{P}(R) = \frac{8}{12}.$$

Now suppose that it is *known* about the person we meet in the street that he was born in a "long month," and we wonder whether he was born in a "month with the letter r." The information given excludes five outcomes of our sample space: it cannot be February, April, June, September, or November. Seven possible outcomes are left, of which only four—those in $R \cap L = \{$Jan, Mar, Oct, Dec$\}$—are favorable, so we reassess the probability as 4/7. We call this the *conditional probability of R given L*, and we write:

$$\mathrm{P}(R \,|\, L) = \frac{4}{7}.$$

This is not the same as $\mathrm{P}(R \cap L)$, which is 1/3. Also note that $\mathrm{P}(R \,|\, L)$ is the proportion that $\mathrm{P}(R \cap L)$ is of $\mathrm{P}(L)$.

QUICK EXERCISE 3.1 Let $N = R^c$ be the event "born in a month without r." What is the conditional probability $P(N \mid L)$?

Recalling the three envelopes on our doormat, consider the events "envelope 1 is the middle one" (call this event A) and "envelope 2 is the middle one" (B). Then $P(A) = P(213 \text{ or } 312) = 1/3$; by symmetry, the same is found for $P(B)$. We say that the envelopes are in order if their order is either 123 or 321. Suppose we know that they are *not* in order, but otherwise we do not know anything; what are the probabilities of A and B, given this information?

Let C be the event that the envelopes are not in order, so: $C = \{123, 321\}^c = \{132, 213, 231, 312\}$. We ask for the probabilities of A and B, given that C occurs. Event C consists of four elements, two of which also belong to A: $A \cap C = \{213, 312\}$, so $P(A \mid C) = 1/2$. The probability of $A \cap C$ is half of $P(C)$. No element of C also belongs to B, so $P(B \mid C) = 0$.

QUICK EXERCISE 3.2 Calculate $P(C \mid A)$ and $P(C^c \mid A \cup B)$.

In general, computing the probability of an event A, given that an event C occurs, means finding which fraction of the probability of C is also in the event A.

DEFINITION. The *conditional probability of A given C* is given by:

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)},$$

provided $P(C) > 0$.

QUICK EXERCISE 3.3 Show that $P(A \mid C) + P(A^c \mid C) = 1$.

This exercise shows that the rule $P(A^c) = 1 - P(A)$ also holds for conditional probabilities. In fact, even more is true: if we have a fixed conditioning event C and define $Q(A) = P(A \mid C)$ for events $A \subset \Omega$, then Q is a probability function and hence satisfies all the rules as described in Chapter 2. The definition of conditional probability agrees with our intuition and it also works in situations where computing probabilities by counting outcomes does not.

A chemical reactor: residence times

Consider a continuously stirred reactor vessel where a chemical reaction takes place. On one side fluid or gas flows in, mixes with whatever is already present in the vessel, and eventually flows out on the other side. One characteristic of each particular reaction setup is the so-called residence time distribution, which tells us how long particles stay inside the vessel before moving on. We consider a continuously stirred tank: the contents of the vessel are perfectly mixed at all times.

Let R_t denote the event "the particle has a residence time longer than t seconds." In Section 5.3 we will see how continuous stirring determines the probabilities; here we just use that in a particular continuously stirred tank, R_t has probability e^{-t}. So:

$$P(R_3) = e^{-3} = 0.04978\ldots$$
$$P(R_4) = e^{-4} = 0.01831\ldots.$$

We can use the definition of conditional probability to find the probability that a particle that has stayed more than 3 seconds will stay more than 4:

$$P(R_4 \mid R_3) = \frac{P(R_4 \cap R_3)}{P(R_3)} = \frac{P(R_4)}{P(R_3)} = \frac{e^{-4}}{e^{-3}} = e^{-1} = 0.36787\ldots.$$

QUICK EXERCISE 3.4 Calculate $P(R_3 \mid R_4^c)$.

For more details on the subject of residence time distributions see, for example, the book on reaction engineering by Fogler ([11]).

3.2 The multiplication rule

From the definition of conditional probability we derive a useful rule by multiplying left and right by $P(C)$.

THE MULTIPLICATION RULE. For any events A and C:

$$P(A \cap C) = P(A \mid C) \cdot P(C).$$

Computing the probability of $A \cap C$ can hence be decomposed into two parts, computing $P(C)$ and $P(A \mid C)$ separately, which is often easier than computing $P(A \cap C)$ directly.

The probability of no coincident birthdays

Suppose you meet two arbitrarily chosen people. What is the probability their birthdays are different? Let B_2 denote the event that this happens. Whatever the birthday of the first person is, there is only one day the second person cannot "pick" as birthday, so:

$$P(B_2) = 1 - \frac{1}{365}.$$

When the same question is asked with *three* people, conditional probabilities become helpful. The event B_3 can be seen as the intersection of the event B_2,

"the first two have different birthdays," with event A_3 "the third person has a birthday that does not coincide with that of one of the first two persons." Using the multiplication rule:

$$P(B_3) = P(A_3 \cap B_2) = P(A_3 \mid B_2)P(B_2).$$

The conditional probability $P(A_3 \mid B_2)$ is the probability that, when two days are already marked on the calendar, a day picked at random is not marked, or

$$P(A_3 \mid B_2) = 1 - \frac{2}{365},$$

and so

$$P(B_3) = P(A_3 \mid B_2)P(B_2) = \left(1 - \frac{2}{365}\right) \cdot \left(1 - \frac{1}{365}\right) = 0.9918.$$

We are already halfway to solving the general question: in a group of n arbitrarily chosen people, what is the probability there are no coincident birthdays? The event B_n of no coincident birthdays among the n persons is the same as: "the birthdays of the first $n-1$ persons are different" (the event B_{n-1}) and "the birthday of the nth person does not coincide with a birthday of any of the first $n-1$ persons" (the event A_n), that is,

$$B_n = A_n \cap B_{n-1}.$$

Applying the multiplication rule yields:

$$P(B_n) = P(A_n \mid B_{n-1}) \cdot P(B_{n-1}) = \left(1 - \frac{n-1}{365}\right) \cdot P(B_{n-1})$$

as person n should avoid $n-1$ days. Applying the same step to $P(B_{n-1})$, $P(B_{n-2})$, etc., we find:

$$
\begin{aligned}
P(B_n) &= \left(1 - \frac{n-1}{365}\right) \cdot P(A_{n-1} \mid B_{n-2}) \cdot P(B_{n-2}) \\
&= \left(1 - \frac{n-1}{365}\right) \cdot \left(1 - \frac{n-2}{365}\right) \cdot P(B_{n-2}) \\
&\quad\ \vdots \\
&= \left(1 - \frac{n-1}{365}\right) \cdots \left(1 - \frac{2}{365}\right) \cdot P(B_2) \\
&= \left(1 - \frac{n-1}{365}\right) \cdots \left(1 - \frac{2}{365}\right) \cdot \left(1 - \frac{1}{365}\right).
\end{aligned}
$$

This can be used to compute the probability for arbitrary n. For example, we find: $P(B_{22}) = 0.5243$ and $P(B_{23}) = 0.4927$. In Figure 3.1 the probability
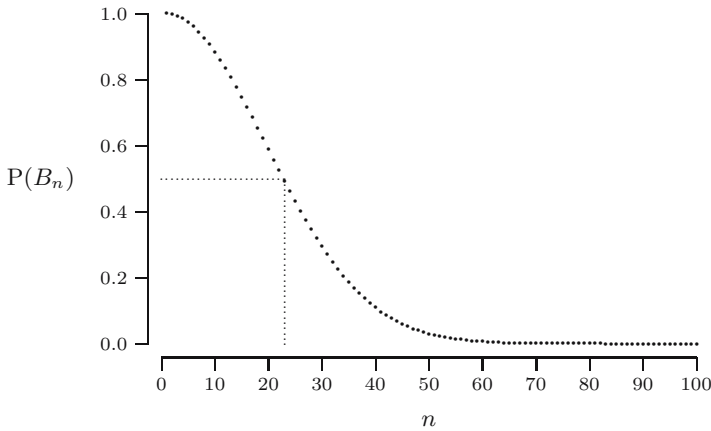
Fig. 3.1. The probability $P(B_n)$ of no coincident birthdays for $n = 1, \ldots, 100$.

$P(B_n)$ is plotted for $n = 1, \ldots, 100$, with dotted lines drawn at $n = 23$ and at probability 0.5. It may be hard to believe, but with just 23 people the probability of all birthdays being different is less than 50%!

QUICK EXERCISE 3.5 Compute the probability that three arbitrary people are born in different months. Can you give the formula for n people?

It matters how one conditions

Conditioning can help to make computations easier, but it matters how it is applied. To compute $P(A \cap C)$ we may condition on C to get

$$P(A \cap C) = P(A \,|\, C) \cdot P(C) \,;$$

or we may condition on A and get

$$P(A \cap C) = P(C \,|\, A) \cdot P(A) \,.$$

Both ways are valid, but often *one* of $P(A \,|\, C)$ and $P(C \,|\, A)$ is easy and the other is not. For example, in the birthday example one could have tried:

$$P(B_3) = P(A_3 \cap B_2) = P(B_2 \,|\, A_3)P(A_3) \,,$$

but just trying to understand the conditional probability $P(B_2 \,|\, A_3)$ already is confusing:

> The probability that the first two persons' birthdays differ given that the third person's birthday does not coincide with the birthday of one of the first two …?

Conditioning should lead to easier probabilities; if not, it is probably the wrong approach.

3.3 The law of total probability and Bayes' rule

We will now discuss two important rules that help probability computations by means of conditional probabilities. We introduce both of them in the next example.

Testing for mad cow disease

In early 2001 the European Commission introduced massive testing of cattle to determine infection with the transmissible form of *Bovine Spongiform Encephalopathy* (BSE) or "mad cow disease." As no test is 100% accurate, most tests have the problem of false positives and false negatives. A *false positive* means that according to the test the cow is infected, but in actuality it is not. A *false negative* means an infected cow is not detected by the test.

Imagine we test a cow. Let B denote the event "the cow has BSE" and T the event "the test comes up positive" (this is test jargon for: according to the test we should believe the cow is infected with BSE). One can "test the test" by analyzing samples from cows that are known to be infected or known to be healthy and so determine the effectiveness of the test. The European Commission had this done for four tests in 1999 (see [19]) and for several more later. The results for what the report calls Test A may be summarized as follows: an infected cow has a 70% chance of testing positive, and a healthy cow just 10%; in formulas:

$$P(T \mid B) = 0.70,$$
$$P(T \mid B^c) = 0.10.$$

Suppose we want to determine the probability $P(T)$ that an arbitrary cow tests positive. The tested cow is either infected or it is not: event T occurs in combination with B or with B^c (there are no other possibilities). In terms of events

$$T = (T \cap B) \cup (T \cap B^c),$$

so that

$$P(T) = P(T \cap B) + P(T \cap B^c),$$

because $T \cap B$ and $T \cap B^c$ are disjoint. Next, apply the multiplication rule (in such a way that the known conditional probabilities appear!):

$$P(T \cap B) = P(T \mid B) \cdot P(B)$$
$$P(T \cap B^c) = P(T \mid B^c) \cdot P(B^c) \tag{3.1}$$

so that

$$P(T) = P(T \mid B) \cdot P(B) + P(T \mid B^c) \cdot P(B^c). \tag{3.2}$$

This is an application of the law of total probability: computing a probability through conditioning on several disjoint events that make up the whole sample

space (in this case two). Suppose[1] $P(B) = 0.02$; then from the last equation we conclude: $P(T) = 0.02 \cdot 0.70 + (1 - 0.02) \cdot 0.10 = 0.112$.

QUICK EXERCISE 3.6 Calculate $P(T)$ when $P(T \mid B) = 0.99$ and $P(T \mid B^c) = 0.05$.

Following is a general statement of the law.

> THE LAW OF TOTAL PROBABILITY. Suppose C_1, C_2, ..., C_m are disjoint events such that $C_1 \cup C_2 \cup \cdots \cup C_m = \Omega$. The probability of an arbitrary event A can be expressed as:
>
> $$P(A) = P(A \mid C_1)P(C_1) + P(A \mid C_2)P(C_2) + \cdots + P(A \mid C_m)P(C_m).$$

Figure 3.2 illustrates the law for $m = 5$. The event A is the disjoint union of $A \cap C_i$, for $i = 1, \ldots, 5$, so $P(A) = P(A \cap C_1) + \cdots + P(A \cap C_5),$ and for each i the multiplication rule states $P(A \cap C_i) = P(A \mid C_i) \cdot P(C_i)$.
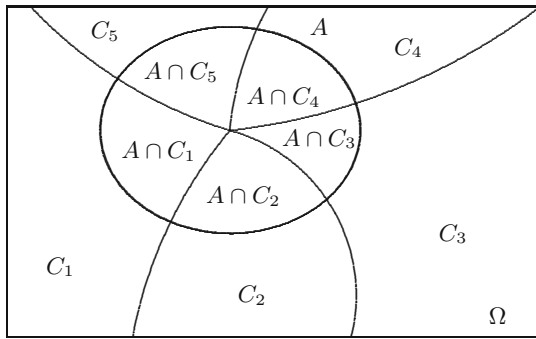
Fig. 3.2. The law of total probability (illustration for $m = 5$).

In the BSE example, we have just two mutually exclusive events: substitute $m = 2$, $C_1 = B$, $C_2 = B^c$, and $A = T$ to obtain (3.2).

Another, perhaps more pertinent, question about the BSE test is the following: suppose my cow tests positive; what is the probability it really has BSE? Translated, this asks for the value of $P(B \mid T)$. The information we were given is $P(T \mid B)$, a conditional probability, but the wrong one. We would like to switch T and B.

Start with the definition of conditional probability and then use equations (3.1) and (3.2):

---

[1] We choose this probability for the sake of the calculations that follow. The true value is unknown and varies from country to country. The BSE risk for the Netherlands for 2003 was estimated to be $P(B) \approx 0.000013$.

$$P(B \mid T) = \frac{P(T \cap B)}{P(T)} = \frac{P(T \mid B) \cdot P(B)}{P(T \mid B) \cdot P(B) + P(T \mid B^c) \cdot P(B^c)}.$$

So with $P(B) = 0.02$ we find

$$P(B \mid T) = \frac{0.70 \cdot 0.02}{0.70 \cdot 0.02 + 0.10 \cdot (1 - 0.02)} = 0.125,$$

and by a similar calculation: $P(B \mid T^c) = 0.0068$. These probabilities reflect that this Test A is not a very good test; a perfect test would result in $P(B \mid T) = 1$ and $P(B \mid T^c) = 0$. In Exercise 3.4 we redo this calculation, replacing $P(B) = 0.02$ with a more realistic number.

What we have just seen is known as Bayes' rule, after the English clergyman Thomas Bayes who derived this in the 18th century. The general statement follows.

> BAYES' RULE. Suppose the events C_1, C_2, ..., C_m are disjoint and $C_1 \cup C_2 \cup \cdots \cup C_m = \Omega$. The conditional probability of C_i, given an arbitrary event A, can be expressed as:
>
> $$P(C_i \mid A) = \frac{P(A \mid C_i) \cdot P(C_i)}{P(A \mid C_1)P(C_1) + P(A \mid C_2)P(C_2) + \cdots + P(A \mid C_m)P(C_m)}.$$

This is the traditional form of Bayes' formula. It follows from

$$P(C_i \mid A) = \frac{P(A \mid C_i) \cdot P(C_i)}{P(A)} \tag{3.3}$$

in combination with the law of total probability applied to $P(A)$ in the denominator. Purists would refer to (3.3) as Bayes' rule, and perhaps they are right.

QUICK EXERCISE 3.7 Calculate $P(B \mid T)$ and $P(B \mid T^c)$ if $P(T \mid B) = 0.99$ and $P(T \mid B^c) = 0.05$.

3.4 Independence

Consider three probabilities from the previous section:

$$P(B) = 0.02,$$
$$P(B \mid T) = 0.125,$$
$$P(B \mid T^c) = 0.0068.$$

If we know nothing about a cow, we would say that there is a 2% chance it is infected. However, if we know it tested positive, we can say there is a 12.5%

chance the cow is infected. On the other hand, if it tested negative, there is only a 0.68% chance. We see that the two events are related in some way: the probability of B *depends* on whether T occurs.

Imagine the opposite: the test is useless. Whether the cow is infected is unrelated to the outcome of the test, and knowing the outcome of the test does not change our probability of B: $P(B\,|\,T) = P(B)$. In this case we would call B independent of T.

> DEFINITION.   An event A is called *independent of B* if
>
> $$P(A\,|\,B) = P(A)\,.$$

From this simple definition many statements can be derived. For example, because $P(A^c\,|\,B) = 1 - P(A\,|\,B)$ and $1 - P(A) = P(A^c)$, we conclude:

$$A \text{ independent of } B \quad \Leftrightarrow \quad A^c \text{ independent of } B. \qquad (3.4)$$

By application of the multiplication rule, if A is independent of B, then $P(A \cap B) = P(A\,|\,B)P(B) = P(A)\,P(B)$. On the other hand, if $P(A \cap B) = P(A)\,P(B)$, then $P(A\,|\,B) = P(A)$ follows from the definition of independence. This shows:

$$A \text{ independent of } B \quad \Leftrightarrow \quad P(A \cap B) = P(A)\,P(B)\,.$$

Finally, by definition of conditional probability, if A is independent of B, then

$$P(B\,|\,A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A) \cdot P(B)}{P(A)} = P(B)\,,$$

that is, B is independent of A. This works in reverse, too, so we have:

$$A \text{ independent of } B \quad \Leftrightarrow \quad B \text{ independent of } A. \qquad (3.5)$$

This statement says that in fact, independence is *a mutual property*. Therefore, the expressions "A is independent of B" and "A and B are independent" are used interchangeably. From the three $\Leftrightarrow$-statements it follows that there are in fact 12 ways to show that A and B are independent; and if they are, there are 12 ways to use that.

> INDEPENDENCE.   To show that A and B are independent it suffices to prove *just one* of the following:
>
> $$P(A\,|\,B) = P(A)\,,$$
> $$P(B\,|\,A) = P(B)\,,$$
> $$P(A \cap B) = P(A)\,P(B)\,,$$
>
> where A may be replaced by A^c and B replaced by B^c, or both. If one of these statements holds, *all* of them are true. If two events are not independent, they are called *dependent*.

Recall the birthday events L "born in a long month" and R "born in a month with the letter r." Let H be the event "born in the first half of the year," so $P(H) = 1/2$. Also, $P(H\,|\,R) = 1/2$. So H and R are independent, and we conclude, for example, $P(R^c\,|\,H^c) = P(R^c) = 1 - 8/12 = 1/3$.

We know that $P(L \cap H) = 1/4$ and $P(L) = 7/12$. Checking $1/2 \times 7/12 \neq 1/4$, you conclude that L and H are dependent.

QUICK EXERCISE 3.8 Derive the statement "R^c is independent of H^c" from "H is independent of R" using rules (3.4) and (3.5).

Since the words dependence and independence have several meanings, one sometimes uses the terms *stochastic* or *statistical* dependence and independence to avoid ambiguity.

> **Remark 3.1 (Physical and stochastic independence).** Stochastic dependence or independence can sometimes be established by inspecting whether there is any physical dependence present. The following statements may be made.
> If events have to do with processes or experiments that have no physical connection, they are always stochastically independent. If they are connected to the same physical process, then, as a rule, they are stochastically dependent, but stochastic independence is possible in exceptional cases. The events H and R are an example.

Independence of two or more events

When more than two events are involved we need a more elaborate definition of independence. The reason behind this is explained by an example following the definition.

> INDEPENDENCE OF TWO OR MORE EVENTS. Events A_1, A_2, ..., A_m are called independent if
>
> $$P(A_1 \cap A_2 \cap \cdots \cap A_m) = P(A_1)\,P(A_2)\cdots P(A_m)$$
>
> *and* this statement *also* holds when any number of the events A_1, ..., A_m are replaced by their complements throughout the formula.

You see that we need to check 2^m equations to establish the independence of m events. In fact, $m + 1$ of those equations are redundant, but we chose this version of the definition because it is easier.

The reason we need to do so much more checking to establish independence for multiple events is that there are subtle ways in which events may depend on each other. Consider the question:

> Is independence for three events A, B, and C the same as: A and B are independent; B and C are independent; and A and C are independent?

The answer is "No," as the following example shows. Perform two independent tosses of a coin. Let A be the event "heads on toss 1," B the event "heads on toss 2," and C "the two tosses are equal."

First, get the probabilities. Of course, $P(A) = P(B) = 1/2$, but also

$$P(C) = P(A \cap B) + P(A^c \cap B^c) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

What about independence? Events A and B are independent by assumption, so check the independence of A and C. Given that the first toss is heads (A occurs), C occurs if and only if the second toss is heads as well (B occurs), so

$$P(C \mid A) = P(B \mid A) = P(B) = \frac{1}{2} = P(C).$$

By symmetry, also $P(C \mid B) = P(C)$, so all pairs taken from A, B, C are independent: the three are called *pairwise independent*. Checking the full conditions for independence, we find, for example:

$$P(A \cap B \cap C) = P(A \cap B) = \frac{1}{4}, \quad \text{whereas} \quad P(A)\, P(B)\, P(C) = \frac{1}{8},$$

and

$$P(A \cap B \cap C^c) = P(\emptyset) = 0, \quad \text{whereas} \quad P(A)\, P(B)\, P(C^c) = \frac{1}{8}.$$

The reason for this is clear: whether C occurs follows deterministically from the outcomes of tosses 1 and 2.

3.5 Solutions to the quick exercises

3.1 $N = \{\text{May, Jun, Jul, Aug}\}$, $L = \{\text{Jan, Mar, May, Jul, Aug, Oct, Dec}\}$, and $N \cap L = \{\text{May, Jul, Aug}\}$. Three out of seven outcomes of L belong to N as well, so $P(N \mid L) = 3/7$.

3.2 The event A is contained in C. So when A occurs, C also occurs; therefore $P(C \mid A) = 1$.
Since $C^c = \{123, 321\}$ and $A \cup B = \{123, 321, 312, 213\}$, one can see that two of the four outcomes of $A \cup B$ belong to C^c as well, so $P(C^c \mid A \cup B) = 1/2$.

3.3 Using the definition we find:

$$P(A \mid C) + P(A^c \mid C) = \frac{P(A \cap C)}{P(C)} + \frac{P(A^c \cap C)}{P(C)} = 1,$$

because C can be split into disjoint parts $A \cap C$ and $A^c \cap C$ and therefore

$$P(A \cap C) + P(A^c \cap C) = P(C).$$

3.4 This asks for the probability that the particle stays more than 3 seconds, given that it does not stay longer than 4 seconds, so 4 or less. From the definition:

$$P(R_3 \mid R_4^c) = \frac{P(R_3 \cap R_4^c)}{P(R_4^c)}.$$

The event $R_3 \cap R_4^c$ describes: longer than 3 but not longer than 4 seconds. Furthermore, R_3 is the disjoint union of the events $R_3 \cap R_4^c$ and $R_3 \cap R_4 = R_4$, so $P(R_3 \cap R_4^c) = P(R_3) - P(R_4) = e^{-3} - e^{-4}$. Using the complement rule: $P(R_4^c) = 1 - P(R_4) = 1 - e^{-4}$. Together:

$$P(R_3 \mid R_4^c) = \frac{e^{-3} - e^{-4}}{1 - e^{-4}} = \frac{0.0315}{0.9817} = 0.0321.$$

3.5 Instead of a calendar of 365 days, we have one with just 12 months. Let C_n be the event n arbitrary persons have different months of birth. Then

$$P(C_3) = \left(1 - \frac{2}{12}\right) \cdot \left(1 - \frac{1}{12}\right) = \frac{55}{72} = 0.7639$$

and it is no surprise that this is much smaller than $P(B_3)$. The general formula is

$$P(C_n) = \left(1 - \frac{n-1}{12}\right) \cdots \left(1 - \frac{2}{12}\right) \cdot \left(1 - \frac{1}{12}\right).$$

Note that it is correct even if n is 13 or more, in which case $P(C_n) = 0$.

3.6 Repeating the calculation we find:

$$P(T \cap B) = 0.99 \cdot 0.02 = 0.0198$$
$$P(T \cap B^c) = 0.05 \cdot 0.98 = 0.0490$$

so $P(T) = P(T \cap B) + P(T \cap B^c) = 0.0198 + 0.0490 = 0.0688.$

3.7 In the solution to Quick exercise 3.5 we already found $P(T \cap B) = 0.0198$ and $P(T) = 0.0688$, so

$$P(B \mid T) = \frac{P(T \cap B)}{P(T)} = \frac{0.0198}{0.0688} = 0.2878.$$

Further, $P(T^c) = 1 - 0.0688 = 0.9312$ and $P(T^c \mid B) = 1 - P(T \mid B) = 0.01$. So, $P(B \cap T^c) = 0.01 \cdot 0.02 = 0.0002$ and

$$P(B \mid T^c) = \frac{0.0002}{0.9312} = 0.00021.$$

3.8 It takes three steps of applying (3.4) and (3.5):

$$
\begin{array}{lll}
H \text{ independent of } R & \Leftrightarrow & H^c \text{ independent of } R \;\; \text{by (3.4)} \\
H^c \text{ independent of } R & \Leftrightarrow & R \text{ independent of } H^c \;\; \text{by (3.5)} \\
R \text{ independent of } H^c & \Leftrightarrow & R^c \text{ independent of } H^c \;\; \text{by (3.4)}.
\end{array}
$$

3.6 Exercises

3.1 ⊞ Your lecturer wants to walk from A to B (see the map). To do so, he first randomly selects one of the paths to C, D, or E. Next he selects randomly one of the possible paths at that moment (so if he first selected the path to E, he can either select the path to A or the path to F), etc. What is the probability that he will reach B after two selections?
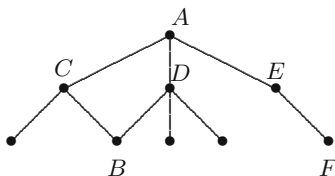
3.2 ⊞ A fair die is thrown twice. A is the event "sum of the throws equals 4," B is "at least one of the throws is a 3."

  a. Calculate $P(A \mid B)$.
  b. Are A and B independent events?

3.3 ⊞ We draw two cards from a regular deck of 52. Let S_1 be the event "the first one is a spade," and S_2 "the second one is a spade."

  a. Compute $P(S_1)$, $P(S_2 \mid S_1)$, and $P(S_2 \mid S_1^c)$.
  b. Compute $P(S_2)$ by conditioning on whether the first card is a spade.

3.4 ⊡ A Dutch cow is tested for BSE, using Test A as described in Section 3.3, with $P(T \mid B) = 0.70$ and $P(T \mid B^c) = 0.10$. Assume that the BSE risk for the Netherlands is the same as in 2003, when it was estimated to be $P(B) = 1.3 \cdot 10^{-5}$. Compute $P(B \mid T)$ and $P(B \mid T^c)$.

3.5 A ball is drawn at random from an urn containing one red and one white ball. If the white ball is drawn, it is put back into the urn. If the red ball is drawn, it is returned to the urn together with two more red balls. Then a second draw is made. What is the probability a red ball was drawn on *both* the first and the second draws?

3.6 We choose a month of the year, in such a manner that each month has the same probability. Find out whether the following events are independent:

  a. the events "outcome is an even numbered month" (i.e., February, April, June, etc.) and "outcome is in the first half of the year."
  b. the events "outcome is an even numbered month" (i.e., February, April, June, etc.) and "outcome is a summer month" (i.e., June, July, August).

3.7 ⊞ Calculate

a. $P(A \cup B)$ if it is given that $P(A) = 1/3$ and $P(B \mid A^c) = 1/4$.

b. $P(B)$ if it is given that $P(A \cup B) = 2/3$ and $P(A^c \mid B^c) = 1/2$.

3.8 ⊞ Spaceman Spiff's spacecraft has a warning light that is supposed to switch on when the freem blasters are overheated. Let W be the event "the warning light is switched on" and F "the freem blasters are overheated." Suppose the probability of freem blaster overheating $P(F)$ is 0.1, that the light is switched on when they actually *are* overheated is 0.99, and that there is a 2% chance that it comes on when nothing is wrong: $P(W \mid F^c) = 0.02$.

a. Determine the probability that the warning light is switched on.

b. Determine the conditional probability that the freem blasters are overheated, given that the warning light is on.

3.9 ⊟ A certain grapefruit variety is grown in two regions in southern Spain. Both areas get infested from time to time with parasites that damage the crop. Let A be the event that region R_1 is infested with parasites and B that region R_2 is infested. Suppose $P(A) = 3/4$, $P(B) = 2/5$ and $P(A \cup B) = 4/5$. If the food inspection detects the parasite in a ship carrying grapefruits from R_1, what is the probability region R_2 is infested as well?

3.10 A student takes a multiple-choice exam. Suppose for each question he either knows the answer or gambles and chooses an option at random. Further suppose that if he knows the answer, the probability of a correct answer is 1, and if he gambles this probability is 1/4. To pass, students need to answer at least 60% of the questions correctly. The student has "studied for a minimal pass," i.e., with probability 0.6 he knows the answer to a question. Given that he answers a question correctly, what is the probability that he actually *knows* the answer?

3.11 A breath analyzer, used by the police to test whether drivers exceed the legal limit set for the blood alcohol percentage while driving, is known to satisfy
$$P(A \mid B) = P(A^c \mid B^c) = p,$$
where A is the event "breath analyzer indicates that legal limit is exceeded" and B "driver's blood alcohol percentage exceeds legal limit." On Saturday night about 5% of the drivers are known to exceed the limit.

a. Describe in words the meaning of $P(B^c \mid A)$.

b. Determine $P(B^c \mid A)$ if $p = 0.95$.

c. How big should p be so that $P(B \mid A) = 0.9$?

3.12 The events A, B, and C satisfy: $P(A \mid B \cap C) = 1/4$, $P(B \mid C) = 1/3$, and $P(C) = 1/2$. Calculate $P(A^c \cap B \cap C)$.

3.13 In Exercise 2.12 we computed the probability of a "dream draw" in the UEFA playoffs lottery by counting outcomes. Recall that there were ten teams in the lottery, five considered "strong" and five considered "weak." Introduce events D_i, "the ith pair drawn is a dream combination," where a "dream combination" is a pair of a strong team with a weak team, and $i = 1, \ldots, 5$.

a. Compute $P(D_1)$.
b. Compute $P(D_2 \mid D_1)$ and $P(D_1 \cap D_2)$.
c. Compute $P(D_3 \mid D_1 \cap D_2)$ and $P(D_1 \cap D_2 \cap D_3)$.
d. Continue the procedure to obtain the probability of a "dream draw": $P(D_1 \cap \cdots \cap D_5)$.

3.14 Recall the Monty Hall problem from Section 1.3. Let R be the event "the prize is behind the door you chose initially," and W the event "you win the prize by switching doors."

a. Compute $P(W \mid R)$ and $P(W \mid R^c)$.
b. Compute $P(W)$ using the law of total probability.

3.15 Two independent events A and B are given, and $P(B \mid A \cup B) = 2/3$, $P(A \mid B) = 1/2$. What is $P(B)$?

3.16 You are diagnosed with an uncommon disease. You know that there only is a 1% chance of getting it. Use the letter D for the event "you have the disease" and T for "the test says so." It is known that the test is imperfect: $P(T \mid D) = 0.98$ and $P(T^c \mid D^c) = 0.95$.

a. Given that you test positive, what is the probability that you really *have* the disease?
b. You obtain a second opinion: an independent repetition of the test. You test positive again. Given this, what is the probability that you really *have* the disease?

3.17 You and I play a tennis match. It is deuce, which means if you win the next two rallies, you win the game; if I win both rallies, I win the game; if we each win one rally, it is deuce again. Suppose the outcome of a rally is independent of other rallies, and you win a rally with probability p. Let W be the event "you win the game," G "the game ends after the next two rallies," and D "it becomes deuce again."

a. Determine $P(W \mid G)$.
b. Show that $P(W) = p^2 + 2p(1 - p)P(W \mid D)$ and use $P(W) = P(W \mid D)$ (why is this so?) to determine $P(W)$.
c. Explain why the answers are the same.

3.18 Suppose A and B are events with $0 < P(A) < 1$ and $0 < P(B) < 1$.

a. If A and B are disjoint, can they be independent?

b. If A and B are independent, can they be disjoint?

c. If $A \subset B$, can A and B be independent?

d. If A and B are independent, can A and $A \cup B$ be independent?

4

Discrete random variables

The sample space associated with an experiment, together with a probability function defined on all its events, is a complete probabilistic description of that experiment. Often we are interested only in certain features of this description. We focus on these features using *random variables*. In this chapter we discuss *discrete* random variables, and in the next we will consider *continuous* random variables. We introduce the Bernoulli, binomial, and geometric random variables.

4.1 Random variables

Suppose we are playing the board game "Snakes and Ladders," where the moves are determined by the sum of two independent throws with a die. An obvious choice of the sample space is

$$\Omega = \{(\omega_1, \omega_2) : \omega_1, \omega_2 \in \{1, 2, \ldots, 6\}\}$$
$$= \{(1, 1), (1, 2), \ldots, (1, 6), (2, 1), \ldots, (6, 5), (6, 6)\}.$$

However, as players of the game, we are *only* interested in the sum of the outcomes of the two throws, i.e., in the value of the function $S : \Omega \to \mathbb{R}$, given by

$$S(\omega_1, \omega_2) = \omega_1 + \omega_2 \quad \text{for } (\omega_1, \omega_2) \in \Omega.$$

In Table 4.1 the possible results of the first throw (top margin), those of the second throw (left margin), and the corresponding values of S (body) are given. Note that the values of S are constant on lines perpendicular to the diagonal. We denote the event that the function S attains the value k by $\{S = k\}$, which is an abbreviation of "the subset of those $\omega = (\omega_1, \omega_2) \in \Omega$ for which $S(\omega_1, \omega_2) = \omega_1 + \omega_2 = k$," i.e.,

$$\{S = k\} = \{(\omega_1, \omega_2) \in \Omega : S(\omega_1, \omega_2) = k\}.$$

Table 4.1. Two throws with a die and the corresponding sum.

		ω_1				
ω_2	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

QUICK EXERCISE 4.1 List the outcomes in the event $\{S = 8\}$.

We denote the probability of the event $\{S = k\}$ by

$$P(S = k),$$

although formally we should write $P(\{S = k\})$ instead of $P(S = k)$. In our example, S attains only the values $k = 2, 3, \ldots, 12$ with positive probability. For example,

$$P(S = 2) = P((1, 1)) = \frac{1}{36},$$

$$P(S = 3) = P(\{(1, 2), (2, 1)\}) = \frac{2}{36},$$

while

$$P(S = 13) = P(\emptyset) = 0,$$

because 13 is an "impossible outcome."

QUICK EXERCISE 4.2 Use Table 4.1 to determine $P(S = k)$ for $k = 4, 5, \ldots, 12$.

Now suppose that for some other game the moves are given by the maximum of two independent throws. In this case we are interested in the value of the function $M : \Omega \to \mathbb{R}$, given by

$$M(\omega_1, \omega_2) = \max\{\omega_1, \omega_2\} \quad \text{for } (\omega_1, \omega_2) \in \Omega.$$

In Table 4.2 the possible results of the first throw (top margin), those of the second throw (left margin), and the corresponding values of M (body) are given. The functions S and M are examples of what we call discrete random variables.

DEFINITION. Let Ω be a sample space. A *discrete random variable* is a function $X : \Omega \to \mathbb{R}$ that takes on a finite number of values $a_1, a_2, \ldots, a_n$ or an infinite number of values $a_1, a_2, \ldots$.

Table 4.2. Two throws with a die and the corresponding maximum.

ω_2	ω_1					
	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	2	3	4	5	6
3	3	3	3	4	5	6
4	4	4	4	4	5	6
5	5	5	5	5	5	6
6	6	6	6	6	6	6

In a way, a discrete random variable X "transforms" a sample space Ω to a more "tangible" sample space $\tilde{\Omega}$, whose events are more directly related to what you are interested in. For instance, S transforms $\Omega = \{(1,1),(1,2),\ldots,$ $(1,6),(2,1),\ldots,(6,5),(6,6)\}$ to $\tilde{\Omega} = \{2,\ldots,12\}$, and M transforms Ω to $\tilde{\Omega} = \{1,\ldots,6\}$. Of course, there is a price to pay: one has to calculate the probabilities of X. Or, to say things more formally, one has to determine the *probability distribution* of X, i.e., to describe how the probability mass is *distributed* over possible values of X.

4.2 The probability distribution of a discrete random variable

Once a discrete random variable X is introduced, the sample space Ω is no longer important. It suffices to list the possible values of X and their corresponding probabilities. This information is contained in the *probability mass function* of X.

> DEFINITION. The *probability mass function* p of a discrete random variable X is the function $p : \mathbb{R} \to [0,1]$, defined by
>
> $$p(a) = P(X = a) \quad \text{for } -\infty < a < \infty.$$

If X is a discrete random variable that takes on the values $a_1, a_2, \ldots$, then

$$p(a_i) > 0, \quad p(a_1) + p(a_2) + \cdots = 1, \quad \text{and } p(a) = 0 \text{ for all other } a.$$

As an example we give the probability mass function p of M.

a	1	2	3	4	5	6
$p(a)$	1/36	3/36	5/36	7/36	9/36	11/36

Of course, $p(a) = 0$ for all other a.

The distribution function of a random variable

As we will see, so-called continuous random variables cannot be specified by giving a probability mass function. However, the *distribution function* of a random variable X (also known as the *cumulative distribution function*) allows us to treat discrete and continuous random variables in the same way.

> DEFINITION.   The *distribution function F* of a random variable X is the function $F : \mathbb{R} \to [0, 1]$, defined by
>
> $$F(a) = \mathrm{P}(X \leq a) \quad \text{for } -\infty < a < \infty.$$

Both the probability mass function and the distribution function of a discrete random variable X contain all the probabilistic information of X; the *probability distribution* of X is determined by either of them. In fact, the distribution function F of a discrete random variable X can be expressed in terms of the probability mass function p of X and vice versa. If X attains values $a_1, a_2, \ldots$, such that

$$p(a_i) > 0, \quad p(a_1) + p(a_2) + \cdots = 1,$$

then

$$F(a) = \sum_{a_i \leq a} p(a_i).$$

We see that, for a discrete random variable X, the distribution function F jumps in each of the a_i, and is constant between successive a_i. The height of the jump at a_i is $p(a_i)$; in this way p can be retrieved from F. For example, see Figure 4.1, where p and F are displayed for the random variable M.
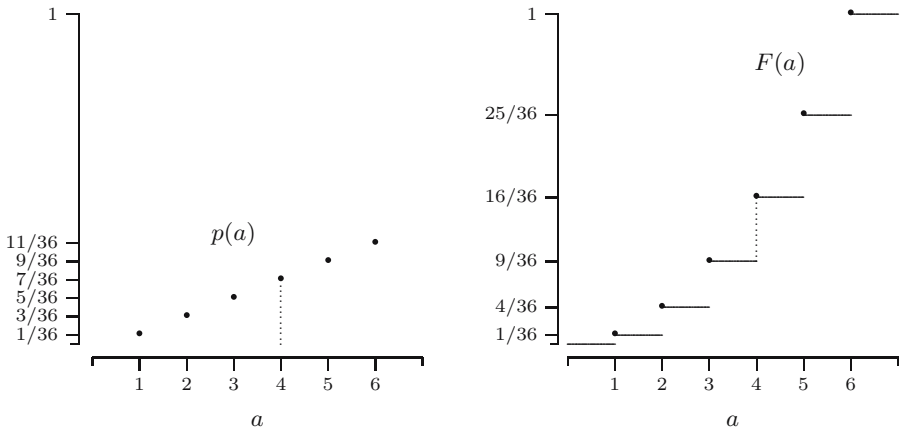
Fig. 4.1. Probability mass function and distribution function of M.

We end this section with three properties of the distribution function F of a random variable X:

1. For $a \leq b$ one has that $F(a) \leq F(b)$. This property is an immediate consequence of the fact that $a \leq b$ implies that the event $\{X \leq a\}$ is contained in the event $\{X \leq b\}$.

2. Since $F(a)$ is a probability, the value of the distribution function is always between 0 and 1. Moreover,

$$\lim_{a \to +\infty} F(a) = \lim_{a \to +\infty} P(X \leq a) = 1$$
$$\lim_{a \to -\infty} F(a) = \lim_{a \to -\infty} P(X \leq a) = 0.$$

3. F is right-continuous, i.e., one has

$$\lim_{\varepsilon \downarrow 0} F(a + \varepsilon) = F(a).$$

This is indicated in Figure 4.1 by bullets. Henceforth we will omit these bullets.

Conversely, any function F satisfying 1, 2, and 3 is the distribution function of some random variable (see Remarks 6.1 and 6.2).

QUICK EXERCISE 4.3 Let X be a discrete random variable, and let a be such that $p(a) > 0$. Show that $F(a) = P(X < a) + p(a)$.

There are many discrete random variables that arise in a natural way. We introduce three of them in the next two sections.

4.3 The Bernoulli and binomial distributions

The Bernoulli distribution is used to model an experiment with only two possible outcomes, often referred to as "success" and "failure", usually encoded as 1 and 0.

> DEFINITION. A discrete random variable X has a *Bernoulli distribution* with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by
>
> $$p_X(1) = P(X = 1) = p \quad \text{and} \quad p_X(0) = P(X = 0) = 1 - p.$$
>
> We denote this distribution by $Ber(p)$.

Note that we wrote p_X instead of p for the probability mass function of X. This was done to emphasize its dependence on X and to avoid possible confusion with the parameter p of the Bernoulli distribution.

Consider the (fictitious) situation that you attend, completely unprepared, a multiple-choice exam. It consists of 10 questions, and each question has four alternatives (of which only one is correct). You will pass the exam if you answer six or more questions correctly. You decide to answer each of the questions in a random way, in such a way that the answer of one question is not affected by the answers of the others. What is the probability that you will pass?

Setting for $i = 1, 2, \ldots, 10$

$$R_i = \begin{cases} 1 & \text{if the } i\text{th answer is correct} \\ 0 & \text{if the } i\text{th answer is incorrect,} \end{cases}$$

the number of correct answers X is given by

$$X = R_1 + R_2 + R_3 + R_4 + R_5 + R_6 + R_7 + R_8 + R_9 + R_{10}.$$

QUICK EXERCISE 4.4 Calculate the probability that you answered the first question correctly and the second one incorrectly.

Clearly, X attains only the values $0, 1, \ldots, 10$. Let us first consider the case $X = 0$. Since the answers to the different questions do not influence each other, we conclude that the events $\{R_1 = a_1\}, \ldots, \{R_{10} = a_{10}\}$ are independent for every choice of the a_i, where each a_i is 0 or 1. We find

$$\begin{aligned} \mathrm{P}(X = 0) &= \mathrm{P}(\text{not a single } R_i \text{ equals } 1) \\ &= \mathrm{P}(R_1 = 0, R_2 = 0, \ldots, R_{10} = 0) \\ &= \mathrm{P}(R_1 = 0)\,\mathrm{P}(R_2 = 0) \cdots \mathrm{P}(R_{10} = 0) \\ &= \left(\frac{3}{4}\right)^{10}. \end{aligned}$$

The probability that we have answered exactly one question correctly equals

$$\mathrm{P}(X = 1) = \frac{1}{4} \cdot \left(\frac{3}{4}\right)^9 \cdot 10,$$

which is the probability that the answer is correct times the probability that the other nine answers are wrong, times the number of ways in which this can occur:

$$\begin{aligned} \mathrm{P}(X = 1) = \quad &\mathrm{P}(R_1 = 1)\,\mathrm{P}(R_2 = 0)\,\mathrm{P}(R_3 = 0) \cdots \mathrm{P}(R_{10} = 0) \\ &+ \mathrm{P}(R_1 = 0)\,\mathrm{P}(R_2 = 1)\,\mathrm{P}(R_3 = 0) \cdots \mathrm{P}(R_{10} = 0) \\ &\qquad\qquad\qquad\qquad \vdots \\ &+ \mathrm{P}(R_1 = 0)\,\mathrm{P}(R_2 = 0)\,\mathrm{P}(R_3 = 0) \cdots \mathrm{P}(R_{10} = 1). \end{aligned}$$

In general we find for $k = 0, 1, \ldots, 10$, again using independence, that

$$P(X = k) = \left(\frac{1}{4}\right)^k \cdot \left(\frac{3}{4}\right)^{10-k} \cdot C_{10,k},$$

which is the probability that k questions were answered correctly times the probability that the other $10-k$ answers are wrong, times the number of ways $C_{10,k}$ this can occur.

So $C_{10,k}$ is the number of different ways in which one can choose k correct answers from the list of 10. We already have seen that $C_{10,0} = 1$, because there is only one way to do everything wrong; and that $C_{10,1} = 10$, because each of the 10 questions may have been answered correctly.

More generally, if we have to choose k different objects out of an ordered list of n objects, and the order in which we pick the objects matters, then for the first object you have n possibilities, and no matter which object you pick, for the second one there are $n - 1$ possibilities. For the third there are $n - 2$ possibilities, and so on, with $n - (k - 1)$ possibilities for the kth. So there are

$$n(n - 1) \cdots (n - (k - 1))$$

ways to choose the k objects.

In how many ways can we choose three questions? When the order matters, there are $10 \cdot 9 \cdot 8$ ways. However, the order in which these three questions are selected does *not* matter: to answer questions 2, 5, and 8 correctly is the same as answering questions 8, 2, and 5 correctly, and so on. The triplet $\{2, 5, 8\}$ can be chosen in $3 \cdot 2 \cdot 1$ different orders, all with the same result. There are six permutations of the numbers 2, 5, and 8 (see page 14).

Thus, compensating for this six-fold overcount, the number $C_{10,3}$ of ways to correctly answer 3 questions out of 10 becomes

$$C_{10,3} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1}.$$

More generally, for $n \geq 1$ and $1 \leq k \leq n$,

$$C_{n,k} = \frac{n(n - 1) \cdots (n - (k - 1))}{k(k - 1) \cdots 2 \cdot 1}.$$

Note that this is equal to

$$\frac{n!}{k!\,(n - k)!},$$

which is usually denoted by $\binom{n}{k}$, so $C_{n,k} = \binom{n}{k}$. Moreover, in accordance with $0! = 1$ (as defined in Chapter 2), we put $C_{n,0} = \binom{n}{0} = 1$.

QUICK EXERCISE 4.5 Show that $\binom{n}{n-k} = \binom{n}{k}$.

Substituting $\binom{10}{k}$ for $C_{10,k}$ we obtain

$$P(X = k) = \binom{10}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{10-k}.$$

Since $P(X \geq 6) = P(X = 6) + \cdots + P(X = 10)$, it is now an easy (but te-
dious) exercise to determine the probability that you will pass. One finds that
$P(X \geq 6) = 0.0197$. It pays to study, doesn't it?!

The preceding random variable X is an example of a random variable with a
binomial distribution with parameters $n = 10$ and $p = 1/4$.

> DEFINITION. A discrete random variable X has a *binomial distri-
> bution* with parameters n and p, where $n = 1, 2, \ldots$ and $0 \leq p \leq 1$,
> if its probability mass function is given by
>
> $$p_X(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n.$$
>
> We denote this distribution by $Bin(n, p)$.

Figure 4.2 shows the probability mass function p_X and distribution function
F_X of a $Bin(10, \frac{1}{4})$ distributed random variable.
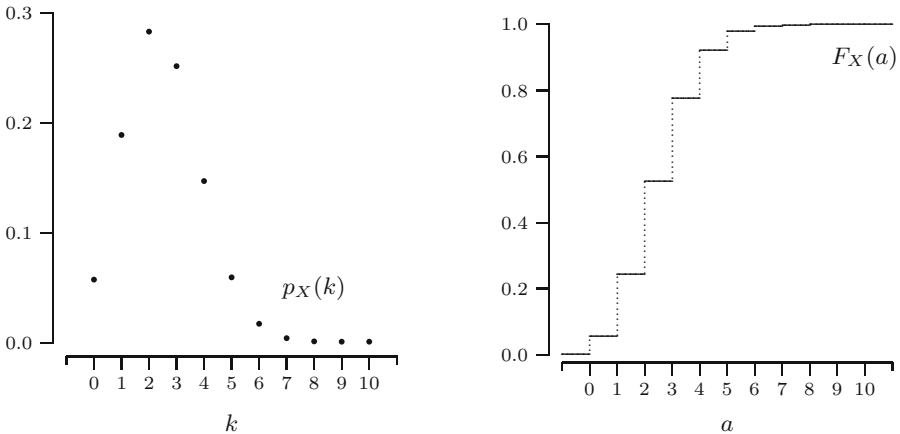
Fig. 4.2. Probability mass function and distribution function of the $Bin(10, \frac{1}{4})$
distribution.

4.4 The geometric distribution

In 1986, Weinberg and Gladen [38] investigated the number of menstrual cy-
cles it took women to become pregnant, measured from the moment they had

decided to become pregnant. We model the number of cycles up to pregnancy by a random variable X.

Assume that the probability that a woman becomes pregnant during a particular cycle is equal to p, for some p with $0 < p \leq 1$, independent of the previous cycles. Then clearly $P(X = 1) = p$. Due to the independence of consecutive cycles, one finds for $k = 1, 2, \ldots$ that

$$P(X = k) = P(\text{no pregnancy in the first } k - 1 \text{ cycles, pregnancy in the } k\text{th})$$
$$= (1 - p)^{k-1} p.$$

This random variable X is an example of a random variable with a geometric distribution with parameter p.

> DEFINITION. A discrete random variable X has a *geometric distribution* with parameter p, where $0 < p \leq 1$, if its probability mass function is given by
>
> $$p_X(k) = P(X = k) = (1 - p)^{k-1} p \quad \text{for } k = 1, 2, \ldots .$$
>
> We denote this distribution by $Geo(p)$.

Figure 4.3 shows the probability mass function p_X and distribution function F_X of a $Geo(\frac{1}{4})$ distributed random variable.
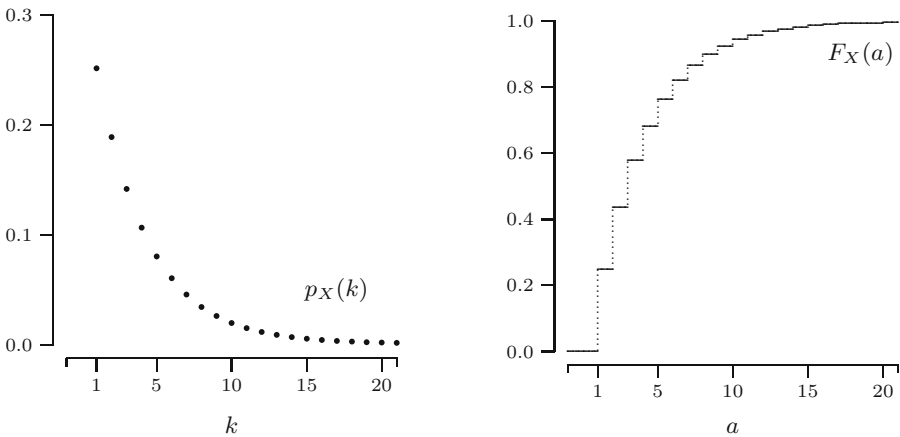
Fig. 4.3. Probability mass function and distribution function of the $Geo(\frac{1}{4})$ distribution.

QUICK EXERCISE 4.6 Let X have a $Geo(p)$ distribution. For $n \geq 0$, show that $P(X > n) = (1 - p)^n$.

The geometric distribution has a remarkable property, which is known as the *memoryless property*.[1] For $n, k = 0, 1, 2, \ldots$ one has

$$P(X > n + k \mid X > k) = P(X > n).$$

We can derive this equality using the result from Quick exercise 4.6:

$$P(X > n + k \mid X > k) = \frac{P(\{X > k + n\} \cap \{X > k\})}{P(X > k)}$$

$$= \frac{P(X > k + n)}{P(X > k)} = \frac{(1 - p)^{n+k}}{(1 - p)^k}$$

$$= (1 - p)^n = P(X > n).$$

4.5 Solutions to the quick exercises

4.1 From Table 4.1, one finds that

$$\{S = 8\} = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}.$$

4.2 From Table 4.1, one determines the following table.

k	4	5	6	7	8	9	10	11	12
$P(S = k)$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

4.3 Since $\{X \le a\} = \{X < a\} \cup \{X = a\}$, it follows that

$$F(a) = P(X \le a) = P(X < a) + P(X = a) = P(X < a) + p(a).$$

Not very interestingly: this also holds if $p(a) = 0$.

4.4 The probability that you answered the first question correctly and the second one incorrectly is given by $P(R_1 = 1, R_2 = 0)$. Due to independence, this is equal to $P(R_1 = 1) P(R_2 = 0) = \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{16}$.

4.5 Rewriting yields

$$\binom{n}{n-k} = \frac{n!}{(n-k)! \, (n - (n - k))!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

---

[1] In fact, the geometric distribution is the only discrete random variable with this property.

4.6 There are two ways to show that $P(X > n) = (1 - p)^n$. The easiest way is to realize that $P(X > n)$ is the probability that we had "no success in the first n trials," which clearly equals $(1 - p)^n$. A more involved way is by calculation:

$$
\begin{aligned}
P(X > n) &= P(X = n + 1) + P(X = n + 2) + \cdots \\
&= (1 - p)^n p + (1 - p)^{n+1} p + \cdots \\
&= (1 - p)^n p \left(1 + (1 - p) + (1 - p)^2 + \cdots \right).
\end{aligned}
$$

If we recall from calculus that

$$
\sum_{k=0}^{\infty} (1 - p)^k = \frac{1}{1 - (1 - p)} = \frac{1}{p},
$$

the answer follows immediately.

4.6 Exercises

4.1 ⊞ Let Z represent the number of times a 6 appeared in two independent throws of a die, and let S and M be as in Section 4.1.

a. Describe the probability distribution of Z, by giving either the probability mass function p_Z of Z or the distribution function F_Z of Z. What type of distribution does Z have, and what are the values of its parameters?

b. List the outcomes in the events $\{M = 2,\ Z = 0\}$, $\{S = 5,\ Z = 1\}$, and $\{S = 8,\ Z = 1\}$. What are their probabilities?

c. Determine whether the events $\{M = 2\}$ and $\{Z = 0\}$ are independent.

4.2 Let X be a discrete random variable with probability mass function p given by:

a	-1	0	1	2
$p(a)$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{2}$

and $p(a) = 0$ for all other a.

a. Let the random variable Y be defined by $Y = X^2$, i.e., if $X = 2$, then $Y = 4$. Calculate the probability mass function of Y.

b. Calculate the value of the distribution functions of X and Y in $a = 1$, $a = 3/4$, and $a = \pi - 3$.

4.3 ⊡ Suppose that the distribution function of a discrete random variable X is given by

$$F(a) = \begin{cases} 0 & \text{for } a < 0 \\ \frac{1}{3} & \text{for } 0 \le a < \frac{1}{2} \\ \frac{1}{2} & \text{for } \frac{1}{2} \le a < \frac{3}{4} \\ 1 & \text{for } a \ge \frac{3}{4}. \end{cases}$$

Determine the probability mass function of X.

4.4 You toss n coins, each showing heads with probability p, independently of the other tosses. Each coin that shows tails is tossed again. Let X be the total number of heads.

a. What type of distribution does X have? Specify its parameter(s).

b. What is the probability mass function of the total number of heads X?

4.5 A fair die is thrown until the sum of the results of the throws exceeds 6. The random variable X is the number of throws needed for this. Let F be the distribution function of X. Determine $F(1)$, $F(2)$, and $F(7)$.

4.6 ⊡ Three times we randomly draw a number from the following numbers:

$$1 \quad 2 \quad 3.$$

If X_i represents the ith draw, $i = 1, 2, 3$, then the probability mass function of X_i is given by

a	1	2	3
$P(X_i = a)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

and $P(X_i = a) = 0$ for all other a. We assume that each draw is independent of the previous draws. Let $\bar{X}$ be the average of X_1, X_2, and X_3, i.e.,

$$\bar{X} = \frac{X_1 + X_2 + X_3}{3}.$$

a. Determine the probability mass function $p_{\bar{X}}$ of $\bar{X}$.

b. Compute the probability that exactly two draws are equal to 1.

4.7 ⊡ A shop receives a batch of 1000 cheap lamps. The odds that a lamp is defective are 0.1%. Let X be the number of defective lamps in the batch.

a. What kind of distribution does X have? What is/are the value(s) of parameter(s) of this distribution?

b. What is the probability that the batch contains no defective lamps? One defective lamp? More than two defective ones?

4.8 ⊡ In Section 1.4 we saw that each space shuttle has six O-rings and that each O-ring fails with probability

$$p(t) = \frac{e^{a+b\cdot t}}{1 + e^{a+b\cdot t}},$$

where $a = 5.085$, $b = -0.1156$, and t is the temperature (in degrees Fahrenheit) at the time of the launch of the space shuttle. At the time of the fatal launch of the *Challenger*, $t = 31$, yielding $p(31) = 0.8178$.

a. Let X be the number of failing O-rings at launch temperature 31°F. What type of probability distribution does X have, and what are the values of its parameters?

b. What is the probability $P(X \geq 1)$ that at least one O-ring fails?

4.9 For simplicity's sake, let us assume that all space shuttles will be launched at 81°F (which is the highest recorded launch temperature in Figure 1.3). With this temperature, the probability of an O-ring failure is equal to $p(81) = 0.0137$ (see Section 1.4 or Exercise 4.8).

a. What is the probability that during 23 launches no O-ring will fail, but that at least one O-ring will fail during the 24th launch of a space shuttle?

b. What is the probability that no O-ring fails during 24 launches?

4.10 ⊞ Early in the morning, a group of m people decides to use the elevator in an otherwise deserted building of 21 floors. Each of these persons chooses his or her floor independently of the others, and—from our point of view—completely at random, so that each person selects a floor with probability $1/21$. Let S_m be the number of times the elevator stops. In order to study S_m, we introduce for $i = 1, 2, \ldots, 21$ random variables R_i, given by

$$R_i = \begin{cases} 1 & \text{if the elevator stops at the } i\text{th floor} \\ 0 & \text{if the elevator does not stop at the } i\text{th floor.} \end{cases}$$

a. Each R_i has a $Ber(p)$ distribution. Show that $p = 1 - \left(\frac{20}{21}\right)^m$.

b. From the way we defined S_m, it follows that

$$S_m = R_1 + R_2 + \cdots + R_{21}.$$

Can we conclude that S_m has a $Bin(21, p)$ distribution, with p as in part **a**? Why or why not?

c. Clearly, if $m = 1$, one has that $P(S_1 = 1) = 1$. Show that for $m = 2$

$$P(S_2 = 1) = \frac{1}{21} = 1 - P(S_2 = 2),$$

and that S_3 has the following distribution.

a	1	2	3
$P(S_3 = a)$	1/441	60/441	380/441

4.11 You decide to play monthly in two different lotteries, and you stop playing as soon as you win a prize in one (or both) lotteries of at least one million euros. Suppose that every time you participate in these lotteries, the probability to win one million (or more) euros is p_1 for one of the lotteries and p_2 for the other. Let M be the number of times you participate in these lotteries until winning at least one prize. What kind of distribution does M have, and what is its parameter?

4.12 ☐ You and a friend want to go to a concert, but unfortunately only one ticket is still available. The man who sells the tickets decides to toss a coin until heads appears. In each toss heads appears with probability p, where $0 < p < 1$, independent of each of the previous tosses. If the number of tosses needed is odd, your friend is allowed to buy the ticket; otherwise you can buy it. Would you agree to this arrangement?

4.13 ⊞ A box contains an unknown number N of identical bolts. In order to get an idea of the size N, we randomly mark one of the bolts from the box. Next we select at random a bolt from the box. If this is the marked bolt we stop, otherwise we return the bolt to the box, and we randomly select a second one, etc. We stop when the selected bolt is the marked one. Let X be the number of times a bolt was selected. Later (in Exercise 21.11) we will try to find an estimate of N. Here we look at the probability distribution of X.

  a. What is the probability distribution of X? Specify its parameter(s)!
  b. The drawback of this approach is that X can attain any of the values $1, 2, 3, \ldots$, so that if N is large we might be sampling from the box for quite a long time. We decide to sample from the box in a slightly different way: after we have randomly marked one of the bolts in the box, we select at random a bolt from the box. If this is the marked one, we stop, otherwise we randomly select a second bolt (we do *not* return the selected bolt). We stop when we select the marked bolt. Let Y be the number of times a bolt was selected.
  Show that $\mathrm{P}(Y = k) = 1/N$ for $k = 1, 2, \ldots, N$ (Y has a so-called *discrete uniform* distribution).
  c. Instead of randomly marking one bolt in the box, we mark m bolts, with m smaller than N. Next, we randomly select r bolts; Z is the number of marked bolts in the sample.
  Show that

$$\mathrm{P}(Z = k) = \frac{\binom{m}{k}\binom{N-m}{r-k}}{\binom{N}{r}}, \qquad \text{for} \quad k = 0, 1, 2, \ldots, r.$$

  (Z has a so-called *hypergeometric* distribution, with parameters m, N, and r.)

4.14 We throw a coin until a head turns up for the second time, where p is the probability that a throw results in a head and we assume that the outcome

of each throw is independent of the previous outcomes. Let X be the number of times we have thrown the coin.

a. Determine $P(X = 2)$, $P(X = 3)$, and $P(X = 4)$.

b. Show that $P(X = n) = (n - 1)p^2(1 - p)^{n-2}$ for $n \geq 2$.

5

Continuous random variables

Many experiments have outcomes that take values on a continuous scale. For example, in Chapter 2 we encountered the load at which a model of a bridge collapses. These experiments have *continuous* random variables naturally associated with them.

5.1 Probability density functions

One way to look at continuous random variables is that they arise by a (never-ending) process of refinement from discrete random variables. Suppose, for example, that a discrete random variable associated with some experiment takes on the value 6.283 with probability p. If we refine, in the sense that we also get to know the fourth decimal, then the probability p is spread over the outcomes $6.2830, 6.2831, \ldots, 6.2839$. Usually this will mean that each of these new values is taken on with a probability that is much smaller than p—the sum of the ten probabilities is p. Continuing the refinement process to more and more decimals, the probabilities of the possible values of the outcomes become smaller and smaller, approaching zero. However, the probability that the possible values lie in some fixed interval $[a, b]$ will settle down. This is closely related to the way sums converge to an integral in the definition of the integral and motivates the following definition.

DEFINITION. A random variable X is *continuous* if for some function $f : \mathbb{R} \to \mathbb{R}$ and for any numbers a and b with $a \leq b$,

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x.$$

The function f has to satisfy $f(x) \geq 0$ for all x and $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$. We call f the *probability density function* (or *probability density*) of X.
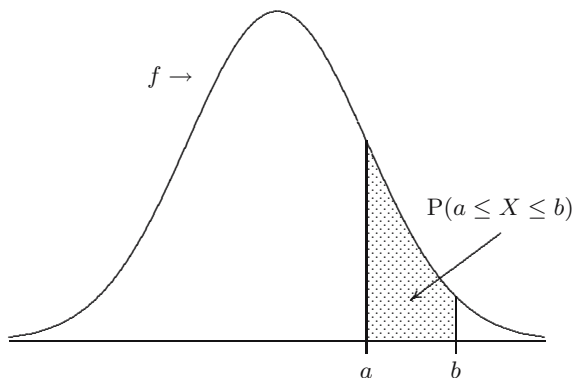
Fig. 5.1. Area under a probability density function f on the interval $[a, b]$.

Note that the probability that X lies in an interval $[a, b]$ is equal to the area under the probability density function f of X over the interval $[a, b]$; this is illustrated in Figure 5.1. So if the interval gets smaller and smaller, the probability will go to zero: for any positive ε

$$P(a - \varepsilon \le X \le a + \varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\, dx,$$

and sending ε to 0, it follows that for any a

$$P(X = a) = 0.$$

This implies that for continuous random variables you may be careless about the precise form of the intervals:

$$P(a \le X \le b) = P(a < X \le b) = P(a < X < b) = P(a \le X < b).$$

What does $f(a)$ represent? Note (see also Figure 5.2) that

$$P(a - \varepsilon \le X \le a + \varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\, dx \approx 2\varepsilon f(a) \tag{5.1}$$

for small positive ε. Hence $f(a)$ can be interpreted as a (relative) measure of how likely it is that X will be near a. However, do not think of $f(a)$ as a probability: $f(a)$ can be arbitrarily large. An example of such an f is given in the following exercise.

QUICK EXERCISE 5.1 Let the function f be defined by $f(x) = 0$ if $x \le 0$ or $x \ge 1$, and $f(x) = 1/(2\sqrt{x})$ for $0 < x < 1$. You can check quickly that f satisfies the two properties of a probability density function. Let X be a random variable with f as its probability density function. Compute the probability that X lies between 10^{-4} and 10^{-2}.
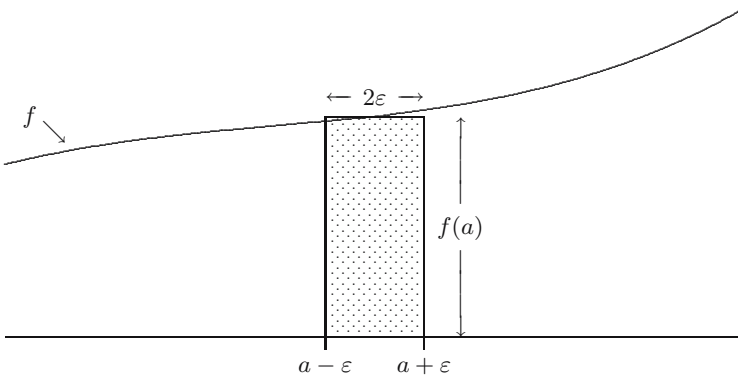
Fig. 5.2. Approximating the probability that X lies ε-close to a.

You should realize that discrete random variables do not have a probability density function f and continuous random variables do not have a probability mass function p, but that *both* have a distribution function $F(a) = \mathrm{P}(X \le a)$. Using the fact that for $a < b$ the event $\{X \le b\}$ is a disjoint union of the events $\{X \le a\}$ and $\{a < X \le b\}$, we can express the probability that X lies in an interval $(a, b]$ directly in terms of F for *both* cases:

$$\mathrm{P}(a < X \le b) = \mathrm{P}(X \le b) - \mathrm{P}(X \le a) = F(b) - F(a).$$

There is a simple relation between the distribution function F and the probability density function f of a continuous random variable. It follows from integral calculus that

$$F(b) = \int_{-\infty}^{b} f(x)\,\mathrm{d}x \quad \text{and}^{1} \quad f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x).$$

Both the probability density function and the distribution function of a continuous random variable X contain all the probabilistic information about X; the *probability distribution* of X is described by either of them.

We illustrate all this with an example. Suppose we want to make a probability model for an experiment that can be described as "an object hits a disc of radius r in a completely arbitrary way" (of course, this is not *you* playing darts—nevertheless we will refer to this example as the darts example). We are interested in the distance X between the hitting point and the center of the disc. Since distances cannot be negative, we have $F(b) = \mathrm{P}(X \le b) = 0$ when $b < 0$. Since the object hits the disc, we have $F(b) = 1$ when $b > r$. That the dart hits the disk in a completely arbitrary way we interpret as that the probability of hitting any region is proportional to the area of that region. In particular, because the disc has area πr^2 and the disc with radius b has area πb^2, we should put

---

[1] This holds for all x where f is continuous.

$$F(b) = P(X \le b) = \frac{\pi b^2}{\pi r^2} = \frac{b^2}{r^2} \quad \text{for } 0 \le b \le r.$$

Then the probability density function f of X is equal to 0 outside the interval $[0, r]$ and

$$f(x) = \frac{d}{dx}F(x) = \frac{1}{r^2}\frac{d}{dx}x^2 = \frac{2x}{r^2} \quad \text{for } 0 \le x \le r.$$

QUICK EXERCISE 5.2 Compute for the darts example the probability that $0 < X \le r/2$, and the probability that $r/2 < X \le r$.

5.2 The uniform distribution

In this section we encounter a continuous random variable that describes an experiment where the outcome is completely arbitrary, except that we know that it lies between certain bounds. Many experiments of physical origin have this kind of behavior. For instance, suppose we measure for a long time the emission of radioactive particles of some material. Suppose that the experiment consists of recording in each hour at what times the particles are emitted. Then the outcomes will lie in the interval [0,60] minutes. If the measurements would concentrate in any way, there is either something wrong with your Geiger counter or you are about to discover some new physical law. Not concentrating in any way means that subintervals of the same length should have the same probability. It is then clear (cf. equation (5.1)) that the probability density function associated with this experiment should be constant on $[0, 60]$. This motivates the following definition.

DEFINITION. A continuous random variable has a *uniform distribution* on the interval $[\alpha, \beta]$ if its probability density function f is given by $f(x) = 0$ if x is not in $[\alpha, \beta]$ and

$$f(x) = \frac{1}{\beta - \alpha} \quad \text{for } \alpha \le x \le \beta.$$

We denote this distribution by $U(\alpha, \beta)$.

QUICK EXERCISE 5.3 Argue that the distribution function F of a random variable that has a $U(\alpha, \beta)$ distribution is given by $F(x) = 0$ if $x < \alpha$, $F(x) = 1$ if $x > \beta$, and $F(x) = (x - \alpha)/(\beta - \alpha)$ for $\alpha \le x \le \beta$.

In Figure 5.3 the probability density function and the distribution function of a $U(0, \frac{1}{3})$ distribution are depicted.
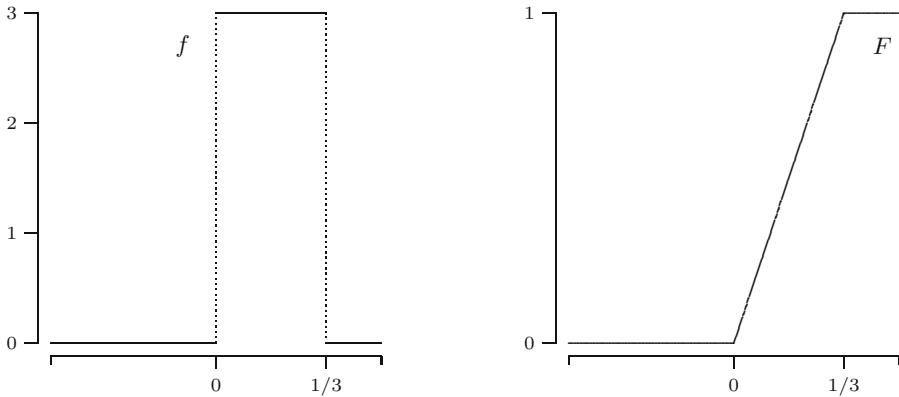
Fig. 5.3. The probability density function and the distribution function of the $U(0, \frac{1}{3})$ distribution.

5.3 The exponential distribution

We already encountered the exponential distribution in the chemical reactor example of Chapter 3. We will give an argument why it appears in that example. Let v be the effluent volumetric flow rate, i.e., the volume that leaves the reactor over a time interval $[0, t]$ is vt (and an equal volume enters the vessel at the other end). Let V be the volume of the reactor vessel. Then in total a fraction $(v/V) \cdot t$ will have left the vessel during $[0, t]$, when t is not too large. Let the random variable T be the residence time of a particle in the vessel. To compute the distribution of T, we divide the interval $[0, t]$ in n small intervals of equal length t/n. Assuming perfect mixing, so that the particle's position is uniformly distributed over the volume, the particle has probability $p = (v/V) \cdot t/n$ to have left the vessel during any of the n intervals of length t/n. If we assume that the behavior of the particle in different time intervals of length t/n is independent, we have, if we call "leaving the vessel" a success, that T has a geometric distribution with success probability p. It follows (see also Quick exercise 4.6) that the probability $P(T > t)$ that the particle is still in the vessel at time t is, for large n, well approximated by

$$(1 - p)^n = \left(1 - \frac{vt}{Vn}\right)^n.$$

But then, letting $n \to \infty$, we obtain (recall a well-known limit from your calculus course)

$$P(T > t) = \lim_{n \to \infty} \left(1 - \frac{vt}{V} \cdot \frac{1}{n}\right)^n = e^{-\frac{v}{V}t}.$$

It follows that the distribution function of T equals $1 - e^{-\frac{v}{V}t}$, and differentiating we obtain that the probability density function f_T of T is equal to

$$f_T(t) = \frac{\mathrm{d}}{\mathrm{d}t}(1 - \mathrm{e}^{-\frac{v}{V}t}) = \frac{v}{V}\mathrm{e}^{-\frac{v}{V}t} \quad \text{for} \quad t \geq 0.$$

This is an example of an exponential distribution, with parameter v/V.

> DEFINITION. A continuous random variable has an *exponential distribution* with parameter λ if its probability density function f is given by $f(x) = 0$ if $x < 0$ and
>
> $$f(x) = \lambda \mathrm{e}^{-\lambda x} \quad \text{for } x \geq 0.$$
>
> We denote this distribution by $Exp(\lambda)$.

The distribution function F of an $Exp(\lambda)$ distribution is given by

$$F(a) = 1 - \mathrm{e}^{-\lambda a} \quad \text{for} \quad a \geq 0.$$

In Figure 5.4 we show the probability density function and the distribution function of the $Exp(0.25)$ distribution.
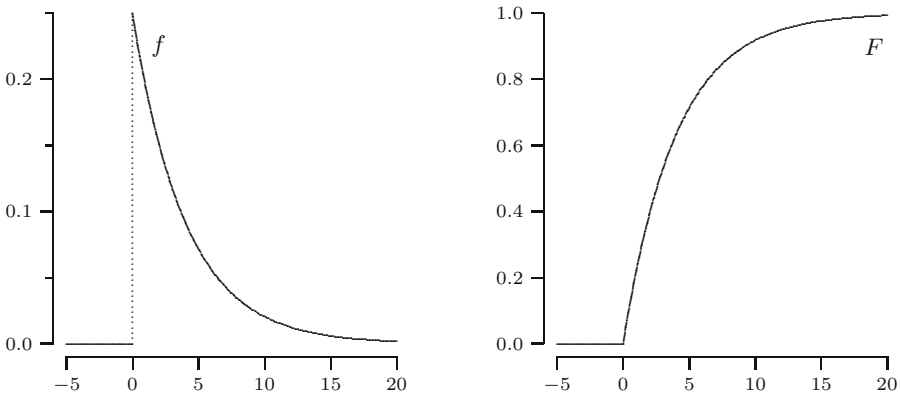
Fig. 5.4. The probability density and the distribution function of the $Exp(0.25)$ distribution.

Since we obtained the exponential distribution directly from the geometric distribution it should not come as a surprise that the exponential distribution *also* satisfies the memoryless property, i.e., if X has an exponential distribution, then for all $s, t > 0$,

$$\mathrm{P}(X > s + t \mid X > s) = \mathrm{P}(X > t).$$

Actually, this follows directly from

$$\mathrm{P}(X > s + t \mid X > s) = \frac{\mathrm{P}(X > s + t)}{\mathrm{P}(X > s)} = \frac{\mathrm{e}^{-\lambda(s+t)}}{\mathrm{e}^{-\lambda s}} = \mathrm{e}^{-\lambda t} = \mathrm{P}(X > t).$$

QUICK EXERCISE 5.4 A study of the response time of a certain computer system yields that the response time in seconds has an exponentially distributed time with parameter 0.25. What is the probability that the response time exceeds 5 seconds?

5.4 The Pareto distribution

More than a century ago the economist Vilfredo Pareto ([20]) noticed that the number of people whose income exceeded level x was well approximated by C/x^α, for some constants C and $\alpha > 0$ (it appears that for all countries α is around 1.5). A similar phenomenon occurs with city sizes, earthquake rupture areas, insurance claims, and sizes of commercial companies. When these quantities are modeled as realizations of random variables X, then their distribution functions are of the type $F(x) = 1 - 1/x^\alpha$ for $x \geq 1$. (Here 1 is a more or less arbitrarily chosen starting point—what matters is the behavior for large x.) Differentiating, we obtain probability densities of the form $f(x) = \alpha/x^{\alpha+1}$. This motivates the following definition.

> DEFINITION. A continuous random variable has a *Pareto distribution* with parameter $\alpha > 0$ if its probability density function f is given by $f(x) = 0$ if $x < 1$ and
>
> $$f(x) = \frac{\alpha}{x^{\alpha+1}} \quad \text{for } x \geq 1.$$
>
> We denote this distribution by $Par(\alpha)$.
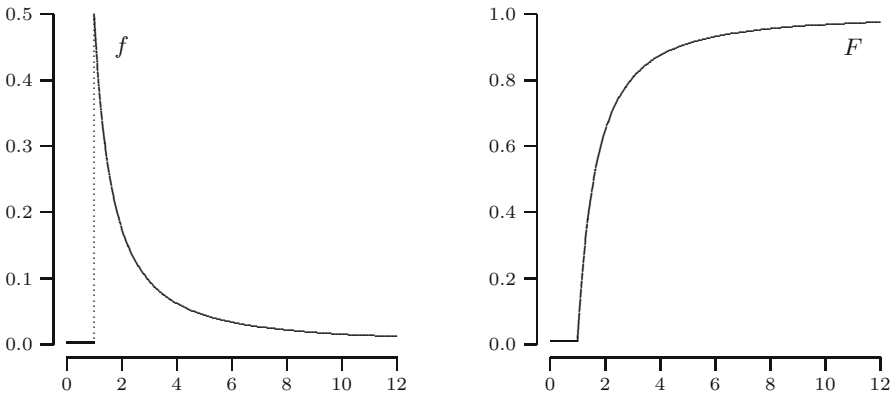
Fig. 5.5. The probability density and the distribution function of the $Par(0.5)$ distribution.

In Figure 5.5 we depicted the probability density f and the distribution function F of the $Par(0.5)$ distribution.

5.5 The normal distribution

The normal distribution plays a central role in probability theory and statistics. One of its first applications was due to C.F. Gauss, who used it in 1809 to model observational errors in astronomy; see [13]. We will see in Chapter 14 that the normal distribution is an important tool to approximate the probability distribution of the average of independent random variables.

> DEFINITION. A continuous random variable has a *normal distribution* with parameters μ and $\sigma^2 > 0$ if its probability density function f is given by
>
> $$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{for } -\infty < x < \infty.$$
>
> We denote this distribution by $N(\mu, \sigma^2)$.

In Figure 5.6 the graphs of the probability density function f and distribution function F of the normal distribution with $\mu = 3$ and $\sigma^2 = 6.25$ are displayed.
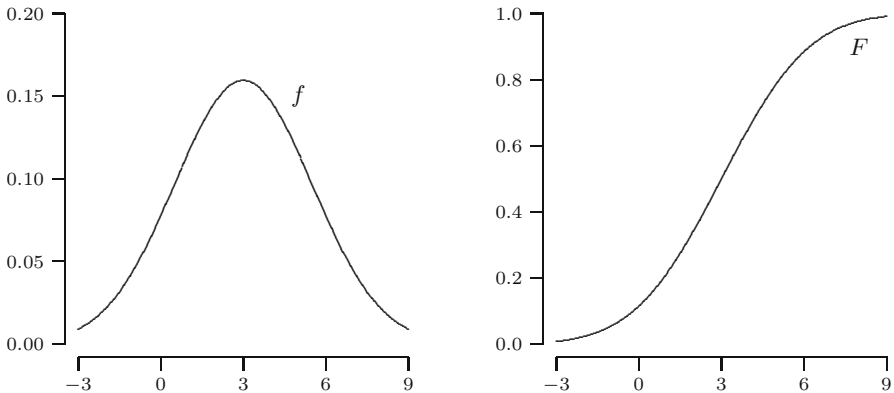
Fig. 5.6. The probability density and the distribution function of the $N(3, 6.25)$ distribution.

If X has an $N(\mu, \sigma^2)$ distribution, then its distribution function is given by

$$F(a) = \int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\, dx \qquad \text{for } -\infty < a < \infty.$$

Unfortunately there is no explicit expression for F; f has no antiderivative. However, as we shall see in Chapter 8, any $N(\mu, \sigma^2)$ distributed random variable can be turned into an $N(0, 1)$ distributed random variable by a simple transformation. As a consequence, a table of the $N(0, 1)$ distribution suffices. The latter is called the *standard* normal distribution, and because of its special role the letter ϕ has been reserved for its probability density function:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \text{for } -\infty < x < \infty.$$

Note that ϕ is symmetric around zero: $\phi(-x) = \phi(x)$ for each x. The corresponding distribution function is denoted by Φ. The table for the standard normal distribution (see Table B.1) does not contain the values of $\Phi(a)$, but rather the so-called *right tail probabilities* $1 - \Phi(a)$. If, for instance, we want to know the probability that a standard normal random variable Z is smaller than or equal to 1, we use that $P(Z \leq 1) = 1 - P(Z \geq 1)$. In the table we find that $P(Z \geq 1) = 1 - \Phi(1)$ is equal to 0.1587. Hence $P(Z \leq 1) = 1 - 0.1587 = 0.8413$. With the table you can handle tail probabilities with numbers a given to two decimals. To find, for instance, $P(Z > 1.07)$, we stay in the same row in the table but move to the seventh column to find that $P(Z > 1.07) = 0.1423$.

QUICK EXERCISE 5.5 Let the random variable Z have a standard normal distribution. Use Table B.1 to find $P(Z \leq 0.75)$. How do you know—without doing any calculations—that the answer should be larger than 0.5?

5.6 Quantiles

Recall the chemical reactor example, where the residence time T, measured in minutes, has an exponential distribution with parameter $\lambda = v/V = 0.25$. As we shall see in the next chapters, a consequence of this choice of λ is that the *mean* time the particle stays in the vessel is 4 minutes. However, from the viewpoint of process control this is not the quantity of interest. Often, there will be some minimal amount of time the particle has to stay in the vessel to participate in the chemical reaction, and we would want that at least 90% of the particles stay in the vessel this minimal amount of time. In other words, we are interested in the number q with the property that $P(T > q) = 0.9$, or equivalently,

$$P(T \leq q) = 0.1.$$

The number q is called the 0.1th quantile or 10th percentile of the distribution. In the case at hand it is easy to determine. We should have

$$P(T \leq q) = 1 - e^{-0.25q} = 0.1.$$

This holds exactly when $e^{-0.25q} = 0.9$ or when $-0.25q = \ln(0.9) = -0.105$. So $q = 0.42$. Hence, although the mean residence time is 4 minutes, 10% of

the particles stays less than 0.42 minute in the vessel, which is just slightly more than 25 seconds! We use the following general definition.

DEFINITION. Let X be a continuous random variable and let p be a number between 0 and 1. The pth *quantile* or $100p$th *percentile* of the distribution of X is the smallest number q_p such that

$$F(q_p) = \mathrm{P}(X \le q_p) = p.$$

The *median* of a distribution is its 50th percentile.

QUICK EXERCISE 5.6 What is the median of the $U(2, 7)$ distribution?

For continuous random variables q_p is often easy to determine. Indeed, if F is *strictly* increasing from 0 to 1 on some interval (which may be infinite to one or both sides), then

$$q_p = F^{\mathrm{inv}}(p),$$

where F^{inv} is the inverse of F. This is illustrated in Figure 5.7 for the $Exp(0.25)$ distribution.
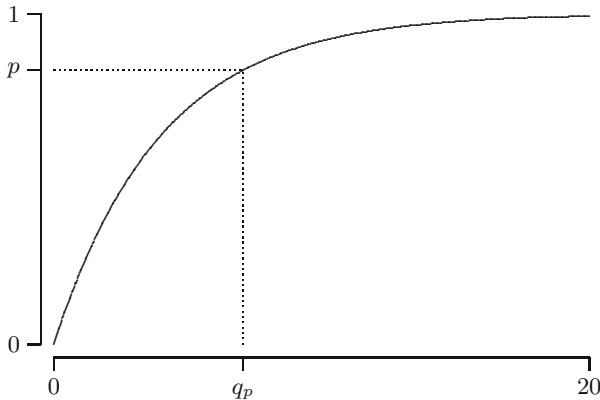
Fig. 5.7. The pth quantile q_p of the $Exp(0.25)$ distribution.

For an exponential distribution it is easy to *compute* quantiles. This is different for the standard normal distribution, where we have to use a table (like Table B.1). For example, the 90th percentile of a standard normal is the number $q_{0.9}$ such that $\Phi(q_{0.9}) = 0.9$, which is the same as $1 - \Phi(q_{0.9}) = 0.1$, and the table gives us $q_{0.9} = 1.28$. This is illustrated in Figure 5.8, with both the probability density function and the distribution function of the standard normal distribution.

QUICK EXERCISE 5.7 Find the 0.95th quantile $q_{0.95}$ of a standard normal distribution, accurate to two decimals.
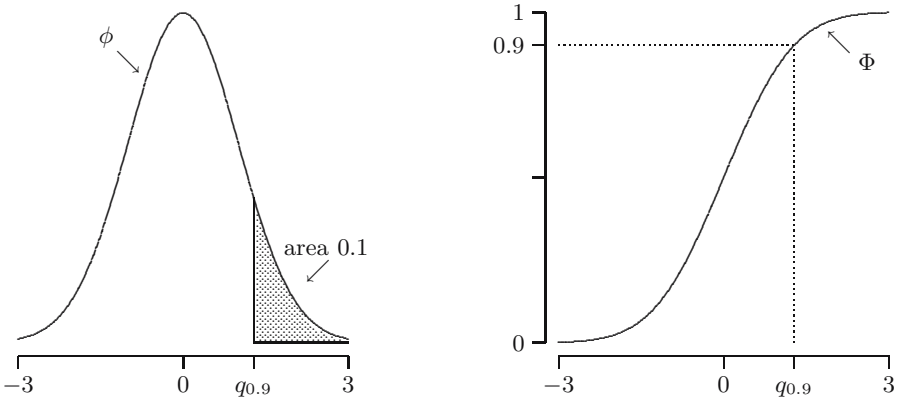
Fig. 5.8. The 90th percentile of the $N(0,1)$ distribution.

5.7 Solutions to the quick exercises

5.1 We know from integral calculus that for $0 \le a \le b \le 1$

$$\int_a^b f(x)\,dx = \int_a^b \frac{1}{2\sqrt{x}}\,dx = \sqrt{b} - \sqrt{a}.$$

Hence $\int_{-\infty}^\infty f(x)\,dx = \int_0^1 1/(2\sqrt{x})\,dx = 1$ (so f is a probability density function—nonnegativity being obvious), and

$$P\left(10^{-4} \le X \le 10^{-2}\right) = \int_{10^{-4}}^{10^{-2}} \frac{1}{2\sqrt{x}}\,dx$$
$$= \sqrt{10^{-2}} - \sqrt{10^{-4}} = 10^{-1} - 10^{-2} = 0.09.$$

Actually, the random variable X arises in a natural way; see equation (7.1).

5.2 We have $P(0 < X \le r/2) = F(r/2) - F(0) = (1/2)^2 - 0^2 = 1/4$, and $P(r/2 < X \le r) = F(r) - F(r/2) = 1 - 1/4 = 3/4$, no matter what the radius of the disc is!

5.3 Since $f(x) = 0$ for $x < \alpha$, we have $F(x) = 0$ if $x < \alpha$. Also, since $f(x) = 0$ for all $x > \beta$, $F(x) = 1$ if $x > \beta$. In between

$$F(x) = \int_{-\infty}^x f(y)\,dy = \int_\alpha^x \frac{1}{\beta - \alpha}\,dy = \left[\frac{y}{\beta - \alpha}\right]_\alpha^x = \frac{x - \alpha}{\beta - \alpha}.$$

In other words; the distribution function increases linearly from the value 0 in α to the value 1 in β.

5.4 If X is the response time, we ask for $P(X > 5)$. This equals

$$P(X > 5) = e^{-0.25 \cdot 5} = e^{-1.25} = 0.2865\ldots.$$

5.5 In the eighth row and sixth column of the table, we find that $1 - \Phi(0.75) = 0.2266$. Hence the answer is $1 - 0.2266 = 0.7734$. Because of the symmetry of the probability density ϕ, half of the mass of a standard normal distribution lies on the negative axis. Hence for any number $a > 0$, it should be true that $P(Z \le a) > P(Z \le 0) = 0.5$.

5.6 The median is the number $q_{0.5} = F^{\text{inv}}(0.5)$. You either see directly that you have got half of the mass to both sides of the middle of the interval, hence $q_{0.5} = (2 + 7)/2 = 4.5$, or you solve with the distribution function:

$$\frac{1}{2} = F(q) = \frac{q - 2}{7 - 2}, \quad \text{and so} \quad q = 4.5.$$

5.7 Since $\Phi(q_{0.95}) = 0.95$ is the same as $1 - \Phi(q_{0.95}) = 0.05$, the table gives us $q_{0.95} = 1.64$, or more precisely, if we interpolate between the fourth and the fifth column; 1.645.

5.8 Exercises

5.1 Let X be a continuous random variable with probability density function

$$f(x) = \begin{cases} \frac{3}{4} & \text{for } 0 \le x \le 1 \\ \frac{1}{4} & \text{for } 2 \le x \le 3 \\ 0 & \text{elsewhere.} \end{cases}$$

a. Draw the graph of f.

b. Determine the distribution function F of X, and draw its graph.

5.2 ⊡ Let X be a random variable that takes values in $[0, 1]$, and is further given by

$$F(x) = x^2 \quad \text{for } 0 \le x \le 1.$$

Compute $P\left(\frac{1}{2} < X \le \frac{3}{4}\right)$.

5.3 Let a continuous random variable X be given that takes values in $[0, 1]$, and whose distribution function F satisfies

$$F(x) = 2x^2 - x^4 \quad \text{for } 0 \le x \le 1.$$

a. Compute $P\left(\frac{1}{4} \le X \le \frac{3}{4}\right)$.

b. What is the probability density function of X?

5.4 ⊞ Jensen, arriving at a bus stop, just misses the bus. Suppose that he decides to walk if the (next) bus takes longer than 5 minutes to arrive. Suppose also that the time in minutes between the arrivals of buses at the bus stop is a continuous random variable with a $U(4, 6)$ distribution. Let X be the time that Jensen will wait.

a. What is the probability that X is less than $4\frac{1}{2}$ (minutes)?

b. What is the probability that X equals 5 (minutes)?

c. Is X a discrete random variable or a continuous random variable?

5.5 ⊡ The probability density function f of a continuous random variable X is given by:

$$f(x) = \begin{cases} cx + 3 & \text{for } -3 \le x \le -2 \\ 3 - cx & \text{for } 2 \le x \le 3 \\ 0 & \text{elsewhere.} \end{cases}$$

a. Compute c.

b. Compute the distribution function of X.

5.6 Let X have an $Exp(0.2)$ distribution. Compute $P(X > 5)$.

5.7 The score of a student on a certain exam is represented by a number between 0 and 1. Suppose that the student passes the exam if this number is at least 0.55. Suppose we model this experiment by a continuous random variable S, the score, whose probability density function is given by

$$f(x) = \begin{cases} 4x & \text{for } 0 \le x \le \frac{1}{2} \\ 4 - 4x & \text{for } \frac{1}{2} \le x \le 1 \\ 0 & \text{elsewhere.} \end{cases}$$

a. What is the probability that the student fails the exam?

b. What is the score that he will obtain with a 50% chance, in other words, what is the 50th percentile of the score distribution?

5.8 ⊞ Consider Quick exercise 5.2. For another dart thrower it is given that *his* distance to the center of the disc Y is described by the following distribution function:

$$G(b) = \sqrt{\frac{b}{r}} \quad \text{for } 0 < b < r$$

and $G(b) = 0$ for $b \le 0$, $G(b) = 1$ for $b \ge r$.

a. Sketch the probability density function $g(y) = \frac{\mathrm{d}}{\mathrm{d}y}G(y)$.

b. Is this person "better" than the person in Quick exercise 5.2?

c. Sketch a distribution function associated to a person who in 90% of his throws hits the disc no further than $0.1 \cdot r$ of the center.

5.9 ⊡ Suppose we choose arbitrarily a point from the square with corners at (2,1), (3,1), (2,2), and (3,2). The random variable A is the area of the triangle with its corners at (2,1), (3,1) and the chosen point (see Figure 5.9).

a. What is the largest area A that can occur, and what is the set of points for which $A \le 1/4$?

$(2,2)$ $(3,2)$

randomly chosen
point

A

$(2,1)$ $(3,1)$
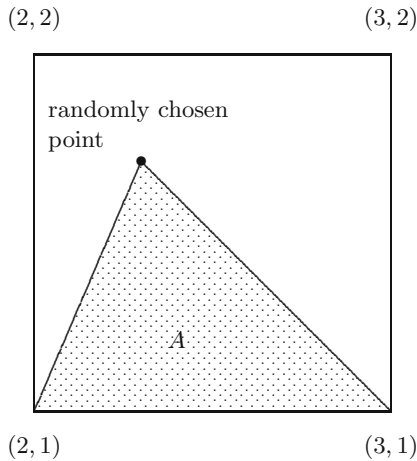
Fig. 5.9. A triangle in a square.

b. Determine the distribution function F of A.

c. Determine the probability density function f of A.

5.10 Consider again the chemical reactor example with parameter $\lambda = 0.5$. We saw in Section 5.6 that 10% of the particles stay in the vessel no longer than about 12 seconds—while the mean residence time is 2 minutes. Which percentage of the particles stay no longer than 2 minutes in the vessel?

5.11 Compute the median of an $Exp(\lambda)$ distribution.

5.12 ⊡ Compute the median of a $Par(1)$ distribution.

5.13 ⊞ We consider a random variable Z with a standard normal distribution.

a. Show why the symmetry of the probability density function ϕ of Z implies that for any a one has $\Phi(-a) = 1 - \Phi(a)$.

b. Use this to compute $P(Z \leq -2)$.

5.14 Determine the 10th percentile of a standard normal distribution.

6

Simulation

Sometimes probabilistic models are so complex that the tools of mathematical analysis are not sufficient to answer all relevant questions about them. Stochastic simulation is an alternative approach: values are generated for the random variables and inserted into the model, thus mimicking outcomes for the whole system. It is shown in this chapter how one can use uniform random number generators to mimic random variables. Also two larger simulation examples are presented.

6.1 What is simulation?

In many areas of science, technology, government, and business, models are used to gain understanding of some part of reality (the portion of interest is often referred to as "the system"). Sometimes these are physical models, such as a scale model of an airplane in a wind tunnel or a scale model of a chemical plant. Other models are abstract, such as macroeconomic models consisting of equations relating things like interest rates, unemployment, and inflation or partial differential equations describing global weather patterns.

In *simulation*, one uses a model to create specific situations in order to study the response of the model to them and then interprets this in terms of what would happen to the system "in the real world." In this way, one can carry out experiments that are impossible, too dangerous, or too expensive to do in the real world—addressing questions like: What happens to the average temperature if we reduce the greenhouse gas emissions globally by 50%? Can the plane still fly if engines 3 and 4 stop in midair? What happens to the distribution of wealth if we halve the tax rate?

More specifically, we focus on situations and problems where randomness or uncertainty or both play a significant or dominant role and should be modeled explicitly. Models for such systems involve random variables, and we speak of *probabilistic* or *stochastic models*. Simulating them is *stochastic simulation*. In

the preceding chapters we have encountered some of the tools of probability theory, and we will encounter others in the chapters to come. With these tools we can compute quantities of interest explicitly for many models. Stochastic simulation of a system means generating values for all the random variables in the model, according to their specified distributions, and recording and analyzing what happens. We refer to the generated values as *realizations* of the random variables.

For us, there are two reasons to learn about stochastic simulation. The first is that for complex systems, simulation can be an alternative to mathematical analysis, sometimes the only one. The second reason is that through simulation, we can get more feeling for random variables, and this is why we study stochastic simulation at this point in the book. We start by asking how we can generate a realization of a random variable.

6.2 Generating realizations of random variables

Simulations are almost always done using computers, which usually have one or more so-called *(pseudo) random number generators*. A call to the random number generator returns a random number between 0 and 1, which mimics a realization of a $U(0,1)$ variable. With this source of uniform (pseudo) randomness we can construct any random variable we want by transforming the outcome, as we shall see.

QUICK EXERCISE 6.1 Describe how you can simulate a coin toss when instead of a coin you have a die. Any ideas on how to simulate a roll of a die if you only have a coin?

Bernoulli random variables

Suppose U has a $U(0,1)$ distribution. To construct a $Ber(p)$ random variable for some $0 < p < 1$, we define

$$X = \begin{cases} 1 & \text{if } U < p, \\ 0 & \text{if } U \geq p \end{cases}$$

so that

$$\mathrm{P}(X = 1) = \mathrm{P}(U < p) = p,$$
$$\mathrm{P}(X = 0) = \mathrm{P}(U \geq p) = 1 - p.$$

This random variable X has a Bernoulli distribution with parameter p.

QUICK EXERCISE 6.2 A random variable Y has outcomes 1, 3, and 4 with the following probabilities: $\mathrm{P}(Y = 1) = 3/5$, $\mathrm{P}(Y = 3) = 1/5$, and $\mathrm{P}(Y = 4) = 1/5$. Describe how to construct Y from a $U(0,1)$ random variable.

Continuous random variables

Suppose we have the distribution function F of a continuous random variable and we wish to construct a random variable with this distribution. We show how to do this if F is strictly increasing from 0 to 1 on an interval. In that case F has an inverse function F^{inv}. Figure 6.1 shows an example: F is strictly increasing on the interval $[2, 10]$; the inverse F^{inv} is a function from the interval $[0, 1]$ to the interval $[2, 10]$.
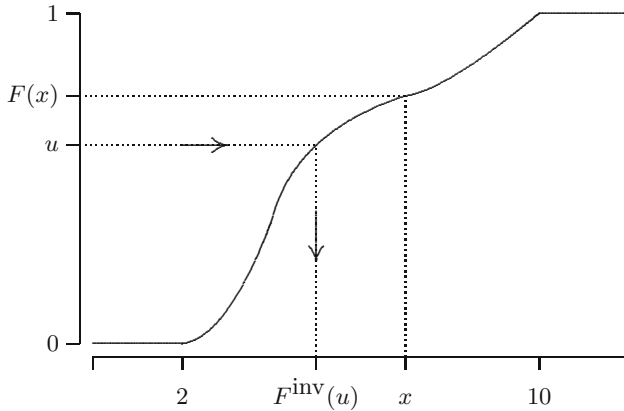
Fig. 6.1. Simulating a continuous random variable using the distribution function.

Note how u relates to $F^{\text{inv}}(u)$ as $F(x)$ relates to x. We see that $u \leq F(x)$ is equivalent with $F^{\text{inv}}(u) \leq x$. If instead of a real number u we consider a $U(0,1)$ random variable U, we obtain that the corresponding events are the same:

$$\{U \leq F(x)\} = \{F^{\text{inv}}(U) \leq x\}. \tag{6.1}$$

We know about the $U(0,1)$ random variable U that $\mathrm{P}(U \leq b) = b$ for any number $0 \leq b \leq 1$. Substituting $b = F(x)$ we see

$$\mathrm{P}(U \leq F(x)) = F(x).$$

From equality (6.1), therefore,

$$\mathrm{P}\big(F^{\text{inv}}(U) \leq x\big) = F(x);$$

in other words, the random variable $F^{\text{inv}}(U)$ has distribution function F.

What remains is to find the function F^{inv}. From Figure 6.1 we see

$$F(x) = u \quad \Leftrightarrow \quad x = F^{\text{inv}}(u),$$

so if we solve the equation $F(x) = u$ for x, we obtain the expression for $F^{\text{inv}}(u)$.

Exponential random variables

We apply this method to the exponential distribution. On the interval $[0, \infty)$, the $Exp(\lambda)$ distribution function is strictly increasing and given by

$$F(x) = 1 - e^{-\lambda x}.$$

To find F^{inv}, we solve the equation $F(x) = u$:

$$
\begin{aligned}
F(x) = u &\Leftrightarrow & 1 - e^{-\lambda x} &= u \\
&\Leftrightarrow & e^{-\lambda x} &= 1 - u \\
&\Leftrightarrow & -\lambda x &= \ln(1 - u) \\
&\Leftrightarrow & x &= -\frac{1}{\lambda}\ln(1 - u),
\end{aligned}
$$

so $F^{\text{inv}}(u) = -\frac{1}{\lambda}\ln(1-u)$ and if U has a $U(0,1)$ distribution, then the random variable X defined by

$$X = F^{\text{inv}}(U) = -\frac{1}{\lambda}\ln(1 - U)$$

has an $Exp(\lambda)$ distribution.

In practice, one replaces $1-U$ with U, because both have a $U(0,1)$ distribution (see Exercise 6.3). Leaving out the subtraction leads to more efficient computer code. So instead of X we may use

$$Y = -\frac{1}{\lambda}\ln(U),$$

which also has an $Exp(\lambda)$ distribution.

QUICK EXERCISE 6.3 A distribution function F is 0 for $x < 1$ and 1 for $x > 3$, and $F(x) = \frac{1}{4}(x-1)^2$ if $1 \le x \le 3$. Let U be a $U(0,1)$ random variable. Construct a random variable with distribution F from U.

> **Remark 6.1 (The general case).** The restriction we imposed earlier, that the distribution function should be strictly increasing, is not really necessary. Furthermore, a distribution function with jumps or a flat section somewhere in the middle is not a problem either. We illustrate this with an example in Figure 6.2.
> This F has a jump at 4 and so for a corresponding X we should have $P(X = 4) = 0.2$, the size of the jump. We see that whenever U is in the interval $[0.3, 0.5]$, it is mapped to 4 by our method, and that this happens with exactly the right probability!
> The flat section of F between 7 and 8 seems to pose a problem: the equation $F(a) = 0.85$ has as its solution any a between 7 and 8, and we cannot define a unique inverse. This, however, does not really matter, because $P(U = 0.85) = 0$, and we can define the inverse $F^{\text{inv}}(0.85)$ in any way we want. Taking the left endpoint, here the number 7, agrees best with the definition of quantiles (see page 66).
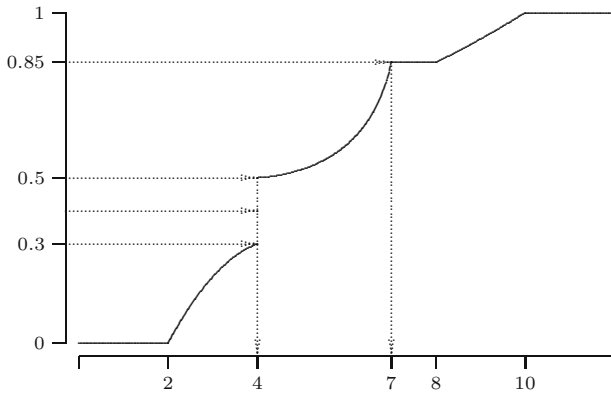
Fig. 6.2. A distribution function with a jump and a flat section.

Remark 6.2 (Existence of random variables). The previous remark supplies a sketchy argument for the fact that any nondecreasing, rightcontinuous function F, with $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$, is the distribution of some random variable.

Generating sequences

For simulations we often want to generate realizations for a large number of random variables. Random number generators have been designed with this purpose in mind: each new call mimics a new $U(0,1)$ random variable. The sequence of numbers thus generated is considered as a realization of a sequence of $U(0,1)$ random variables U_1, U_2, U_3,... with the special property that the events $\{U_i \le a_i\}$ are independent[1] for every choice of the a_i.

6.3 Comparing two jury rules

At the Olympic Games there are several sports events that are judged by a jury, including gymnastics, figure skating, and ice dancing. During the 2002 winter games a dispute arose concerning the gold medal in ice dancing: there were allegations that the Russian team had bribed a French jury member, thereby causing the Russian pair to win just ahead of the Canadians. We look into operating rules for juries, although we leave the effects of bribery to the exercises (Exercise 6.11).

Suppose we have a jury of seven members, and for each performance each juror assigns a grade. The seven grades are to be transformed into a final score. Two rules to do this are under consideration, and we want to choose

---

[1] In Chapter 9 we return to the question of independence between random variables.

the better one. For the first one, the highest and lowest scores are removed and the final score is the average of the remaining five. For the second rule, the scores are put in ascending order and the middle one is assigned as final score. Before you continue reading, consider which rule is better and how you can verify this.

A probabilistic model

For our investigation we assume that the scores the jurors assign deviate by some random amount from the true or deserved score. We model the score that juror i assigns when the performance deserves a score g by

$$Y_i = g + Z_i \qquad \text{for } i = 1, \ldots, 7, \tag{6.2}$$

where $Z_1, \ldots, Z_7$ are random variables with values around zero. Let h_1 and h_2 be functions implementing the two rules:

$$h_1(y_1, \ldots, y_7) = \text{average of the middle five of } y_1, \ldots, y_7,$$
$$h_2(y_1, \ldots, y_7) = \text{middle value of } y_1, \ldots, y_7.$$

We are interested in deviations from the deserved score g:

$$T = h_1(Y_1, \ldots, Y_7) - g,$$
$$M = h_2(Y_1, \ldots, Y_7) - g. \tag{6.3}$$

The distributions of T and M depend on the individual jury grades, and through those, on the juror-deviations $Z_1, Z_2, \ldots, Z_7$, which we model as $U(-0.5, 0.5)$ variables. This more or less finishes the modeling phase: we have given a stochastic model that mimics the workings of a jury and have defined, in terms of the variables in the model, the random variables T and M that represent the errors that result after application of the jury rules.

In any serious application, the model should be *validated*. This means that one tries to gather evidence to convince oneself and others that the model adequately reflects the workings of the real system. In this chapter we are more interested in showing what you can do with simulation once you have a model, so we skip the validation.

The next phase is analysis: which of the deviations is closer to zero? Because T and M are random variables, we would have to clarify what we mean by that, and answering the question certainly involves computing probabilities about T and M. We cannot do this with what we have learned so far, but we know how to simulate, so this is what we do.

Simulation

To generate a realization of a $U(-0.5, 0.5)$ random variable, we only need to subtract 0.5 from the result we obtain from a call to the random generator.

We do this 7 times and insert the resulting values in (6.2) as jury deviations $Z_1, \ldots, Z_7$, and substitute them in equations (6.3) to obtain T and M (the value of g is irrelevant: it drops out of the calculation):

$$
\begin{aligned}
T &= \text{average of the middle five of } Z_1, \ldots, Z_7, \\
M &= \text{middle value of } Z_1, \ldots, Z_7.
\end{aligned}
\tag{6.4}
$$

In simulation terminology, this is called a *run*: we have gone through the whole procedure once, inserting realizations for the random variables. If we repeat the whole procedure, we have a second run; see Table 6.1 for the results of five runs.

Table 6.1. Simulation results for the two jury rules.

Run	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	T	M
1	-0.45	-0.08	-0.38	0.11	-0.42	0.48	0.02	-0.15	-0.08
2	-0.37	-0.18	0.05	-0.10	0.01	0.28	0.31	0.01	0.01
3	0.08	0.07	0.47	-0.21	-0.33	-0.22	-0.48	-0.12	-0.21
4	0.24	0.08	-0.11	0.19	-0.03	0.02	0.44	0.10	0.08
5	0.10	0.18	-0.39	-0.24	-0.36	-0.25	0.20	-0.11	-0.24

QUICK EXERCISE 6.4 The next realizations for $Z_1, \ldots, Z_7$ are: -0.05, 0.26, 0.25, 0.39, 0.22, 0.23, 0.13. Determine the corresponding realizations of T and M.

Table 6.1 can be used to check some computations. We also see that the realization of T was closest to zero in runs 3 and 5, the realization of M was closest to zero in runs 1 and 4, and they were (about) the same in run 2. There is no clear conclusion from this, and even if there was, one could wonder whether the next five runs would yield the same picture. Because the whole process mimics randomness, one has to expect some variation—or perhaps a lot. In later chapters we will get a better understanding of this variation; for the moment we just say that judgment based on a large number of runs is better. We do one thousand runs and exchange the table for pictures. Figure 6.3 depicts, for juror 1, a histogram of all the deviations from the true score g. For each interval of length 0.05 we have counted the number of runs for which the deviation of juror 1 fell in that interval. These numbers vary from about 40 to about 60.

This is just to get an idea about the results for an individual juror. In Figure 6.4 we see histograms for the final scores. Comparing the histograms, it seems that the realizations of T are more concentrated near zero than those of M.
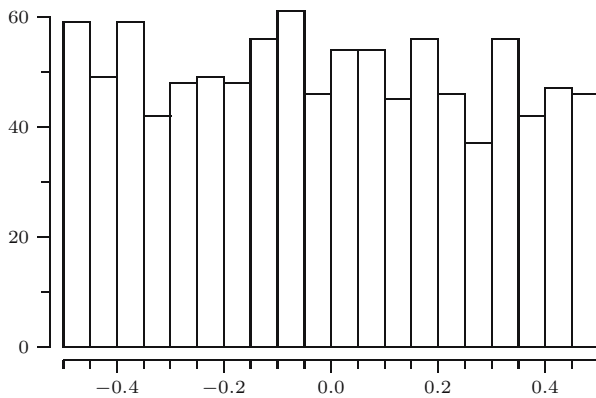
Fig. 6.3. Deviations of juror 1 from the deserved score, one thousand runs.
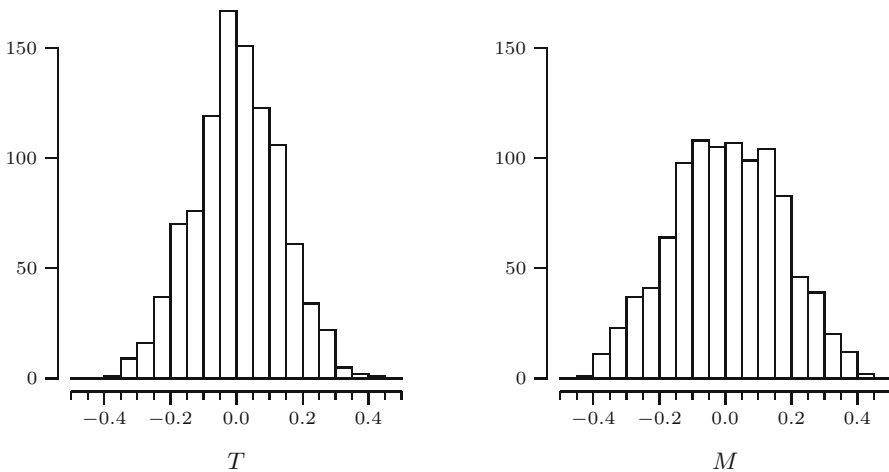
Fig. 6.4. One thousand realizations of T and M.

However, the two histograms do not tell us anything about the relation between T and M, so we plot the realizations of pairs (T, M) for all one thousand runs (Figure 6.5). From this plot we see that in most cases M and T go in the same direction: if T is positive, then usually M is also positive, and the same goes for negative values. In terms of the final scores, both rules generally overvalue and undervalue the performance simultaneously. On closer examination, with help of the line drawn from $(-0.5, -0.5)$ to $(0.5, 0.5)$, we see that the T values tend to be a little closer to zero than the M values.

This suggests that we make a histogram that shows the difference of the absolute deviations from true score. For rule 1 this absolute deviation is $|T|$, for rule 2 it is $|M|$. If the difference $|M| - |T|$ is positive, then T is closer to zero than M, and the difference tells us by how much. A negative difference
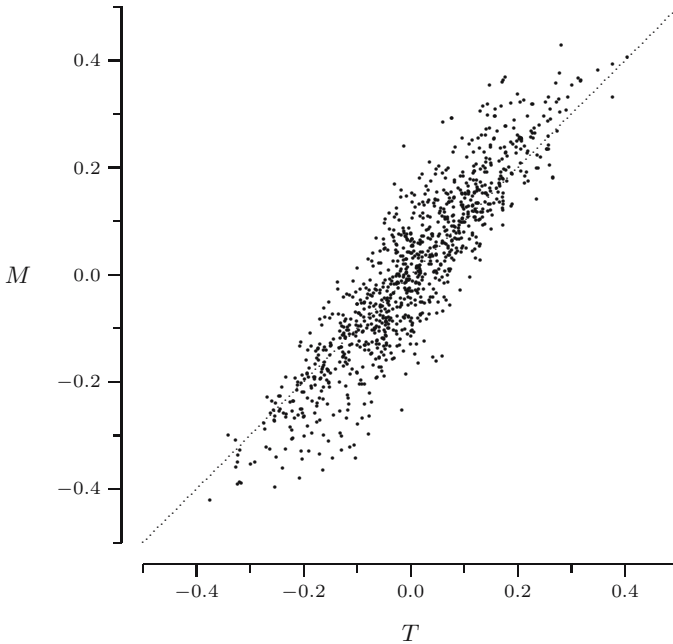
Fig. 6.5. Plot of the points (T, M), one thousand runs.

means that M was closer. In Figure 6.6 all the differences are shown in a histogram. The bars to the right of zero represent 696 runs. So, in about 70% of the runs, rule 1 resulted in a final score that is closer to the true score than rule 2. In about 30% of the cases, rule 2 was better, but generally by a smaller amount, as we see from the histogram.
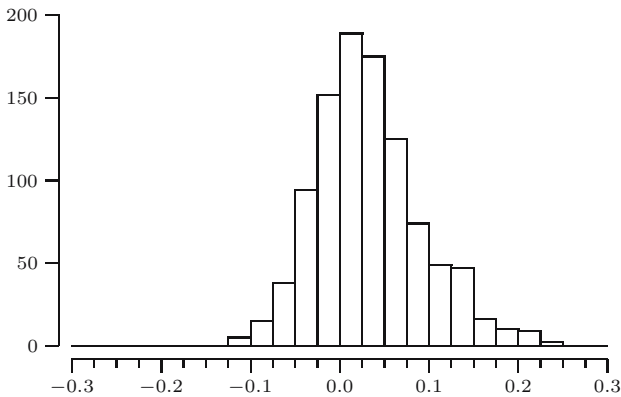
Fig. 6.6. Differences $|M| - |T|$ for one thousand runs.

6.4 The single-server queue

There are many situations in life where you stand in a line waiting for some service: when you want to withdraw money from a cash dispenser, borrow books at the library, be admitted to the emergency room at the hospital, or pump gas at the gas station. Many other queueing situations are hidden: an email message you send might be queued at the local server until it has sent all messages that were submitted ahead of yours; searching the Internet, your browser sends and receives packets of information that are queued at various stages and locations; in assembly lines, partly finished products move from station to station, each time waiting for the next component to be added.

We are going to study one simple queueing model, the so-called single-server queue: it has one *server* or service mechanism, and the arriving customers await their turn in order of their arrival. For definiteness, think of an oasis with one big water well. People arrive at the well with bottles, jerry cans, and other types of containers, to pump water. The supply of water is large, but the pump capacity is limited. The pump is about to be replaced, and while it is clear that a larger pump capacity will result in shorter waiting times, more powerful pumps are also more expensive. Therefore, to prepare a decision that balances costs and benefits, we wish to investigate the relationship between pump capacity and system performance.

Modeling the system

A stochastic model is in order: some general characteristics are known, such as how many people arrive per day and how much water they take on average, but the individual arrival times and amounts are unpredictable. We introduce random variables to describe them: let T_1 be the time between the start at time zero and the arrival of the first customer, T_2 the time between the arrivals of the first and the second customer, T_3 the time between the second and the third, etc.; these are called the *interarrival times*. Let S_i be the length of time that customer i needs to use the pump; in standard terminology this is called the *service time*. This is our description so far:

$$\begin{array}{llll} \text{Arrivals at:} & T_1 & T_1 + T_2 & T_1 + T_2 + T_3 & \text{etc.} \\ \text{Service times:} & S_1 & S_2 & S_3 & \text{etc.} \end{array}$$

The pump capacity v (liters per minute) is not a random variable but a model parameter or decision variable, whose "best" value we wish to determine. So if customer i requires R_i liters of water, then her service time is

$$S_i = \frac{R_i}{v}.$$

To complete the model description, we need to specify the distribution of the random variables T_i and R_i:

Interarrival times: every T_i has an $Exp(0.5)$ distribution (minutes);
Service requirement: every R_i has a $U(2,5)$ distribution (liters).

This particular choice of distributions would have to be supported by evidence that they are suited for the system at hand: a validation step as suggested for the jury model is appropriate here as well. For many arrival type processes, however, the exponential distribution is reasonable as a model for the inter-arrival times (see Chapter 12). The particular uniform distribution chosen for the required amount of water says that all amounts between 2 and 5 liters are equally likely. So there is no sheik who owns a 5000-liter water truck in "our" oasis.

To evaluate system performance, we want to extract from the model the waiting times of the customers and how busy it is at the pump.

Waiting times

Let W_i denote the waiting time of customer i. The first customer is lucky; the system starts empty, and so $W_1 = 0$. For customer i the waiting time depends on how long customer $i-1$ spent in the system compared to the time between their respective arrivals. We see that if the interarrival time T_i is long, relatively speaking, then customer i arrives *after* the departure of customer $i-1$, and so $W_i = 0$:
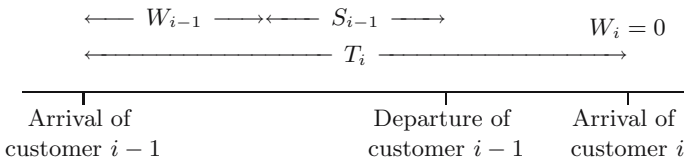
On the other hand, if customer i arrives *before* the departure, the waiting time W_i equals whatever remains of $W_{i-1} + S_{i-1}$:
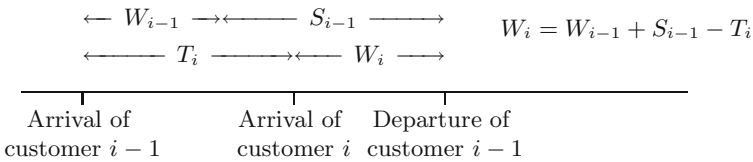
Summarizing the two cases, we see obtain:

$$W_i = \max\{W_{i-1} + S_{i-1} - T_i, 0\}. \tag{6.5}$$

To carry out a simulation, we start at time zero and generate realizations of the interarrival times (the T_i) and service requirements (the R_i) for as long as we want, computing the other quantities that follow from the model on the way. Table 6.2 shows the values generated this way, for two pump capacities ($v = 2$ and 3) for the first six customers. Note that in both cases we use the same realizations of T_i and R_i.

Table 6.2. Results of a short simulation.

	Input realizations			$v = 2$		$v = 3$	
i	T_i	Arr.time	R_i	S_i	W_i	S_i	W_i
1	0.24	0.24	4.39	2.20	0	1.46	0
2	1.97	2.21	4.00	2.00	0.23	1.33	0
3	1.73	3.94	2.33	1.17	0.50	0.78	0
4	2.82	6.76	4.03	2.01	0	1.34	0
5	1.01	7.77	4.17	2.09	1.00	1.39	0.33
6	1.09	8.86	4.24	2.12	1.99	1.41	0.63

QUICK EXERCISE 6.5 The next four realizations are T_7: 1.86; R_7: 4.79; T_8: 1.08; and R_8: 2.33. Complete the corresponding rows of the table.

Longer simulations produce so many numbers that we will drown in them unless we think of something. First, we summarize the waiting times of the first n customers with their average:

$$\bar{W}_n = \frac{W_1 + W_2 + \cdots + W_n}{n}. \tag{6.6}$$

Then, instead of giving a table, we plot the pairs $(n, \bar{W}_n)$, for $n = 1, 2, \ldots$ until the end of the simulation. In Figure 6.7 we see that both lines bounce up and down a bit. Toward the end, the average waiting time for pump capacity 3 is about 0.5 and for $v = 2$ about 2. In a longer simulation we would see each of the averages converge to a limiting value (a consequence of the so-called law of large numbers, the topic of Chapter 13).
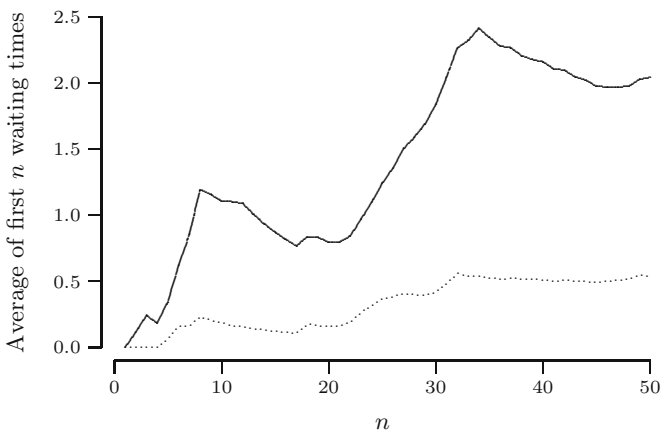
Fig. 6.7. Averaged waiting times at the well, for pump capacity 2 and 3.

Work-in-system

To show how busy it is at the pump one could record how many customers are waiting in the queue and plot this quantity against time. A slightly different approach is to record at every moment how much work there is in the system, that is, how much time it would take to serve everyone present at that moment. For example, if I am halfway through filling my 4-liter jerry can and three persons are waiting who require 2, 3, and 5 liters, respectively, then there are 12 liters to go; at $v = 2$, there is 6 minutes of work in the system, and at $v = 3$ just 4.

The amount of work in the system just before a customer arrives equals the waiting time of that customer, because it is exactly the time it takes to finish the work for everybody ahead of her. The work-in-system at time t tells us how long the wait would be *if* somebody were to arrive at t. For this reason, this quantity is also called the *virtual waiting time*.

Figure 6.8 shows the work-in-system as a function of time for the first 15 minutes, using the same realizations that were the basis for Table 6.2. In the top graph, corresponding to $v = 2$, the work in the system jumps to 2.20 (which is the realization of $R_1/2$) at $t = 0.24$, when the first customer arrives. So at $t = 2.21$, which is 1.97 later, there is $2.20 - 1.97 = 0.23$ minute of work left; this is the waiting time for customer 2, who brings an amount of work of 2.00 minutes, so the peak at 1.97 is $0.23 + 2.00 = 2.23$, etc. In the bottom graph we see the work-in-system reach zero more often, because the individual (work) amounts are $2/3$ of what they are when $v = 2$. More often, arriving
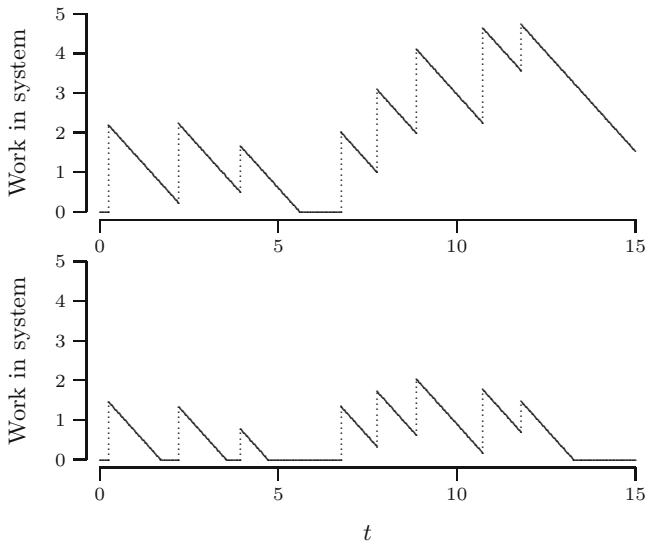
Fig. 6.8. Work in system: top, $v = 2$; bottom, $v = 3$.
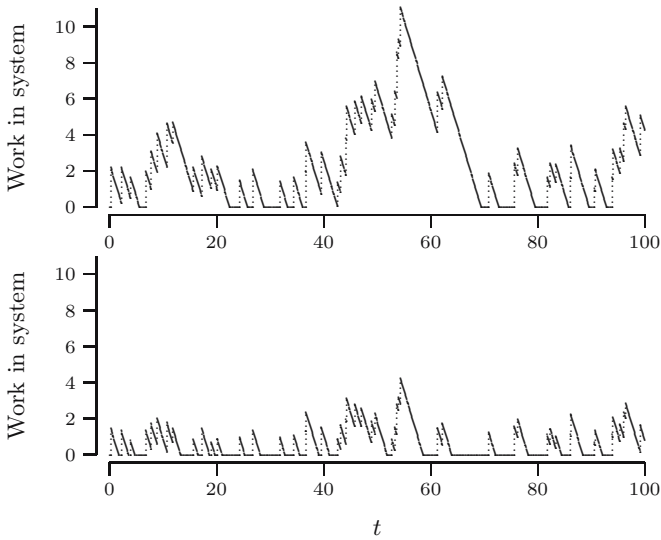
Fig. 6.9. Work in system: top, $v = 2$; bottom, $v = 3$.

customers find the queue empty and the pump not in use; they do not have to wait.

In Figure 6.9 the work-in-system is depicted as a function of time for the first 100 minutes of our run. At pump capacity 2 the virtual waiting time peaks at close to 11 minutes after about 55 minutes, whereas with $v = 3$ the corresponding peak is only about 4 minutes. There also is a marked difference in the proportion of time the system is empty.

6.5 Solutions to the quick exercises

6.1 To simulate the coin, choose any three of the six possible outcomes of the die, report heads if one of these three outcomes turns up, and report tails otherwise. For example, heads if the outcome is odd, tails if it is even.

To simulate the die using a coin is more difficult; one solution is as follows. Toss the coin three times and use the following conversion table to map the result:

Coins	HHH	HHT	HTH	HTT	THH	THT
Die	1	2	3	4	5	6

Repeat the coin tosses if you get TTH or TTT.

6.2 Let the $U(0,1)$ variable be U and set:

$$Y = \begin{cases} 1 & \text{if } U < \frac{3}{5}, \\ 3 & \text{if } \frac{3}{5} \le U < \frac{4}{5}, \\ 4 & \text{if } U \ge \frac{4}{5}. \end{cases}$$

So, for example, $P(Y = 3) = P\left(\frac{3}{5} \le U < \frac{4}{5}\right) = \frac{1}{5}$.

6.3 The given distribution function F is strictly increasing between 1 and 3, so we use the method with F^{inv}. Solve the equation $F(x) = \frac{1}{4}(x-1)^2 = u$ for x. This yields $x = 1 + 2\sqrt{u}$, so we can set $X = 1 + 2\sqrt{U}$. If you need to be convinced, determine F_X.

6.4 In ascending order the values are $-0.05, 0.13, 0.22, 0.23, 0.25, 0.26, 0.39$, so for M we find 0.23, and for T $(0.13 + 0.22 + 0.23 + 0.25 + 0.26)/5 = 0.22$.

6.5 We find:

		Input realizations		$v = 2$		$v = 3$	
i	T_i	Arr.time	R_i	S_i	W_i	S_i	W_i
7	1.86	10.72	4.79	2.39	2.25	1.60	0.18
8	1.08	11.80	2.33	1.16	3.57	0.78	0.70

6.6 Exercises

6.1 Let U have a $U(0,1)$ distribution.

a. Describe how to simulate the outcome of a roll with a die using U.

b. Define Y as follows: round $6U + 1$ down to the nearest integer. What are the possible outcomes of Y and their probabilities?

6.2 ⊡ We simulate the random variable $X = 1 + 2\sqrt{U}$ constructed in Quick exercise 6.3. As realization for U we obtain from the pseudo random generator the number $u = 0.3782739$.

a. What is the corresponding realization x of the random variable X?

b. If the next call to the random generator yields $u = 0.3$, will the corresponding realization for X be larger or smaller than the value you found in **a**?

c. What is the probability the next draw will be smaller than the value you found in **a**?

6.3 Let U have a $U(0,1)$ distribution. Show that $Z = 1 - U$ has a $U(0,1)$ distribution by deriving the probability density function or the distribution function.

6.4 Let F be the distribution function as given in Quick exercise 6.3: $F(x)$ is 0 for $x < 1$ and 1 for $x > 3$, and $F(x) = \frac{1}{4}(x-1)^2$ if $1 \leq x \leq 3$. In the answer it is claimed that $X = 1 + 2\sqrt{U}$ has distribution function F, where U is a $U(0,1)$ random variable. Verify this by computing $P(X \leq a)$ and checking that this equals $F(a)$, for any a.

6.5 ⊞ We have seen that if U has a $U(0,1)$ distribution, then $X = -\ln U$ has an $Exp(1)$ distribution. Check this by verifying that $P(X \leq a) = 1 - e^{-a}$ for $a \geq 0$.

6.6 ⊡ Somebody messed up the random number generator in your computer: instead of uniform random numbers it generates numbers with an $Exp(2)$ distribution. Describe how to construct a $U(0,1)$ random variable U from an $Exp(2)$ distributed X.
Hint: look at how you obtain an $Exp(2)$ random variable from a $U(0,1)$ random variable.

6.7 ⊞ In models for the lifetimes of mechanical components one sometimes uses random variables with distribution functions from the so-called Weibull family. Here is an example: $F(x) = 0$ for $x < 0$, and

$$F(x) = 1 - e^{-5x^2} \qquad \text{for } x \geq 0.$$

Construct a random variable Z with this distribution from a $U(0,1)$ variable.

6.8 A random variable X has a $Par(3)$ distribution, so with distribution function F with $F(x) = 0$ for $x < 1$, and $F(x) = 1 - x^{-3}$ for $x \geq 1$. For details on the Pareto distribution see Section 5.4. Describe how to construct X from a $U(0,1)$ random variable.

6.9 ⊡ In Quick exercise 6.1 we simulated a die by tossing three coins. Recall that we might need several attempts before succeeding.

  a. What is the probability that we succeed on the first try?
  b. Let N be the number of tries that we need. Determine the distribution of N.

6.10 ⊞ There is usually more than one way to simulate a particular random variable. In this exercise we consider two ways to generate geometric random variables.

  a. We give you a sequence of independent $U(0,1)$ random variables U_1, U_2, ... From this sequence, construct a sequence of Bernoulli random vari-

ables. From the sequence of Bernoulli random variables, construct a (single) $Geo(p)$ random variable.

b. It is possible to generate a $Geo(p)$ random variable using just *one* $U(0,1)$ random variable. If calls to the random number generator take a lot of CPU time, this would lead to faster simulation programs. Set $\lambda = -\ln(1 - p)$ and let Y have a $Exp(\lambda)$ distribution. We obtain Z from Y by rounding to the nearest integer *greater* than Y. Note that Z is a discrete random variable, whereas Y is a continuous one. Show that, nevertheless, the event $\{Z > n\}$ is the same as $\{Y > n\}$. Use this to compute $P(Z > n)$ from the distribution of Y. What is the distribution of Z? (See Quick exercise 4.6.)

6.11 Reconsider the jury example (see Section 6.3). Suppose the first jury member is bribed to vote in favor of the present candidate.

a. How should you now model Y_1? Describe how you can investigate which of the two rules is less sensitive to the effect of the bribery.

b. The International Skating Union decided to adopt a rule similar to the following: *randomly* discard two of the jury scores, then average the remaining scores. Describe how to investigate this rule. Do you expect this rule to be more sensitive to the bribery than the two rules already discussed, or less sensitive?

6.12 ⊞ A tiny financial model. To investigate investment strategies, consider the following:

You can choose to invest your money in one particular stock or put it in a savings account. Your initial capital is € 1000. The interest rate r is 0.5% per month and does not change. The initial stock price is € 100. Your stochastic model for the stock price is as follows: next month the price is the same as this month with probability 1/2, with probability 1/4 it is 5% lower, and with probability 1/4 it is 5% higher. This principle applies for every new month. There are no transaction costs when you buy or sell stock.

Your investment strategy for the next 5 years is: convert all your money to stock when the price drops below € 95, and sell all stock and put the money in the bank when the stock price exceeds € 110.

Describe how to simulate the results of this strategy for the model given.

6.13 We give you an unfair coin and you do not know $P(H)$ for this coin. Can you simulate a fair coin, and how many tosses do you need for each fair coin toss?

7

Expectation and variance

Random variables are complicated objects, containing a lot of information on the experiments that are modeled by them. If we want to summarize a random variable by a *single number*, then this number should undoubtedly be its *expected value*. The expected value, also called the *expectation* or *mean*, gives the center—in the sense of average value—of the distribution of the random variable. If we allow a second number to describe the random variable, then we look at its *variance*, which is a measure of spread of the distribution of the random variable.

7.1 Expected values

An oil company needs drill bits in an exploration project. Suppose that it is known that (after rounding to the nearest hour) drill bits of the type used in this particular project will last 2, 3, or 4 hours with probabilities 0.1, 0.7, and 0.2. If a drill bit is replaced by one of the same type each time it has worn out, how long could exploration be continued if in total the company would reserve 10 drill bits for the exploration job? What most people would do to answer this question is to take the weighted average

$$0.1 \cdot 2 + 0.7 \cdot 3 + 0.2 \cdot 4 = 3.1,$$

and conclude that the exploration could continue for 10×3.1, or 31 hours. This weighted average is what we call the *expected value* or *expectation* of the random variable X whose distribution is given by

$$P(X = 2) = 0.1, \quad P(X = 3) = 0.7, \quad P(X = 4) = 0.2.$$

It might happen that the company is unlucky and that each of the 10 drill bits has worn out after two hours, in which case exploration ends after 20 hours. At the other extreme, they may be lucky and drill for 40 hours on these 10

bits. However, it is a mathematical fact that the conclusion about a 31-hour total drilling time is correct in the following sense: for a large number n of drill bits the total running time will be around n times 3.1 hours with high probability. In the example, where $n = 10$, we have, for instance, that drilling will continue for 29, 30, 31, 32, or 33 hours with probability more than 0.86, while the probability that it will last only for 20, 21, 22, 23, or 24 hours is less than 0.00006. We will come back to this in Chapters 13 and 14. This example illustrates the following definition.

DEFINITION. The *expectation* of a discrete random variable X taking the values $a_1, a_2, \ldots$ and with probability mass function p is the number
$$\mathrm{E}[X] = \sum_i a_i \mathrm{P}(X = a_i) = \sum_i a_i p(a_i).$$

We also call $\mathrm{E}[X]$ the *expected value* or *mean* of X. Since the expectation is determined by the probability distribution of X only, we also speak of the expectation or mean of the distribution.

QUICK EXERCISE 7.1 Let X be the discrete random variable that takes the values 1, 2, 4, 8, and 16, each with probability 1/5. Compute the expectation of X.

Looking at an expectation as a weighted average gives a more physical interpretation of this notion, namely as the center of gravity of weights $p(a_i)$ placed at the points a_i. For the random variable associated with the drill bit, this is illustrated in Figure 7.1.
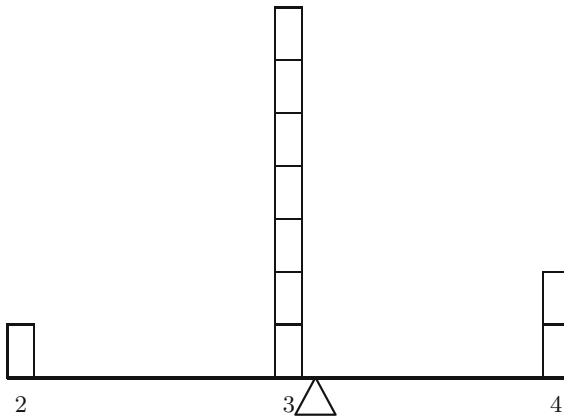
Fig. 7.1. Expected value as center of gravity.

This point of view also leads the way to how one should define the expected value of a continuous random variable. Let, for example, X be a continuous random variable whose probability density function f is zero outside the interval $[0, 1]$. It seems reasonable to approximate X by the *discrete* random variable Y, taking the values

$$\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$$

with as probabilities the masses that X assigns to the intervals $[\frac{k-1}{n}, \frac{k}{n}]$:

$$P\left(Y = \frac{k}{n}\right) = P\left(\frac{k-1}{n} \leq X \leq \frac{k}{n}\right) = \int_{(k-1)/n}^{k/n} f(x)\, dx.$$

We have a good idea of the size of this probability. For large n, it can be approximated well in terms of f:

$$P\left(Y = \frac{k}{n}\right) = \int_{k/n-1/n}^{k/n} f(x)\, dx \approx \frac{1}{n} f\left(\frac{k}{n}\right).$$

The "center-of-gravity" interpretation suggests that the expectation $E[Y]$ of Y should approximate the expectation $E[X]$ of X. We have

$$E[Y] = \sum_{k=1}^{n} \frac{k}{n} P\left(Y = \frac{k}{n}\right) \approx \sum_{k=1}^{n} \frac{k}{n} f\left(\frac{k}{n}\right) \frac{1}{n}.$$

By the definition of a definite integral, for large n the right-hand side is close to

$$\int_{0}^{1} x f(x)\, dx.$$

This motivates the following definition.

> DEFINITION. The *expectation* of a continuous random variable X with probability density function f is the number
>
> $$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

We also call $E[X]$ the *expected value* or *mean* of X. Note that $E[X]$ is indeed the center of gravity of the mass distribution described by the function f:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \frac{\int_{-\infty}^{\infty} x f(x)\, dx}{\int_{-\infty}^{\infty} f(x)\, dx}.$$

This is illustrated in Figure 7.2.
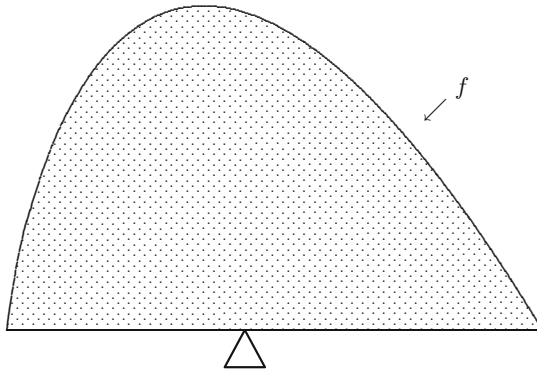
Fig. 7.2. Expected value as center of gravity, continuous case.

QUICK EXERCISE 7.2 Compute the expectation of a random variable U that is uniformly distributed over $[2, 5]$.

Remark 7.1 (The expected value may not exist!). In the definitions in this section we have been rather careless about the *convergence* of sums and integrals. Let us take a closer look at the integral $I = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$. Since a probability density function cannot take negative values, we have $I = I^- + I^+$ with $I^- = \int_{-\infty}^{0} x f(x)\,\mathrm{d}x$ a negative and $I^+ = \int_{0}^{\infty} x f(x)\,\mathrm{d}x$ a positive number. However, it may happen that I^- equals $-\infty$ or I^+ equals $+\infty$. If both $I^- = -\infty$ and $I^+ = +\infty$, then we say that the expected value *does not exist*. An example of a continuous random variable for which the expected value does not exist is the random variable with the Cauchy distribution (see also page 161), having probability density function

$$f(x) = \frac{1}{\pi(1 + x^2)} \quad \text{for } -\infty < x < \infty.$$

For this random variable

$$I^+ = \int_0^{\infty} x \cdot \frac{1}{\pi(1 + x^2)}\,\mathrm{d}x = \left[\frac{1}{2\pi} \ln(1 + x^2)\right]_0^{\infty} = +\infty,$$

$$I^- = \int_{-\infty}^{0} x \cdot \frac{1}{\pi(1 + x^2)}\,\mathrm{d}x = \left[\frac{1}{2\pi} \ln(1 + x^2)\right]_{-\infty}^{0} = -\infty.$$

If I^- is finite but $I^+ = +\infty$, then we say that the expected value is infinite. A distribution that has an infinite expectation is the Pareto distribution with parameter $\alpha = 1$ (see Exercise 7.11). The remarks we made on the integral in the definition of $E[X]$ for continuous X apply similarly to the *sum* in the definition of $E[X]$ for discrete random variables X.

7.2 Three examples

The geometric distribution

If you buy a lottery ticket every week and you have a chance of 1 in 10 000 of winning the jackpot, what is the *expected* number of weeks you have to buy tickets before you get the jackpot? The answer is: 10 000 weeks (almost two centuries!). The number of weeks is modeled by a random variable with a geometric distribution with parameter $p = 10^{-4}$.

THE EXPECTATION OF A GEOMETRIC DISTRIBUTION. Let X have a geometric distribution with parameter p; then

$$E[X] = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = \frac{1}{p}.$$

Here $\sum_{k=1}^{\infty} kp(1-p)^{k-1} = 1/p$ follows from the formula $\sum_{k=1}^{\infty} kx^{k-1} = 1/(1-x)^2$ that has been derived in your calculus course. We will see a simple (probabilistic) way to obtain the value of this sum in Chapter 11.

The exponential distribution

In Section 5.6 we considered the chemical reactor example, where the residence time T, measured in minutes, has an $Exp(0.5)$ distribution. We claimed that this implies that the mean time a particle stays in the vessel is 2 minutes. More generally, we have the following.

THE EXPECTATION OF AN EXPONENTIAL DISTRIBUTION. Let X have an exponential distribution with parameter λ; then

$$E[X] = \int_0^{\infty} x\lambda e^{-\lambda x}\, dx = \frac{1}{\lambda}.$$

The integral has been determined in your calculus course (with the technique of integration by parts).

The normal distribution

Here, using that the normal density integrates to 1 and applying the substitution $z = (x - \mu)/\sigma$,

$$E[X] = \int_{-\infty}^{\infty} x\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\, dx = \mu + \int_{-\infty}^{\infty} (x-\mu)\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\, dx$$

$$= \mu + \sigma\int_{-\infty}^{\infty} z\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}\, dz = \mu,$$

where the integral is 0, because the integrand is an odd function. We obtained the following rule.

THE EXPECTATION OF A NORMAL DISTRIBUTION. Let X be an $N(\mu, \sigma^2)$ distributed random variable. Then

$$E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \mu.$$

7.3 The change-of-variable formula

Often one does not want to compute the expected value of a random variable X but rather of a function of X, as, for example, X^2. We then need to determine the distribution of $Y = X^2$, for example by computing the distribution function F_Y of Y (this is an example of the general problem of how distributions change under transformations—this topic is the subject of Chapter 8). For a concrete example, suppose an architect wants maximal variety in the sizes of buildings: these should be of the same width and depth X, but X is uniformly distributed between 0 and 10 meters. What is the distribution of the area X^2 of a building; in particular, will this distribution be (anything near to) uniform? Let us compute F_Y; for $0 \le a \le 100$:

$$F_Y(a) = P(X^2 \le a) = P(X \le \sqrt{a}) = \frac{\sqrt{a}}{10}.$$

Hence the probability density function f_Y of the area is, for $0 < y < 100$ meters squared, given by

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} \frac{\sqrt{y}}{10} = \frac{1}{20\sqrt{y}}. \tag{7.1}$$

This means that the buildings with small areas are heavily overrepresented, because f_Y explodes near 0—see also Figure 7.3, in which we plotted f_Y.

Surprisingly, this is not very visible in Figure 7.4, an example where we should believe our calculations more than our eyes. In the figure the locations of the buildings are generated by a Poisson process, the subject of Chapter 12. Suppose that a contractor has to make an offer on the price of the foundations of the buildings. The amount of concrete he needs will be proportional to the area X^2 of a building. So his problem is: what is the expected area of a building? With f_Y from (7.1) he finds

$$E[X^2] = E[Y] = \int_0^{100} y \cdot \frac{1}{20\sqrt{y}} dy = \int_0^{100} \frac{\sqrt{y}}{20} dy = \left[\frac{1}{20}\frac{2}{3}y^{\frac{3}{2}}\right]_0^{100} = 33\frac{1}{3} \text{ m}^2.$$
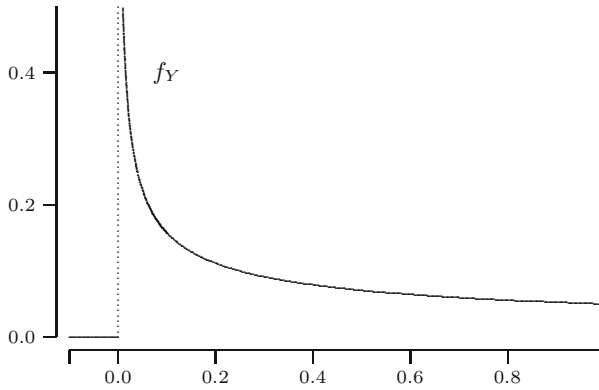
Fig. 7.3. The probability density of the square of a $U(0,10)$ random variable.

It is interesting to note that we really *need* to do this calculation, because the expected area is *not* simply the product of the expected width and the expected depth, which is $25 \, \text{m}^2$. However, there is a much easier way in which the contractor could have obtained this result. He could have argued that the value of the *area* is x^2 when x is the width, and that he should take the weighted average of *those* values, where the weight at width x is given by the value $f_X(x)$ of the probability density of X. Then he would have computed

$$\mathrm{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 f_X(x) \, \mathrm{d}x = \int_0^{10} x^2 \cdot \frac{1}{10} \, \mathrm{d}x = \left[\frac{1}{30} x^3\right]_0^{10} = 33\tfrac{1}{3} \, \text{m}^2.$$

It is indeed a mathematical theorem that this is *always* a correct way to compute expected values of functions of random variables.
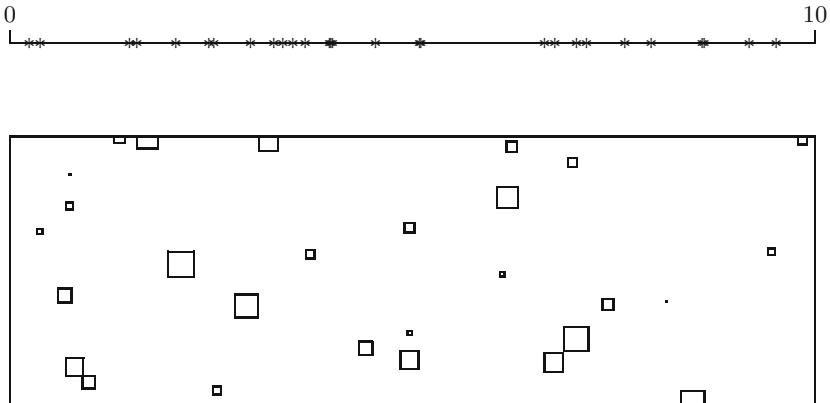
Fig. 7.4. Top: widths of the buildings between 0 and 10 meters. Bottom: corresponding buildings in a 100×300 m area.

THE CHANGE-OF-VARIABLE FORMULA. Let X be a random variable, and let $g : \mathbb{R} \to \mathbb{R}$ be a function.
If X is discrete, taking the values $a_1, a_2, \ldots$, then

$$E[g(X)] = \sum_i g(a_i)P(X = a_i).$$

If X is continuous, with probability density function f, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

QUICK EXERCISE 7.3 Let X have a $Ber(p)$ distribution. Compute $E\left[2^X\right]$.

An operation that occurs very often in practice is a change of units, e.g., from Fahrenheit to Celsius. What happens then to the expectation? Here we have to apply the formula with the function $g(x) = rx + s$, where r and s are real numbers. When X has a continuous distribution, the change-of-variable formula yields:

$$
\begin{aligned}
E[rX + s] &= \int_{-\infty}^{\infty} (rx + s)f(x)\,dx \\
&= r\int_{-\infty}^{\infty} xf(x)\,dx + s\int_{-\infty}^{\infty} f(x)\,dx \\
&= rE[X] + s.
\end{aligned}
$$

A similar computation with integrals replaced by sums gives the same result for discrete random variables.

7.4 Variance

Suppose you are offered an opportunity for an investment whose expected return is € 500. If you are given the extra information that this expected value is the result of a 50% chance of a € 450 return and a 50% chance of a € 550 return, then you would not hesitate to spend € 450 on this investment. However, if the expected return were the result of a 50% chance of a € 0 return and a 50% chance of a € 1000 return, then most people would be reluctant to spend such an amount. This demonstrates that the spread (around the mean) of a random variable is of great importance. Usually this is measured by the expected squared deviation from the mean.

DEFINITION.   The *variance* $\mathrm{Var}(X)$ of a random variable X is the number
$$\mathrm{Var}(X) = E\left[(X - E[X])^2\right].$$

Note that the variance of a random variable is always positive (or 0). Furthermore, there is the question of existence and finiteness (cf. Remark 7.1). In practical situations one often considers the *standard deviation* defined by $\sqrt{\text{Var}(X)}$, because it has the same dimension as $\text{E}[X]$.

As an example, let us compute the variance of a normal distribution. If X has an $N(\mu, \sigma^2)$ distribution, then:

$$\text{Var}(X) = \text{E}\left[(X - \text{E}[X])^2\right]$$

$$= \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, dx$$

$$= \sigma^2 \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2} \, dz.$$

Here we substituted $z = (x - \mu)/\sigma$. Using integration by parts one finds that

$$\int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2} \, dz = 1.$$

We have found the following property.

> VARIANCE OF A NORMAL DISTRIBUTION. Let X be an $N(\mu, \sigma^2)$ distributed random variable. Then
>
> $$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, dx = \sigma^2.$$

QUICK EXERCISE 7.4 Let us call the two returns discussed above Y_1 and Y_2, respectively. Compute the variance and standard deviation of Y_1 and Y_2.

It is often not practical to compute $\text{Var}(X)$ directly from the definition, but one uses the following rule.

> AN ALTERNATIVE EXPRESSION FOR THE VARIANCE. For any random variable X,
>
> $$\text{Var}(X) = \text{E}\left[X^2\right] - (\text{E}[X])^2.$$

To see that this rule holds, we apply the change-of-variable formula. Suppose X is a continuous random variable with probability density function f (the discrete case runs completely analogously). Using the change-of-variable formula, well-known properties of the integral, and $\int_{-\infty}^{\infty} f(x)\, dx = 1$, we find

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$$

$$= \int_{-\infty}^{\infty} (x - \mathrm{E}[X])^2 f(x)\,\mathrm{d}x$$

$$= \int_{-\infty}^{\infty} \left(x^2 - 2x\mathrm{E}[X] + (\mathrm{E}[X])^2\right) f(x)\,\mathrm{d}x$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - 2\mathrm{E}[X]\int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x + (\mathrm{E}[X])^2 \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$$

$$= \mathrm{E}\left[X^2\right] - 2(\mathrm{E}[X])^2 + (\mathrm{E}[X])^2$$

$$= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2.$$

With this rule we make two steps: first we compute $\mathrm{E}[X]$, then we compute $\mathrm{E}[X^2]$. The latter is called *the second moment* of X. Let us compare the computations, using the definition and this rule for the drill bit example. Recall that for this example X takes the values $2, 3$, and 4 with probabilities $0.1, 0.7$, and 0.2. We found that $\mathrm{E}[X] = 3.1$. According to the definition

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - 3.1)^2\right] = 0.1 \cdot (2 - 3.1)^2 + 0.7 \cdot (3 - 3.1)^2 + 0.2 \cdot (4 - 3.1)^2$$

$$= 0.1 \cdot (-1.1)^2 + 0.7 \cdot (-0.1)^2 + 0.2 \cdot (0.9)^2$$

$$= 0.1 \cdot 1.21 + 0.7 \cdot 0.01 + 0.2 \cdot 0.81$$

$$= 0.121 + 0.007 + 0.162$$

$$= 0.29.$$

Using the rule is neater and somewhat faster:

$$\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (3.1)^2 = 0.1 \cdot 2^2 + 0.7 \cdot 3^2 + 0.2 \cdot 4^2 - 9.61$$

$$= 0.1 \cdot 4 + 0.7 \cdot 9 + 0.2 \cdot 16 - 9.61$$

$$= 0.4 + 6.3 + 3.2 - 9.61$$

$$= 0.29.$$

What happens to the variance if we change units? At the end of the previous section we showed that $\mathrm{E}[rX + s] = r\mathrm{E}[X] + s$. This can be used to obtain the corresponding rule for the variance under change of units (see also Exercise 7.15).

> EXPECTATION AND VARIANCE UNDER CHANGE OF UNITS. For any random variable X and any real numbers r and s,
>
> $$\mathrm{E}[rX + s] = r\mathrm{E}[X] + s, \quad \text{and} \quad \mathrm{Var}(rX + s) = r^2\mathrm{Var}(X).$$

Note that the variance is insensitive to the shift over s. Can you understand why this must be true without doing any computations?

7.5 Solutions to the quick exercises

7.1 We have

$$\mathrm{E}[X] = \sum_i a_i \mathrm{P}(X = a_i) = 1 \cdot \frac{1}{5} + 2 \cdot \frac{1}{5} + 4 \cdot \frac{1}{5} + 8 \cdot \frac{1}{5} + 16 \cdot \frac{1}{5} = \frac{31}{5} = 6.2.$$

7.2 The probability density function f of U is given by $f(x) = 0$ outside $[2, 5]$ and $f(x) = 1/3$ for $2 \le x \le 5$; hence

$$\mathrm{E}[U] = \int_{-\infty}^{\infty} x f(x) \, \mathrm{d}x = \int_2^5 \frac{1}{3} x \, \mathrm{d}x = \left[\frac{1}{6} x^2 \right]_2^5 = 3\tfrac{1}{2}.$$

7.3 Using the change-of-variable formula we obtain

$$\mathrm{E}\left[2^X\right] = \sum_i 2^{a_i} \mathrm{P}(X = a_i)$$
$$= 2^0 \cdot \mathrm{P}(X = 0) + 2^1 \cdot \mathrm{P}(X = 1)$$
$$= 1 \cdot (1 - p) + 2 \cdot p = 1 - p + 2p = 1 + p.$$

You could also have noted that $Y = 2^X$ has a distribution given by $\mathrm{P}(Y = 1) = 1 - p, \mathrm{P}(Y = 2) = p$; hence

$$\mathrm{E}\left[2^X\right] = \mathrm{E}[Y] = 1 \cdot \mathrm{P}(Y = 1) + 2 \cdot \mathrm{P}(Y = 2) = 1 \cdot (1 - p) + 2 \cdot p = 1 + p.$$

7.4 We have

$$\mathrm{Var}(Y_1) = \tfrac{1}{2}(450 - 500)^2 + \tfrac{1}{2}(550 - 500)^2 = 50^2 = 2500,$$

so Y_1 has standard deviation € 50 and

$$\mathrm{Var}(Y_2) = \tfrac{1}{2}(0 - 500)^2 + \tfrac{1}{2}(1000 - 500)^2 = 500^2 = 250\,000,$$

so Y_2 has standard deviation € 500.

7.6 Exercises

7.1 ⊡ Let T be the outcome of a roll with a fair die.

 a. Describe the probability distribution of T, that is, list the outcomes and the corresponding probabilities.
 b. Determine $\mathrm{E}[T]$ and $\mathrm{Var}(T)$.

7.2 ⊡ The probability distribution of a discrete random variable X is given by

$$\mathrm{P}(X = -1) = \tfrac{1}{5}, \quad \mathrm{P}(X = 0) = \tfrac{2}{5}, \quad \mathrm{P}(X = 1) = \tfrac{2}{5}.$$

a. Compute $E[X]$.

b. Give the probability distribution of $Y = X^2$ and compute $E[Y]$ using the distribution of Y.

c. Determine $E[X^2]$ using the change-of-variable formula. Check your answer against the answer in **b**.

d. Determine $\text{Var}(X)$.

7.3 For a certain random variable X it is known that $E[X] = 2$, $\text{Var}(X) = 3$. What is $E[X^2]$?

7.4 Let X be a random variable with $E[X] = 2$, $\text{Var}(X) = 4$. Compute the expectation and variance of $3 - 2X$.

7.5 ⊡ Determine the expectation and variance of the $Ber(p)$ distribution.

7.6 ⊞ The random variable Z has probability density function $f(z) = 3z^2/19$ for $2 \leq z \leq 3$ and $f(z) = 0$ elsewhere. Determine $E[Z]$. Before you do the calculation: will the answer lie closer to 2 than to 3 or the other way around?

7.7 Given is a random variable X with probability density function f given by $f(x) = 0$ for $x < 0$, and for $x > 1$, and $f(x) = 4x - 4x^3$ for $0 \leq x \leq 1$. Determine the expectation and variance of the random variable $2X + 3$.

7.8 ⊡ Given is a continuous random variable X whose distribution function F satisfies $F(x) = 0$ for $x < 0$, $F(x) = 1$ for $x > 1$, and $F(x) = x(2 - x)$ for $0 \leq x \leq 1$. Determine $E[X]$.

7.9 Let U be a random variable with a $U(\alpha, \beta)$ distribution.

a. Determine the expectation of U.

b. Determine the variance of U.

7.10 ⊡ Let X have an exponential distribution with parameter λ.

a. Determine $E[X]$ and $E[X^2]$ using partial integration.

b. Determine $\text{Var}(X)$.

7.11 ⊡ In this exercise we take a look at the mean of a Pareto distribution.

a. Determine the expectation of a $Par(2)$ distribution.

b. Determine the expectation of a $Par(\frac{1}{2})$ distribution.

c. Let X have a $Par(\alpha)$ distribution. Show that $E[X] = \alpha/(\alpha - 1)$ if $\alpha > 1$.

7.12 For which α is the variance of a $Par(\alpha)$ distribution finite? Compute the variance for these α.

7.13 Remember that we found on page 95 that the expected area of a building was $33\frac{1}{3}$ m^2, whereas the square of the expected width was only 25 m^2. This phenomenon is more general: show that for any random variable X one has $\mathrm{E}\left[X^2\right] \geq \left(\mathrm{E}[X]\right)^2$.
Hint: you might use that $\mathrm{Var}(X) \geq 0$.

7.14 Suppose we choose arbitrarily a point from the square with corners at (2,1), (3,1), (2,2), and (3,2). The random variable A is the area of the triangle with its corners at (2,1), (3,1), and the chosen point. (See also Exercise 5.9 and Figure 7.5.) Compute $\mathrm{E}[A]$.

$(2,2)$ $\qquad\qquad\qquad\qquad\qquad\qquad$ $(3,2)$

randomly chosen point

A

$(2,1)$ $\qquad\qquad\qquad\qquad\qquad\qquad$ $(3,1)$
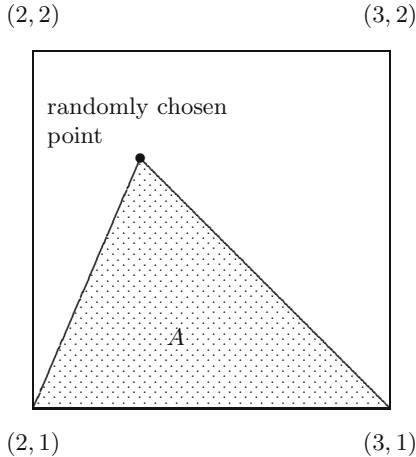
Fig. 7.5. A triangle in a 1×1 square.

7.15 ⊞ Let X be a random variable and r and s any real numbers. Use the change-of-units rule $\mathrm{E}[rX + s] = r\mathrm{E}[X] + s$ for the expectation to obtain **a** and **b**.

  a. Show that $\mathrm{Var}(rX) = r^2\mathrm{Var}(X)$.
  b. Show that $\mathrm{Var}(X + s) = \mathrm{Var}(X)$.
  c. Combine parts **a** and **b** to show that

$$\mathrm{Var}(rX + s) = r^2\mathrm{Var}(X).$$

7.16 ⊡ The probability density function f of the random variable X used in Figure 7.2 is given by $f(x) = 0$ outside $(0,1)$ and $f(x) = -4x\ln(x)$ for $0 < x < 1$. Compute the position of the balancing point in the figure, that is, compute the expectation of X.

7.17 ⊞ Let U be a discrete random variable taking the values $a_1, \ldots, a_r$ with probabilities $p_1, \ldots, p_r$.

  a. Suppose all $a_i \geq 0$, but that $\mathrm{E}[U]{=}0$. Show then

$$a_1 = a_2 = \cdots = a_r = 0.$$

In other words; $P(U = 0) = 1$.

b. Suppose that V is a random variable taking the values $b_1, \ldots, b_r$ with probabilities $p_1, \ldots, p_r$. Show that $\mathrm{Var}(V) = 0$ implies

$$P(V = E[V]) = 1.$$

Hint: apply **a** with $U = (V - E[V])^2$.

Computations with random variables

There are many ways to make new random variables from old ones. Of course this is not a goal in itself; usually new variables are created naturally in the process of solving a practical problem. The expectations and variances of such new random variables can be calculated with the change-of-variable formula. However, often one would like to know the *distributions* of the new random variables. We shall show how to determine these distributions, how to compare expectations of random variables and their transformed versions (Jensen's inequality), and how to determine the distributions of maxima and minima of several random variables.

8.1 Transforming discrete random variables

The problem we consider in this section and the next is how the distribution of a random variable X changes if we apply a function g to it, thus obtaining a new random variable Y:
$$Y = g(X).$$

When X is a discrete random variable this is usually not too hard to do: it is just a matter of bookkeeping. We illustrate this with an example. Imagine an airline company that sells tickets for a flight with 150 available seats. It has no idea about how many tickets it will sell. Suppose, to keep the example simple, that the number X of tickets that will be sold can be anything from 1 to 200. Moreover, suppose that each possibility has equal probability to occur, i.e., $P(X = j) = 1/200$ for $j = 1, 2, \ldots, 200$. The real interest of the airline company is in the random variable Y, which is the number of passengers that have to be refused. What is the distribution of Y? To answer this, note that nobody will be refused when the passengers fit in the plane, hence

$$P(Y = 0) = P(X \le 150) = \frac{150}{200} = \frac{3}{4}.$$

For the other values, $k = 1, 2 \ldots, 50$

$$P(Y = k) = P(X = 150 + k) = \frac{1}{200}.$$

Note that in this example the function g is given by $g(x) = \max\{x - 150, 0\}$.

QUICK EXERCISE 8.1 Let Z be the number of passengers that will be in the plane. Determine the probability distribution of Z. What is the function g in this case?

8.2 Transforming continuous random variables

We now turn to continuous random variables. Since single values occur with probability zero for a continuous random variable, the approach above does not work. The strategy now is to *first* determine the distribution function of the transformed random variable $Y = g(X)$ and then the probability density by differentiating. We shall illustrate this with the following example (actually we saw an example of such a computation in Section 7.3 with the function $g(x) = x^2$).

We consider two methods that traffic police employ to determine whether you deserve a fine for speeding. From experience, the traffic police think that vehicles are driving at speeds ranging from 60 to 90 km/hour at a certain road section where the speed limit is 80 km/hour. They assume that the speed of the cars is uniformly distributed over this interval. The first method is measuring the speed at a fixed spot in the road section. With this method the police will find that about $(90 - 80)/(90 - 60) = 1/3$ of the cars will be fined.

For the second method, cameras are put at the beginning and end of a 1-km road section, and a driver is fined if he spends less than a certain amount of time in the road section. Cars driving at 60 km/hour need one minute, those driving at 90 km/hour only 40 seconds. Let us therefore model the time T an arbitrary car spends in the section by a uniform distribution over $(40, 60)$ seconds. What is the speed V we deduce from this travelling time? Note that for $40 \leq t \leq 60$,

$$P(T \leq t) = \frac{t - 40}{20}.$$

Since there are 3600 seconds in an hour we have that

$$V = g(T) = \frac{3600}{T}.$$

We therefore find for the distribution function $F_V(v) = P(V \leq v)$ of the speed V that

$$F_V(v) = \mathrm{P}\left(\frac{3600}{T} \le v\right) = \mathrm{P}\left(T \ge \frac{3600}{v}\right) = 1 - \frac{(3600/v) - 40}{20} = 3 - \frac{180}{v}$$

for all speeds v between 60 and 90. We can now obtain the probability density f_V of V by differentiating:

$$f_V(v) = \frac{\mathrm{d}}{\mathrm{d}v}F_V(v) = \frac{\mathrm{d}}{\mathrm{d}v}\left(3 - \frac{180}{v}\right) = \frac{180}{v^2}$$

for $60 \le v \le 90$.

It is amusing to note that with the second model the traffic police write fewer speeding tickets because

$$\mathrm{P}(V > 80) = 1 - \mathrm{P}(V \le 80) = 1 - \left(3 - \frac{180}{80}\right) = \frac{1}{4}.$$

(With the first model we found probability $1/3$ that a car drove faster than 80 km/hour.) This is related to a famous result in road traffic research, which is succinctly phrased as: "space mean speed $<$ time mean speed" (see [37]). It is also related to Jensen's inequality, which we introduce in Section 8.3.

Similar to the way this is done in the traffic example, one can determine the distribution of $Y = 1/X$ for any X with a continuous distribution. The outcome will be that if X has density f_X, then the density f_Y of Y is given by

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_Y(y) = \frac{1}{y^2}f_X\left(\frac{1}{y}\right) \qquad \text{for } y < 0 \text{ and } y > 0.$$

One can give $f_Y(0)$ any value; often one puts $f_Y(0) = 0$.

QUICK EXERCISE 8.2 Let X have a continuous distribution with probability density $f_X(x) = 1/[\pi(1 + x^2)]$. What is the distribution of $Y = 1/X$?

We turn to a second example. A very common transformation is a change of units, for instance, from Celsius to Fahrenheit. If X is temperature expressed in degrees Celsius, then $Y = \frac{9}{5}X + 32$ is the temperature in degrees Fahrenheit. Let F_X and F_Y be the distribution functions of X and Y. Then we have for any a

$$F_Y(a) = \mathrm{P}(Y \le a) = \mathrm{P}\left(\frac{9}{5}X + 32 \le a\right)$$

$$= \mathrm{P}\left(X \le \frac{5}{9}(a - 32)\right) = F_X\left(\frac{5}{9}(a - 32)\right).$$

By differentiating F_Y (using the chain rule), we obtain the probability density $f_Y(y) = \frac{5}{9}f_X\left(\frac{5}{9}(y - 32)\right)$. We can do this for more general changes of units, and we obtain the following useful rule.

CHANGE-OF-UNITS TRANSFORMATION. Let X be a continuous random variable with distribution function F_X and probability density function f_X. If we change units to $Y = rX + s$ for real numbers $r > 0$ and s, then

$$F_Y(y) = F_X\left(\frac{y-s}{r}\right) \quad \text{and} \quad f_Y(y) = \frac{1}{r}f_X\left(\frac{y-s}{r}\right).$$

As an example, let X be a random variable with an $N(\mu, \sigma^2)$ distribution, and let $Y = rX + s$. Then this rule gives us

$$f_Y(y) = \frac{1}{r}f_X\left(\frac{y-s}{r}\right) = \frac{1}{r\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}((y-r\mu-s)/r\sigma)^2}$$

for $-\infty < y < \infty$. On the right-hand side we recognize the probability density of a normal distribution with parameters $r\mu + s$ and $r^2\sigma^2$. This illustrates the following rule.

NORMAL RANDOM VARIABLES UNDER CHANGE OF UNITS. Let X be a random variable with an $N(\mu, \sigma^2)$ distribution. For any $r \neq 0$ and any s, the random variable $rX + s$ has an $N(r\mu + s, r^2\sigma^2)$ distribution.

Note that if X has an $N(\mu, \sigma^2)$ distribution, then with $r = 1/\sigma$ and $s = -\mu/\sigma$ we conclude that

$$Z = \frac{1}{\sigma}X + \left(-\frac{\mu}{\sigma}\right) = \frac{X-\mu}{\sigma}$$

has an $N(0, 1)$ distribution. As a consequence

$$F_X(a) = P(X \le a) = P(\sigma Z + \mu \le a) = P\left(Z \le \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right).$$

So any probability for an $N(\mu, \sigma^2)$ distributed random variable X can be expressed in terms of an $N(0, 1)$ distributed random variable Z.

QUICK EXERCISE 8.3 Compute the probabilities $P(X \le 5)$ and $P(X \ge 2)$ for X with an $N(4, 25)$ distribution.

8.3 Jensen's inequality

Without actually computing the distribution of $g(X)$ we can often tell how $E[g(X)]$ relates to $g(E[X])$. For the change-of-units transformation $g(x) = rx + s$ we know that $E[g(X)] = g(E[X])$ (see Section 7.3). It is a common

error to equate these two sides for *other* functions g. In fact, equality will *very rarely* occur for nonlinear g.

For example, suppose that a company that produces microelectronic parts has a target production of 240 chips per day, but the yield has only been 40, 60, and 80 chips on three consecutive days. The average production over the three days then is 60 chips, so on average the production should have been 4 times higher to reach the target. However, one can also look at this in the following way: on the three days the production should have been $240/40 = 6$, $240/60 = 4$, and $240/80 = 3$ times higher. On average that is

$$\tfrac{1}{3}(6 + 4 + 3) = \tfrac{13}{3} = 4.3333$$

times higher! What happens here can be explained (take for X the part of the target production that is realized, where you give equal probabilities to the three outcomes $1/6$, $1/4$, and $1/3$) by the fact that if X is a random variable taking positive values, then always

$$\frac{1}{\mathrm{E}\left[X\right]} < \mathrm{E}\left[\frac{1}{X}\right],$$

unless $\mathrm{Var}(X) = 0$, which only happens if X is not random at all (cf. Exercise 7.17). This inequality is the case $g(x) = 1/x$ on $(0, \infty)$ of the following result that holds for general convex functions g.

> JENSEN'S INEQUALITY.   Let g be a convex function, and let X be a random variable. Then
>
> $$g(\mathrm{E}\left[X\right]) \leq \mathrm{E}\left[g(X)\right].$$

Recall from calculus that a twice differentiable function g is *convex* on an interval I if $g''(x) \geq 0$ for all x in I, and *strictly* convex if $g''(x) > 0$ for all x in I. When X takes its values in an interval I (this can, for instance, be $I = (-\infty, \infty)$), and g is strictly convex on I, then *strict* inequality holds: $g(\mathrm{E}\left[X\right]) < \mathrm{E}\left[g(X)\right]$, unless X is not random.

In Figure 8.1 we illustrate the way in which this result can be obtained for the special case of a random variable X that takes two values, a and b. In the figure, X takes these two values with probability $3/4$ and $1/4$ respectively. Convexity of g forces any line segment connecting two points on the graph of g to lie above the part of the graph between these two points. So if we choose the line segment from $(a, g(a))$ to $(b, g(b))$, then it follows that the point

$$(\mathrm{E}\left[X\right], \mathrm{E}\left[g(X)\right]) = \left(\tfrac{3}{4}a + \tfrac{1}{4}b, \tfrac{3}{4}g(a) + \tfrac{1}{4}g(b)\right) = \tfrac{3}{4}(a, g(a)) + \tfrac{1}{4}(b, g(b))$$

on this line lies "above" the point $(\mathrm{E}\left[X\right], g(\mathrm{E}\left[X\right])$ on the graph of g. Hence $\mathrm{E}\left[g(X)\right] \geq g(\mathrm{E}\left[X\right])$.
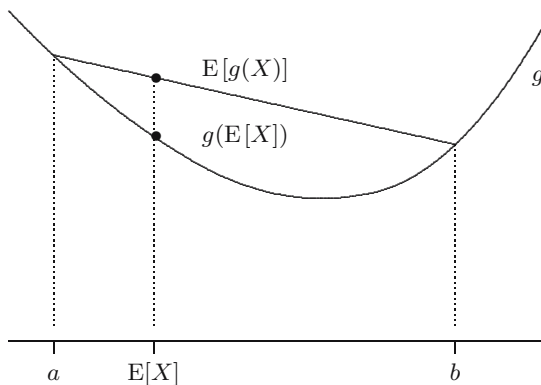
Fig. 8.1. Jensen's inequality.

A simple example is given by $g(x) = x^2$. This function is convex ($g''(x) = 2$ for all x), and hence

$$(\mathrm{E}[X])^2 \le \mathrm{E}\left[X^2\right].$$

Note that this is exactly the same as saying that $\mathrm{Var}(X) \ge 0$, which we have already seen in Section 7.4.

QUICK EXERCISE 8.4 Let X be a random variable with $\mathrm{Var}(X) > 0$. Which is true: $\mathrm{E}\left[e^{-X}\right] < e^{-\mathrm{E}[X]}$ or $\mathrm{E}\left[e^{-X}\right] > e^{-\mathrm{E}[X]}$?

8.4 Extremes

In many situations the maximum (or minimum) of a sequence $X_1, X_2, \ldots, X_n$ of random variables is the variable of interest. For instance, let $X_1, X_2, \ldots, X_{365}$ be the water level of a river during the days of a particular year for a particular location. Suppose there will be flooding if the level exceeds a certain height—usually the height of the dykes. The question whether flooding occurs during a year is completely answered by looking at the maximum of $X_1, X_2, \ldots, X_{365}$. If one wants to predict occurrence of flooding in the future, the probability distribution of this maximum is of great interest. Similar models arise, for instance, when one is interested in possible damage from a series of shocks or in the extent of a contamination plume in the subsurface.

We want to find the distribution of the random variable

$$Z = \max\{X_1, X_2, \ldots, X_n\}.$$

We can determine the distribution function of Z by realizing that the maximum of the X_i is smaller than a number a if and only if *all* X_i are smaller than a:

$$F_Z(a) = \mathrm{P}(Z \leq a) = \mathrm{P}(\max\{X_1, \ldots, X_n\} \leq a) = \mathrm{P}(X_1 \leq a, \ldots, X_n \leq a).$$

Now suppose that the events $\{X_i \leq a_i\}$ are independent for every choice of the a_i. In this case we call the random variables *independent* (see also Chapter 9, where we study independence of random variables). In particular, the events $\{X_i \leq a\}$ are independent for all a. It then follows that

$$F_Z(a) = \mathrm{P}(X_1 \leq a, \ldots, X_n \leq a) = \mathrm{P}(X_1 \leq a) \cdots \mathrm{P}(X_n \leq a).$$

Hence, if all random variables have the same distribution function F, then the following result holds.

> THE DISTRIBUTION OF THE MAXIMUM.   Let $X_1, X_2, \ldots, X_n$ be n independent random variables with the same distribution function F, and let $Z = \max\{X_1, X_2, \ldots, X_n\}$. Then
>
> $$F_Z(a) = (F(a))^n.$$

QUICK EXERCISE 8.5 Let $X_1, X_2, \ldots, X_n$ be independent random variables, all with a $U(0, 1)$ distribution. Let $Z = \max\{X_1, \ldots, X_n\}$. Compute the distribution function and the probability density function of Z.

What can we say about the distribution of the minimum? Let

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

We can now find the distribution function F_V of V by observing that the minimum of the X_i is *larger* than a number a if and only if all X_i are *larger* than a. The trick is to switch to the complement of the event $\{V \leq a\}$:

$$F_V(a) = \mathrm{P}(V \leq a) = 1 - \mathrm{P}(V > a) = 1 - \mathrm{P}(\min\{X_1, \ldots, X_n\} > a)$$
$$= 1 - \mathrm{P}(X_1 > a, \ldots, X_n > a).$$

So using independence and switching back again, we obtain

$$F_V(a) = 1 - \mathrm{P}(X_1 > a, \ldots, X_n > a) = 1 - \mathrm{P}(X_1 > a) \cdots \mathrm{P}(X_n > a)$$
$$= 1 - (1 - \mathrm{P}(X_1 \leq a)) \cdots (1 - \mathrm{P}(X_n \leq a)).$$

We have found the following result for the minimum.

> THE DISTRIBUTION OF THE MINIMUM.   Let $X_1, X_2, \ldots, X_n$ be n independent random variables with the same distribution function F, and let $V = \min\{X_1, X_2, \ldots, X_n\}$. Then
>
> $$F_V(a) = 1 - (1 - F(a))^n.$$

QUICK EXERCISE 8.6 Let $X_1, X_2, \ldots, X_n$ be independent random variables, all with a $U(0, 1)$ distribution. Let $V = \min\{X_1, \ldots, X_n\}$. Compute the distribution function and the probability density function of V.

8.5 Solutions to the quick exercises

8.1 Clearly Z can take the values $1, \ldots, 150$. The value 150 is special: the plane is full if 150 or more people buy a ticket. Hence $P(Z = 150) = P(X \geq 150) = 51/200$. For the other values we have $P(Z = i) = P(X = i) = 1/200$, for $i = 1, \ldots, 149$. Clearly, here $g(x) = \min\{150, x\}$.

8.2 The probability density of $Y = 1/X$ is

$$f_Y(y) = \frac{1}{y^2} \frac{1}{\pi(1 + (\frac{1}{y})^2)} = \frac{1}{\pi(1 + y^2)}.$$

We see that $1/X$ has the same distribution as X! (This distribution is called the standard Cauchy distribution, it will be introduced in Chapter 11.)

8.3 First define $Z = (X - 4)/5$, which has an $N(0, 1)$ distribution. Then from Table B.1

$$P(X \leq 5) = P\left(Z \leq \frac{5 - 4}{5}\right) = P(Z \leq 0.20) = 1 - 0.4207 = 0.5793.$$

Similarly, using the symmetry of the normal distribution,

$$P(X \geq 2) = P\left(Z \geq \frac{2 - 4}{5}\right) = P(Z \geq -0.40) = P(Z \leq 0.40) = 0.6554.$$

8.4 If $g(x) = e^{-x}$, then $g''(x) = e^{-x} > 0$; hence g is strictly convex. It follows from Jensen's inequality that

$$e^{-E[X]} \leq E\left[e^{-X}\right].$$

Moreover, if $\text{Var}(X) > 0$, then the inequality is strict.

8.5 The distribution function of the X_i is given by $F(x) = x$ on $[0, 1]$. Therefore the distribution function F_Z of the maximum Z is equal to $F_Z(a) = (F(a))^n = a^n$. Its probability density function is

$$f_Z(z) = \frac{d}{dz} F_Z(z) = nz^{n-1} \quad \text{for } 0 \leq z \leq 1.$$

8.6 The distribution function of the X_i is given by $F(x) = x$ on $[0, 1]$. Therefore the distribution function F_V of the minimum V is equal to $F_V(a) = 1 - (1 - a)^n$. Its probability density function is

$$f_V(v) = \frac{d}{dv} F_V(v) = n(1 - v)^{n-1} \quad \text{for } 0 \leq v \leq 1.$$

8.6 Exercises

8.1 ⊡ Often one is interested in the distribution of the deviation of a random variable X from its mean $\mu = \mathrm{E}[X]$. Let X take the values $80, 90, 100, 110$, and 120, all with probability 0.2; then $\mathrm{E}[X] = \mu = 100$. Determine the distribution of $Y = |X - \mu|$. That is, specify the values Y can take and give the corresponding probabilities.

8.2 ⊞ Suppose X has a uniform distribution over the points $\{1, 2, 3, 4, 5, 6\}$ and that $g(x) = \sin(\frac{\pi}{2}x)$.

a. Determine the distribution of $Y = g(X) = \sin(\frac{\pi}{2}X)$, that is, specify the values Y can take and give the corresponding probabilities.

b. Let $Z = \cos(\frac{\pi}{2}X)$. Determine the distribution of Z.

c. Determine the distribution of $W = Y^2 + Z^2$. Warning: in this example there is a very special dependency between Y and Z, and in general it is much harder to determine the distribution of a random variable that is a function of *two* other random variables. This is the subject of Chapter 11.

8.3 ⊡ The continuous random variable U is uniformly distributed over $[0, 1]$.

a. Determine the distribution function of $V = 2U + 7$. What kind of distribution does V have?

b. Determine the distribution function of $V = rU + s$ for all real numbers $r > 0$ and s. See Exercise 8.9 for what happens for negative r.

8.4 Transforming exponential distributions.

a. Let X have an $Exp(\frac{1}{2})$ distribution. Determine the distribution function of $\frac{1}{2}X$. What kind of distribution does $\frac{1}{2}X$ have?

b. Let X have an $Exp(\lambda)$ distribution. Determine the distribution function of λX. What kind of distribution does λX have?

8.5 ⊡ Let X be a continuous random variable with probability density function

$$f_X(x) = \begin{cases} \frac{3}{4}x(2 - x) & \text{for } 0 \le x \le 2 \\ 0 & \text{elsewhere.} \end{cases}$$

a. Determine the distribution function F_X.

b. Let $Y = \sqrt{X}$. Determine the distribution function F_Y.

c. Determine the probability density of Y.

8.6 Let X be a continuous random variable with probability density f_X that takes only positive values and let $Y = 1/X$.

a. Determine $F_Y(y)$ and show that

$$f_Y(y) = \frac{1}{y^2} f_X\left(\frac{1}{y}\right) \quad \text{for } y > 0.$$

b. Let $Z = 1/Y$. Using **a**, determine the probability density f_Z of Z, in terms of f_X.

8.7 Let X have a $Par(\alpha)$ distribution. Determine the distribution function of $\ln X$. What kind of a distribution does $\ln X$ have?

8.8 ⊡ Let X have an $Exp(1)$ distribution, and let α and λ be positive numbers. Determine the distribution function of the random variable

$$W = \frac{X^{1/\alpha}}{\lambda}.$$

The distribution of the random variable W is called the Weibull distribution with parameters α and λ.

8.9 Let X be a continuous random variable. Express the distribution function and probability density of the random variable $Y = -X$ in terms of those of X.

8.10 ⊞ Let X be an $N(3,4)$ distributed random variable. Use the rule for normal random variables under change of units and Table B.1 to determine the probabilities $P(X \geq 3)$ and $P(X \leq 1)$.

8.11 ⊞ Let X be a random variable, and let g be a twice differentiable function with $g''(x) \leq 0$ for all x. Such a function is called a *concave* function. Show that for concave functions always

$$g(E[X]) \geq E[g(X)].$$

8.12 ⊞ Let X be a random variable with the following probability mass function:

x	0	1	100	10000
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

a. Determine the distribution of $Y = \sqrt{X}$.

b. Which is larger $E\left[\sqrt{X}\right]$ or $\sqrt{E[X]}$?

   Hint: use Exercise 8.11, or start by showing that the function $g(x) = -\sqrt{x}$ is convex.

c. Compute $\sqrt{E[X]}$ and $E\left[\sqrt{X}\right]$ to check your answer (and to see that it makes a big difference!).

8.13 Let W have a $U(\pi, 2\pi)$ distribution. What is larger: $E[\sin(W)]$ or $\sin(E[W])$? Check your answer by computing these two numbers.

8.14 In this exercise we take a look at Jensen's inequality for the function $g(x) = x^3$ (which is neither convex nor concave on $(-\infty, \infty)$).

a. Can you find a (discrete) random variable X with $\text{Var}(X) > 0$ such that

$$\text{E}\left[X^3\right] = (\text{E}\left[X\right])^3?$$

b. Under what kind of conditions on a random variable X will the inequality $\text{E}\left[X^3\right] > (\text{E}\left[X\right])^3$ certainly hold?

8.15 Let $X_1, X_2, \ldots, X_n$ be independent random variables, all with a $U(0,1)$ distribution. Let $Z = \max\{X_1, \ldots, X_n\}$ and $V = \min\{X_1, \ldots, X_n\}$.

a. Compute $\text{E}\left[\max\{X_1, X_2\}\right]$ and $\text{E}\left[\min\{X_1, X_2\}\right]$.

b. Compute $\text{E}\left[Z\right]$ and $\text{E}\left[V\right]$ for general n.

c. Can you argue directly (using the symmetry of the uniform distribution (see Exercise 6.3) and not the result of the computation in **b**) that $1 - \text{E}\left[\max\{X_1, \ldots, X_n\}\right] = \text{E}\left[\min\{X_1, \ldots, X_n\}\right]$?

8.16 In this exercise we derive a kind of Jensen inequality for the minimum.

a. Let a and b be real numbers. Show that

$$\min\{a, b\} = \frac{1}{2}(a + b - |a - b|).$$

b. Let X and Y be independent random variables with the same distribution and finite expectation. Deduce from **a** that

$$\text{E}\left[\min\{X, Y\}\right] = \text{E}\left[X\right] - \frac{1}{2}\text{E}\left[|X - Y|\right].$$

c. Show that
$$\text{E}\left[\min\{X, Y\}\right] \leq \min\{\text{E}\left[X\right], \text{E}\left[Y\right]\}.$$

Remark: this is not so interesting, since $\min\{\text{E}\left[X\right], \text{E}\left[Y\right]\} = \text{E}\left[X\right] = \text{E}\left[Y\right]$, but we will see in the exercises of Chapter 11 that this inequality is also true for X and Y, which do not have the same distribution.

8.17 Let $X_1, \ldots, X_n$ be n independent random variables with the same distribution function F.

a. Convince yourself that for any numbers $x_1, \ldots, x_n$ it is true that

$$\min\{x_1, \ldots, x_n\} = -\max\{-x_1, \ldots, -x_n\}.$$

b. Let $Z = \max\{X_1, X_2, \ldots, X_n\}$ and $V = \min\{X_1, X_2, \ldots, X_n\}$. Use Exercise 8.9 and the observation in **a** to deduce the formula

$$F_V(a) = 1 - (1 - F(a))^n$$

directly from the formula

$$F_Z(a) = (F(a))^n.$$

8.18 ⊡ Let $X_1, X_2, \ldots, X_n$ be independent random variables, all with an $Exp(\lambda)$ distribution. Let $V = \min\{X_1, \ldots, X_n\}$. Determine the distribution function of V. What kind of distribution is this?

8.19 ⊞ From the "north pole" N of a circle with diameter 1, a point Q on the circle is mapped to a point t on the line by its projection from N, as illustrated in Figure 8.2.
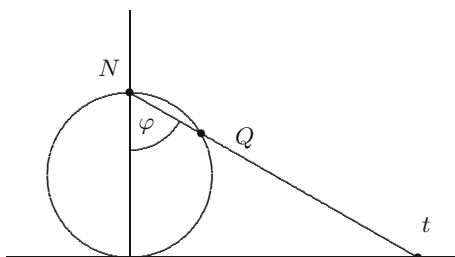
Fig. 8.2. Mapping the circle to the line.

Suppose that the point Q is uniformly chosen on the circle. This is the same as saying that the angle φ is uniformly chosen from the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$ (can you see this?). Let X be this angle, so that X is uniformly distributed over the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$. This means that $P(X \le \varphi) = 1/2 + \varphi/\pi$ (cf. Quick exercise 5.3). What will be the distribution of the projection of Q on the line? Let us call this random variable Z. Then it is clear that the event $\{Z \le t\}$ is equal to the event $\{X \le \varphi\}$, where t and φ correspond to each other under the projection. This means that $\tan(\varphi) = t$, which is the same as saying that $\arctan(t) = \varphi$.

a. What part of the circle is mapped to the interval $[1, \infty)$?

b. Compute the distribution function of Z using the correspondence between t and φ.

c. Compute the probability density function of Z.

The distribution of Z is called the Cauchy distribution (which will be discussed in Chapter 11).

9

Joint distributions and independence

Random variables related to the same experiment often influence one another. In order to capture this, we introduce the *joint distribution* of two or more random variables. We also discuss the notion of *independence* for random variables, which models the situation where random variables do not influence each other. As with single random variables we treat these topics for discrete and continuous random variables separately.

9.1 Joint distributions of discrete random variables

In a census one is usually interested in several variables, such as income, age, and gender. In itself these variables are interesting, but when two (or more) are studied simultaneously, detailed information is obtained on the society where the census is performed. For instance, studying income, age, and gender jointly might give insight to the emancipation of women.

Without mentioning it explicitly, we already encountered several examples of joint distributions of discrete random variables. For example, in Chapter 4 we defined two random variables S and M, the sum and the maximum of two independent throws of a die.

QUICK EXERCISE 9.1 List the elements of the event $\{S = 7, M = 4\}$ and compute its probability.

In general, the joint distribution of two discrete random variables X and Y, defined on the *same* sample space Ω, is given by prescribing the probabilities of all possible values of the pair (X, Y).

DEFINITION. The *joint probability mass function* p of two discrete random variables X and Y is the function $p : \mathbb{R}^2 \to [0, 1]$, defined by

$$p(a, b) = P(X = a, Y = b) \quad \text{for} \ -\infty < a, b < \infty.$$

To stress the dependence on (X, Y), we sometimes write $p_{X,Y}$ instead of p.

If X and Y take on the values $a_1, a_2, \ldots, a_k$ and $b_1, b_2, \ldots, b_\ell$, respectively, the joint distribution of X and Y can simply be described by listing all the possible values of $p(a_i, b_j)$. For example, for the random variables S and M from Chapter 4 we obtain Table 9.1.

Table 9.1. Joint probability mass function $p(a, b) = P(S = a, M = b)$.

				b		
a	1	2	3	4	5	6
2	1/36	0	0	0	0	0
3	0	2/36	0	0	0	0
4	0	1/36	2/36	0	0	0
5	0	0	2/36	2/36	0	0
6	0	0	1/36	2/36	2/36	0
7	0	0	0	2/36	2/36	2/36
8	0	0	0	1/36	2/36	2/36
9	0	0	0	0	2/36	2/36
10	0	0	0	0	1/36	2/36
11	0	0	0	0	0	2/36
12	0	0	0	0	0	1/36

From this table we can retrieve the distribution of S and of M. For example, because

$$\{S = 6\} = \{S = 6, M = 1\} \cup \{S = 6, M = 2\} \cup \cdots \cup \{S = 6, M = 6\},$$

and because the six events

$$\{S = 6, M = 1\}, \{S = 6, M = 2\}, \ldots, \{S = 6, M = 6\}$$

are mutually exclusive, we find that

$$
\begin{aligned}
p_S(6) = P(S = 6) &= P(S = 6, M = 1) + \cdots + P(S = 6, M = 6) \\
&= p(6, 1) + p(6, 2) + \cdots + p(6, 6) \\
&= 0 + 0 + \frac{1}{36} + \frac{2}{36} + \frac{2}{36} + 0 \\
&= \frac{5}{36}.
\end{aligned}
$$

Table 9.2. Joint distribution and marginal distributions of S and M.

			b				
a	1	2	3	4	5	6	$p_S(a)$
2	1/36	0	0	0	0	0	1/36
3	0	2/36	0	0	0	0	2/36
4	0	1/36	2/36	0	0	0	3/36
5	0	0	2/36	2/36	0	0	4/36
6	0	0	1/36	2/36	2/36	0	5/36
7	0	0	0	2/36	2/36	2/36	6/36
8	0	0	0	1/36	2/36	2/36	5/36
9	0	0	0	0	2/36	2/36	4/36
10	0	0	0	0	1/36	2/36	3/36
11	0	0	0	0	0	2/36	2/36
12	0	0	0	0	0	1/36	1/36
$p_M(b)$	1/36	3/36	5/36	7/36	9/36	11/36	1

Thus we see that the probabilities of S can be obtained by taking the sum of the joint probabilities in the rows of Table 9.1. This yields the probability distribution of S, i.e., all values of $p_S(a)$ for $a = 2, \ldots, 12$. We speak of the *marginal* distribution of S. In Table 9.2 we have added this distribution in the right "margin" of the table. Similarly, summing over the columns of Table 9.1 yields the marginal distribution of M, in the bottom margin of Table 9.2.

The joint distribution of two random variables contains a lot more information than the two marginal distributions. This can be illustrated by the fact that in many cases the joint probability mass function of X and Y cannot be retrieved from the marginal probability mass functions p_X and p_Y. A simple example is given in the following quick exercise.

QUICK EXERCISE 9.2 Let X and Y be two discrete random variables, with joint probability mass function p, given by the following table, where ε is an arbitrary number between $-1/4$ and $1/4$.

	b		
a	0	1	$p_X(a)$
0	$1/4 - \varepsilon$	$1/4 + \varepsilon$	...
1	$1/4 + \varepsilon$	$1/4 - \varepsilon$	...
$p_Y(b)$	...	...	...

Complete the table, and conclude that we cannot retrieve p from p_X and p_Y.

The joint distribution function

As in the case of a single random variable, the distribution function enables us to treat pairs of discrete and pairs of continuous random variables in the same way.

> DEFINITION. The *joint distribution function* F of two random variables X and Y is the function $F : \mathbb{R}^2 \to [0,1]$ defined by
>
> $$F(a,b) = \mathrm{P}(X \le a, Y \le b) \quad \text{for } -\infty < a, b < \infty.$$

QUICK EXERCISE 9.3 Compute $F(5,3)$ for the joint distribution function F of the pair (S, M).

The distribution functions F_X and F_Y can be obtained from the joint distribution function of X and Y. As before, we speak of the *marginal* distribution functions. The following rule holds.

> FROM JOINT TO MARGINAL DISTRIBUTION FUNCTION. Let F be the joint distribution function of random variables X and Y. Then the marginal distribution function of X is given for each a by
>
> $$F_X(a) = \mathrm{P}(X \le a) = F(a, +\infty) = \lim_{b \to \infty} F(a,b), \qquad (9.1)$$
>
> and the marginal distribution function of Y is given for each b by
>
> $$F_Y(b) = \mathrm{P}(Y \le b) = F(+\infty, b) = \lim_{a \to \infty} F(a,b). \qquad (9.2)$$

9.2 Joint distributions of continuous random variables

We saw in Chapter 5 that the probability that a single continuous random variable X lies in an interval $[a,b]$, is equal to the area under the probability density function f of X over the interval (see also Figure 5.1). For the joint distribution of continuous random variables X and Y the situation is analogous: the probability that the pair (X,Y) falls in the rectangle $[a_1, b_1] \times [a_2, b_2]$ is equal to the volume under the joint probability density function $f(x,y)$ of (X,Y) over the rectangle. This is illustrated in Figure 9.1, where a chunk of a joint probability density function $f(x,y)$ is displayed for x between -0.5 and 1 and for y between -1.5 and 1. Its volume represents the probability $\mathrm{P}(-0.5 \le X \le 1, -1.5 \le Y \le 1)$. As the volume under f on $[-0.5, 1] \times [-1.5, 1]$ is equal to the integral of f over this rectangle, this motivates the following definition.
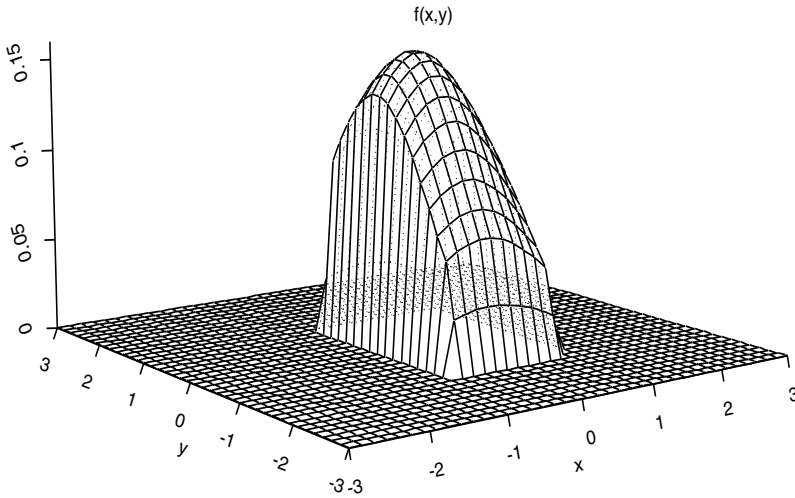
Fig. 9.1. Volume under a joint probability density function f on the rectangle $[-0.5, 1] \times [-1.5, 1]$.

DEFINITION.   Random variables X and Y have a *joint continuous distribution* if for some function $f : \mathbb{R}^2 \to \mathbb{R}$ and for all numbers a_1, a_2 and b_1, b_2 with $a_1 \leq b_1$ and $a_2 \leq b_2$,

$$P(a_1 \leq X \leq b_1, a_2 \leq Y \leq b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y) \, dx \, dy.$$

The function f has to satisfy $f(x, y) \geq 0$ for all x and y, and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1$. We call f the *joint probability density function* of X and Y.

As in the one-dimensional case there is a simple relation between the joint distribution function F and the joint probability density function f:

$$F(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f(x, y) \, dx \, dy \quad \text{and} \quad f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

A joint probability density function of two random variables is also called a *bivariate probability density*. An explicit example of such a density is the function

$$f(x, y) = \frac{30}{\pi} e^{-50x^2 - 50y^2 + 80xy}$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$; see Figure 9.2. This is an example of a bivariate normal density (see Remark 11.2 for a full description of bivariate normal distributions).

We illustrate a number of properties of joint continuous distributions by means of the following simple example. Suppose that X and Y have joint probability
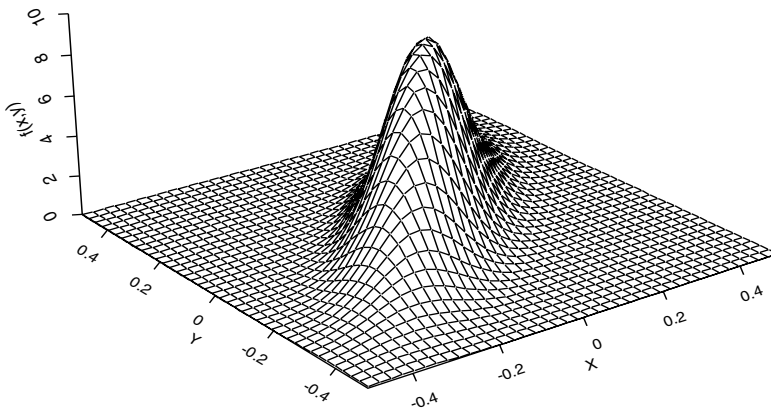
Fig. 9.2. A bivariate normal probability density function.

density function

$$f(x,y) = \frac{2}{75}\left(2x^2y + xy^2\right) \quad \text{for } 0 \le x \le 3 \text{ and } 1 \le y \le 2,$$

and $f(x,y) = 0$ otherwise; see Figure 9.3.
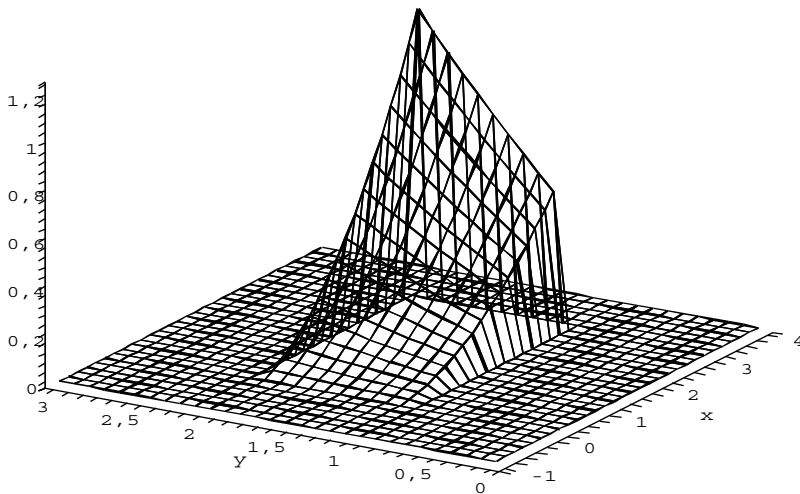
Fig. 9.3. The probability density function $f(x,y) = \frac{2}{75}\left(2x^2y + xy^2\right)$.

As an illustration of how to compute joint probabilities:

$$P\left(1 \leq X \leq 2, \frac{4}{3} \leq Y \leq \frac{5}{3}\right) = \int_1^2 \int_{\frac{4}{3}}^{\frac{5}{3}} f(x, y) \, dx \, dy$$

$$= \frac{2}{75} \int_1^2 \left(\int_{\frac{4}{3}}^{\frac{5}{3}} (2x^2 y + xy^2) \, dy\right) dx$$

$$= \frac{2}{75} \int_1^2 \left(x^2 + \frac{61}{81}x\right) dx = \frac{187}{2025}.$$

Next, for a between 0 and 3 and b between 1 and 2, we determine the expression of the joint distribution function. Since $f(x, y) = 0$ for $x < 0$ or $y < 1$,

$$F(a, b) = P(X \leq a, Y \leq b) = \int_{-\infty}^a \left(\int_{-\infty}^b f(x, y) \, dy\right) dx$$

$$= \frac{2}{75} \int_0^a \left(\int_1^b (2x^2 y + xy^2) \, dy\right) dx$$

$$= \frac{1}{225} \left(2a^3 b^2 - 2a^3 + a^2 b^3 - a^2\right).$$

Note that for either a outside $[0, 3]$ or b outside $[1, 2]$, the expression for $F(a, b)$ is different. For example, suppose that a is between 0 and 3 and b is larger than 2. Since $f(x, y) = 0$ for $y > 2$, we find for any $b \geq 2$:

$$F(a, b) = P(X \leq a, Y \leq b) = P(X \leq a, Y \leq 2) = F(a, 2) = \frac{1}{225} \left(6a^3 + 7a^2\right).$$

Hence, applying (9.1) one finds the marginal distribution function of X:

$$F_X(a) = \lim_{b \to \infty} F(a, b) = \frac{1}{225} \left(6a^3 + 7a^2\right)$$

for a between 0 and 3.

QUICK EXERCISE 9.4 Show that $F_Y(b) = \frac{1}{75}\left(3b^3 + 18b^2 - 21\right)$ for b between 1 and 2.

The probability density of X can be found by differentiating F_X:

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx}\left(\frac{1}{225}\left(6x^3 + 7x^2\right)\right) = \frac{2}{225}\left(9x^2 + 7x\right)$$

for x between 0 and 3. It is also possible to obtain the probability density function of X directly from $f(x, y)$. Recall that we determined marginal probabilities of discrete random variables by summing over the joint probabilities (see Table 9.2). In a similar way we can find f_X. For x between 0 and 3,

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \frac{2}{75} \int_1^2 \left(2x^2 y + xy^2\right) dy = \frac{2}{225}\left(9x^2 + 7x\right).$$

This illustrates the following rule.

> FROM JOINT TO MARGINAL PROBABILITY DENSITY FUNCTION. Let f be the joint probability density function of random variables X and Y. Then the *marginal* probability densities of X and Y can be found as follows:
>
> $$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx.$$

Hence the probability density function of each of the random variables X and Y can easily be obtained by "integrating out" the other variable.

QUICK EXERCISE 9.5 Determine $f_Y(y)$.

9.3 More than two random variables

To determine the joint distribution of n random variables $X_1, X_2, \ldots, X_n$, all defined on the *same* sample space Ω, we have to describe how the probability mass is distributed over all possible values of $(X_1, X_2, \ldots, X_n)$. In fact, it suffices to specify the *joint distribution function* F of $X_1, X_2, \ldots, X_n$, which is defined by

$$F(a_1, a_2, \ldots, a_n) = P(X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n)$$

for $-\infty < a_1, a_2, \ldots, a_n < \infty$.

In case the random variables $X_1, X_2, \ldots, X_n$ are *discrete*, the joint distribution can also be characterized by specifying the *joint probability mass function p* of $X_1, X_2, \ldots, X_n$, defined by

$$p(a_1, a_2, \ldots, a_n) = P(X_1 = a_1, X_2 = a_2, \ldots, X_n = a_n)$$

for $-\infty < a_1, a_2, \ldots, a_n < \infty$.

Drawing without replacement

Let us illustrate the use of the joint probability mass function with an example. In the weekly Dutch National Lottery Show, 6 balls are drawn from a vase that contains balls numbered from 1 to 41. Clearly, the first number takes values $1, 2, \ldots, 41$ with equal probabilities. Is this also the case for—say—the third ball?

Let us consider a more general situation. Suppose a vase contains balls numbered $1, 2, \ldots, N$. We draw n balls *without replacement* from the vase. Note that n cannot be larger than N. Each ball is selected with equal probability, i.e., in the first draw each ball has probability $1/N$, in the second draw each of the $N-1$ remaining balls has probability $1/(N-1)$, and so on. Let X_i denote the number on the ball in the i-th draw, for $i = 1, 2, \ldots, n$. In order to obtain the marginal probability mass function of X_i, we first compute the joint probability mass function of $X_1, X_2, \ldots, X_n$. Since there are $N(N-1)\cdots(N-n+1)$ possible combinations for the values of $X_1, X_2, \ldots, X_n$, each having the same probability, the joint probability mass function is given by

$$p(a_1, a_2, \ldots, a_n) = \mathrm{P}(X_1 = a_1, X_2 = a_2, \ldots, X_n = a_n)$$
$$= \frac{1}{N(N-1)\cdots(N-n+1)},$$

for all *distinct* values $a_1, a_2, \ldots, a_n$ with $1 \le a_j \le N$. Clearly $X_1, X_2, \ldots, X_n$ influence each other. Nevertheless, the marginal distribution of each X_i is the same. This can be seen as follows. Similar to obtaining the marginal probability mass functions in Table 9.2, we can find the marginal probability mass function of X_i by summing the joint probability mass function over all possible values of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$:

$$p_{X_i}(k) = \sum p(a_1, \ldots, a_{i-1}, k, a_{i+1}, \ldots, a_n)$$
$$= \sum \frac{1}{N(N-1)\cdots(N-n+1)},$$

where the sum runs over all distinct values $a_1, a_2, \ldots, a_n$ with $1 \le a_j \le N$ and $a_i = k$. Since there are $(N-1)(N-2)\cdots(N-n+1)$ such combinations, we conclude that the marginal probability mass function of X_i is given by

$$p_{X_i}(k) = (N-1)(N-2)\cdots(N-n+1) \cdot \frac{1}{N(N-1)\cdots(N-n+1)} = \frac{1}{N},$$

for $k = 1, 2, \ldots, N$. We see that the marginal probability mass function of each X_i is the *same*, assigning equal probability $1/N$ to each possible value.

In case the random variables $X_1, X_2, \ldots, X_n$ are *continuous*, the joint distribution is defined in a similar way as in the case of two variables. We say that the random variables $X_1, X_2, \ldots, X_n$ have a *joint continuous distribution* if for some function $f : \mathbb{R}^n \to \mathbb{R}$ and for all numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ with $a_i \le b_i$,

$$\mathrm{P}(a_1 \le X_1 \le b_1, a_2 \le X_2 \le b_2, \ldots, a_n \le X_n \le b_n)$$
$$= \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n} f(x_1, x_2, \ldots, x_n) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \cdots \mathrm{d}x_n.$$

Again f has to satisfy $f(x_1, x_2, \ldots, x_n) \ge 0$ and f has to integrate to 1. We call f the *joint probability density* of $X_1, X_2, \ldots, X_n$.

9.4 Independent random variables

In earlier chapters we have spoken of independence of random variables, anticipating a formal definition. On page 46 we postulated that the events

$$\{R_1 = a_1\}, \{R_2 = a_2\}, \dots, \{R_{10} = a_{10}\}$$

related to the Bernoulli random variables $R_1, \dots, R_{10}$ are independent. How should one define independence of random variables? Intuitively, random variables X and Y are independent if every event involving only X is independent of every event involving only Y. Since for two discrete random variables X and Y, any event involving X and Y is the union of events of the type $\{X = a, Y = b\}$, an adequate definition for independence would be

$$P(X = a, Y = b) = P(X = a) P(Y = b), \tag{9.3}$$

for all possible values a and b. However, this definition is useless for continuous random variables. Both the discrete and the continuous case are covered by the following definition.

> DEFINITION. The random variables X and Y, with joint distribution function F, are *independent* if
>
> $$P(X \leq a, Y \leq b) = P(X \leq a) P(Y \leq b),$$
>
> that is,
> $$F(a, b) = F_X(a) F_Y(b) \tag{9.4}$$
>
> for all possible values a and b. Random variables that are not independent are called *dependent*.

Note that independence of X and Y guarantees that the joint probability of $\{X \leq a, Y \leq b\}$ factorizes. More generally, the following is true: if X and Y are independent, then

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B), \tag{9.5}$$

for all suitable A and B, such as intervals and points. As a special case we can take $A = \{a\}$, $B = \{b\}$, which yields that for independent X and Y the probability of $\{X = a, Y = b\}$ equals the product of the marginal probabilities. In fact, for *discrete* random variables the definition of independence can be reduced—after cumbersome computations—to equality (9.3). For continuous random variables X and Y we find, differentiating both sides of (9.4) with respect to x and y, that

$$f(x, y) = f_X(x) f_Y(y).$$

QUICK EXERCISE 9.6 Determine for which value of ε the discrete random variables X and Y from Quick exercise 9.2 are independent.

More generally, random variables $X_1, X_2, \ldots, X_n$, with joint distribution function F, are *independent* if for all values $a_1, \ldots, a_n$,

$$F(a_1, a_2, \ldots, a_n) = F_{X_1}(a_1)F_{X_2}(a_2)\cdots F_{X_n}(a_n).$$

As in the case of two discrete random variables, the discrete random variables $X_1, X_2, \ldots, X_n$ are independent if

$$\mathrm{P}(X_1 = a_1, \ldots, X_n = a_n) = \mathrm{P}(X_1 = a_1)\cdots\mathrm{P}(X_n = a_n),$$

for all possible values $a_1, \ldots, a_n$. Thus we see that the definition of independence for discrete random variables is in agreement with our intuitive interpretation given earlier in (9.3).

In case of independent continuous random variables $X_1, X_2, \ldots, X_n$ with joint probability density function f, differentiating the joint distribution function with respect to all the variables gives that

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n) \tag{9.6}$$

for all values $x_1, \ldots, x_n$. By integrating both sides over $(-\infty, a_1] \times (-\infty, a_2] \times \cdots \times (-\infty, a_n]$, we find the definition of independence. Hence in the continuous case, (9.6) is equivalent to the definition of independence.

9.5 Propagation of independence

A natural question is whether transformed independent random variables are again independent. We start with a simple example. Let X and Y be two independent random variables with joint distribution function F. Take an interval $I = (a, b]$ and define random variables U and V as follows:

$$U = \begin{cases} 1 & \text{if } X \in I \\ 0 & \text{if } X \notin I, \end{cases} \quad \text{and} \quad V = \begin{cases} 1 & \text{if } Y \in I \\ 0 & \text{if } Y \notin I. \end{cases}$$

Are U and V independent? Yes, they are! By using (9.5) and the independence of X and Y, we can write

$$\begin{aligned} \mathrm{P}(U = 0, V = 1) &= \mathrm{P}(X \in I^c, Y \in I) \\ &= \mathrm{P}(X \in I^c)\,\mathrm{P}(Y \in I) \\ &= \mathrm{P}(U = 0)\,\mathrm{P}(V = 1). \end{aligned}$$

By a similar reasoning one finds that for *all* values a and b,

$$P(U = a, V = b) = P(U = a) P(V = b).$$

This illustrates the fact that for independent random variables $X_1, X_2, \ldots, X_n$, the random variables $Y_1, Y_2, \ldots, Y_n$, where each Y_i is determined by X_i only, inherit the independence from the X_i. The general rule is given here.

> PROPAGATION OF INDEPENDENCE. Let $X_1, X_2, \ldots, X_n$ be independent random variables. For each i, let $h_i : \mathbb{R} \to \mathbb{R}$ be a function and define the random variable
>
> $$Y_i = h_i(X_i).$$
>
> Then $Y_1, Y_2, \ldots, Y_n$ are also independent.

Often one uses this rule with all functions the same: $h_i = h$. For instance, in the preceding example,

$$h(x) = \begin{cases} 1 & \text{if } x \in I \\ 0 & \text{if } x \notin I. \end{cases}$$

The rule is also useful when we need different transformations for different X_i. We already saw an example of this in Chapter 6. In the single-server queue example in Section 6.4, the $Exp(0.5)$ random variables $T_1, T_2, \ldots$ and $U(2, 5)$ random variables $S_1, S_2, \ldots$ are required to be independent. They are generated according to the technique described in Section 6.2. With a sequence $U_1, U_2, \ldots$ of independent $U(0, 1)$ random variables we can accomplish independence of the T_i and S_i as follows:

$$T_i = F^{\text{inv}}(U_{2i-1}) \quad \text{and} \quad S_i = G^{\text{inv}}(U_{2i}),$$

where F and G are the distribution functions of the $Exp(0.5)$ distribution and the $U(2, 5)$ distribution. The propagation-of-independence rule now guarantees that all random variables $T_1, S_1, T_2, S_2, \ldots$ are independent.

9.6 Solutions to the quick exercises

9.1 The only possibilities with the sum equal to 7 and the maximum equal to 4 are the combinations $(3, 4)$ and $(4, 3)$. They both have probability $1/36$, so that $P(S = 7, M = 4) = 2/36$.

9.2 Since $p_X(0)$, $p_X(1)$, $p_Y(0)$, and $p_Y(1)$ are all equal to $1/2$, knowing only p_X and p_Y yields no information on ε whatsoever. You have to be a student at Hogwarts to be able to get the values of p right!

9.3 Since S and M are discrete random variables, $F(5, 3)$ is the sum of the probabilities $P(S = a, M = b)$ of all combinations (a, b) with $a \le 5$ and $b \le 3$. From Table 9.2 we see that this sum is $8/36$.

9.4 For a between 0 and 3 and for b between 1 and 2, we have seen that

$$F(a,b) = \frac{1}{225}\left(2a^3b^2 - 2a^3 + a^2b^3 - a^2\right).$$

Since $f(x,y) = 0$ for $x > 3$, we find for any $a \geq 3$ and b between 1 and 2:

$$F(a,b) = \mathrm{P}(X \leq a, Y \leq b) = \mathrm{P}(X \leq 3, Y \leq b)$$
$$= F(3,b) = \frac{1}{75}\left(3b^3 + 18b^2 - 21\right).$$

As a result, applying (9.2) yields that $F_Y(b) = \lim_{a\to\infty} F(a,b) = F(3,b) = \frac{1}{75}\left(3b^3 + 18b^2 - 21\right)$, for b between 1 and 2.

9.5 For y between 1 and 2, we have seen that $F_Y(y) = \frac{1}{75}\left(3y^3 + 18y^2 - 21\right)$. Differentiating with respect to y yields that

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_Y(y) = \frac{1}{25}(3y^2 + 12y),$$

for y between 1 and 2 (and $f_Y(y) = 0$ otherwise). The probability density function of Y can also be obtained directly from $f(x,y)$. For y between 1 and 2:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x = \frac{2}{75}\int_0^3 (2x^2y + xy^2)\,\mathrm{d}x$$
$$= \frac{2}{75}\left[\frac{2}{3}x^3y + \frac{1}{2}x^2y^2\right]_{x=0}^{x=3} = \frac{1}{25}(3y^2 + 12y).$$

Since $f(x,y) = 0$ for values of y not between 1 and 2, we have that $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x = 0$ for these y's.

9.6 The number ε is between $-1/4$ and $1/4$. Now X and Y are independent in case $p(i,j) = \mathrm{P}(X = i, Y = j) = \mathrm{P}(X = i)\,\mathrm{P}(Y = j) = p_X(i)p_Y(j)$, for all $i,j = 0,1$. If $i = j = 0$, we should have

$$\frac{1}{4} - \varepsilon = p(0,0) = p_X(0)\,p_Y(0) = \frac{1}{4}.$$

This implies that $\varepsilon = 0$. Furthermore, for all other combinations (i,j) one can check that for $\varepsilon = 0$ also $p(i,j) = p_X(i)\,p_Y(j)$, so that X and Y are independent. If $\varepsilon \neq 0$, we have $p(0,0) \neq p_X(0)\,p_Y(0)$, so that X and Y are dependent.

9.7 Exercises

9.1 The joint probabilities $\mathrm{P}(X = a, Y = b)$ of discrete random variables X and Y are given in the following table (which is based on the *magical square* in Albrecht Dürer's engraving *Melencolia I* in Figure 9.4). Determine the marginal probability distributions of X and Y, i.e., determine the probabilities $\mathrm{P}(X = a)$ and $\mathrm{P}(Y = b)$ for $a,b = 1,2,3,4$.

Fig. 9.4. Albrecht Dürer's *Melencolia I*.

Albrecht Dürer (German, 1471-1528) *Melencolia I*, 1514. Engraving. Bequest of William P. Chapman, Jr., Class of 1895. Courtesy of the Herbert F. Johnson Museum of Art, Cornell University.

	a			
b	1	2	3	4
1	16/136	3/136	2/136	13/136
2	5/136	10/136	11/136	8/136
3	9/136	6/136	7/136	12/136
4	4/136	15/136	14/136	1/136

9.2 ⊞ The joint probability distribution of two discrete random variables X and Y is partly given in the following table.

		a		
b	0	1	2	$P(Y = b)$
-1	...	...	...	$1/2$
1	...	$1/2$	...	$1/2$
$P(X = a)$	$1/6$	$2/3$	$1/6$	1

 a. Complete the table.

 b. Are X and Y dependent or independent?

9.3 Let X and Y be two random variables, with joint distribution the *Melencolia* distribution, given by the table in Exercise 9.1. What is

 a. $P(X = Y)$?

 b. $P(X + Y = 5)$?

 c. $P(1 < X \leq 3, 1 < Y \leq 3)$?

 d. $P((X, Y) \in \{1, 4\} \times \{1, 4\})$?

9.4 This exercise will be easy for those familiar with Japanese puzzles called *nonograms*. The marginal probability distributions of the discrete random variables X and Y are given in the following table:

			a			
b	1	2	3	4	5	$P(Y = b)$
1						$5/14$
2						$4/14$
3						$2/14$
4						$2/14$
5						$1/14$
$P(X = a)$	$1/14$	$5/14$	$4/14$	$2/14$	$2/14$	1

Moreover, for a and b from 1 to 5 the joint probability $P(X = a, Y = b)$ is either 0 or $1/14$. Determine the joint probability distribution of X and Y.

9.5 ⊡ Let η be an unknown real number, and let the joint probabilities $P(X = a, Y = b)$ of the discrete random variables X and Y be given by the following table:

	a		
b	-1	0	1
4	$\eta - \frac{1}{16}$	$\frac{1}{4} - \eta$	0
5	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{8}$
6	$\eta + \frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4} - \eta$

a. Which are the values η can attain?

b. Is there a value of η for which X and Y are independent?

9.6 ⊡ Let X and Y be two independent $Ber(\frac{1}{2})$ random variables. Define random variables U and V by:

$$U = X + Y \quad \text{and} \quad V = |X - Y|.$$

a. Determine the joint and marginal probability distributions of U and V.

b. Find out whether U and V are dependent or independent.

9.7 To investigate the relation between hair color and eye color, the hair color and eye color of 5383 persons was recorded. The data are given in the following table:

Eye color	Hair color		
	Fair/red	Medium	Dark/black
Light	1168	825	305
Dark	573	1312	1200

Source: B. Everitt and G. Dunn. *Applied multivariate data analysis.* Second edition Hodder Arnold, 2001; Table 4.12. Reproduced by permission of Hodder & Stoughton.

Eye color is encoded by the values 1 (Light) and 2 (Dark), and hair color by 1 (Fair/red), 2 (Medium), and 3 (Dark/black). By dividing the numbers in the table by 5383, the table is turned into a joint probability distribution for random variables X (hair color) taking values 1 to 3 and Y (eye color) taking values 1 and 2.

a. Determine the joint and marginal probability distributions of X and Y.

b. Find out whether X and Y are dependent or independent.

9.8 ⊞ Let X and Y be independent random variables with probability distributions given by

$$P(X = 0) = P(X = 1) = \tfrac{1}{2} \quad \text{and} \quad P(Y = 0) = P(Y = 2) = \tfrac{1}{2}.$$

a. Compute the distribution of $Z = X + Y$.

b. Let $\tilde{Y}$ and $\tilde{Z}$ be independent random variables, where $\tilde{Y}$ has the same distribution as Y, and $\tilde{Z}$ the same distribution as Z. Compute the distribution of $\tilde{X} = \tilde{Z} - \tilde{Y}$.

9.9 ⊞ Suppose that the joint distribution function of X and Y is given by

$$F(x, y) = 1 - e^{-2x} - e^{-y} + e^{-(2x+y)} \quad \text{if } x > 0,\, y > 0,$$

and $F(x, y) = 0$ otherwise.

a. Determine the marginal distribution functions of X and Y.

b. Determine the joint probability density function of X and Y.

c. Determine the marginal probability density functions of X and Y.

d. Find out whether X and Y are independent.

9.10 ⊡ Let X and Y be two continuous random variables with joint probability density function

$$f(x, y) = \frac{12}{5}xy(1 + y) \quad \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1,$$

and $f(x, y) = 0$ otherwise.

a. Find the probability $\mathrm{P}\big(\frac{1}{4} \le X \le \frac{1}{2}, \frac{1}{3} \le Y \le \frac{2}{3}\big)$.

b. Determine the joint distribution function of X and Y for a and b between 0 and 1.

c. Use your answer from **b** to find $F_X(a)$ for a between 0 and 1.

d. Apply the rule on page 122 to find the probability density function of X from the joint probability density function $f(x, y)$. Use the result to verify your answer from **c**.

e. Find out whether X and Y are independent.

9.11 ⊞ Let X and Y be two continuous random variables, with the same joint probability density function as in Exercise 9.10. Find the probability $\mathrm{P}(X < Y)$ that X is smaller than Y.

9.12 The joint probability density function f of the pair (X, Y) is given by

$$f(x, y) = K(3x^2 + 8xy) \quad \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 2,$$

and $f(x, y) = 0$ for all other values of x and y. Here K is some positive constant.

a. Find K.

b. Determine the probability $\mathrm{P}(2X \le Y)$.

9.13 ☐ On a disc with origin $(0,0)$ and radius 1, a point (X,Y) is selected by throwing a dart that hits the disc in an arbitrary place. This is best described by the joint probability density function f of X and Y, given by

$$f(x,y) = \begin{cases} c & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

where c is some positive constant.

a. Determine c.

b. Let $R = \sqrt{X^2 + Y^2}$ be the distance from (X,Y) to the origin. Determine the distribution function F_R.

c. Determine the marginal density function f_X. Without doing any calculations, what can you say about f_Y?

9.14 An arbitrary point (X,Y) is drawn from the square $[-1,1] \times [-1,1]$. This means that for any region G in the plane, the probability that (X,Y) is in G, is given by the area of $G \cap \square$ divided by the area of $\square$, where $\square$ denotes the square $[-1,1] \times [-1,1]$:

$$P((X,Y) \in G) = \frac{\text{area of } G \cap \square}{\text{area of } \square}.$$

a. Determine the joint probability density function of the pair (X,Y).

b. Check that X and Y are two independent, $U(-1,1)$ distributed random variables.

9.15 ⊞ Let the pair (X,Y) be drawn arbitrarily from the triangle Δ with vertices $(0,0)$, $(0,1)$, and $(1,1)$.

a. Use Figure 9.5 to show that the joint distribution function F of the pair (X,Y) satisfies

$$F(a,b) = \begin{cases} 0 & \text{for } a \text{ or } b \text{ less than } 0 \\ a(2b-a) & \text{for } (a,b) \text{ in the triangle } \Delta \\ b^2 & \text{for } b \text{ between 0 and 1 and } a \text{ larger than } b \\ 2a - a^2 & \text{for } a \text{ between 0 and 1 and } b \text{ larger than 1} \\ 1 & \text{for } a \text{ and } b \text{ larger than 1.} \end{cases}$$

b. Determine the joint probability density function f of the pair (X,Y).

c. Show that $f_X(x) = 2 - 2x$ for x between 0 and 1 and that $f_Y(y) = 2y$ for y between 0 and 1.

9.16 (Continuation of Exercise 9.15) An arbitrary point (U,V) is drawn from the unit square $[0,1] \times [0,1]$. Let X and Y be defined as in Exercise 9.15. Show that $\min\{U,V\}$ has the same distribution as X and that $\max\{U,V\}$ has the same distribution as Y.
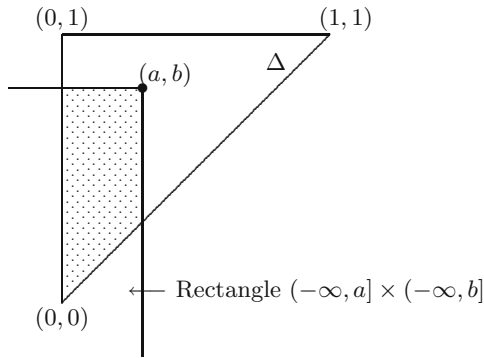
Fig. 9.5. Drawing (X, Y) from $(-\infty, a] \times (-\infty, b] \cap \Delta$.

9.17 Let U_1 and U_2 be two independent random variables, both uniformly distributed over $[0, a]$. Let $V = \min\{U_1, U_2\}$ and $Z = \max\{U_1, U_2\}$. Show that the joint distribution function of V and Z is given by

$$F(s, t) = \mathrm{P}(V \leq s, Z \leq t) = \frac{t^2 - (t - s)^2}{a^2} \quad \text{for } 0 \leq s \leq t \leq a.$$

Hint: note that $V \leq s$ and $Z \leq t$ happens exactly when both $U_1 \leq t$ and $U_2 \leq t$, but *not* both $s < U_1 \leq t$ and $s < U_2 \leq t$.

9.18 Suppose a vase contains balls numbered $1, 2, \ldots, N$. We draw n balls *without replacement* from the vase. Each ball is selected with equal probability, i.e., in the first draw each ball has probability $1/N$, in the second draw each of the $N - 1$ remaining balls has probability $1/(N - 1)$, and so on. For $i = 1, 2, \ldots, n$, let X_i denote the number on the ball in the ith draw. We have shown that the marginal probability mass function of X_i is given by

$$p_{X_i}(k) = \frac{1}{N}, \quad \text{for} \quad k = 1, 2, \ldots, N.$$

a. Show that
$$\mathrm{E}[X_i] = \frac{N + 1}{2}.$$

b. Compute the variance of X_i. You may use the identity

$$1 + 4 + 9 + \cdots + N^2 = \frac{1}{6} N(N + 1)(2N + 1).$$

9.19 ☐ Let X and Y be two continuous random variables, with joint probability density function

$$f(x, y) = \frac{30}{\pi} e^{-50x^2 - 50y^2 + 80xy}$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$; see also Figure 9.2.

a. Determine positive numbers a, b, and c such that

$$50x^2 - 80xy + 50y^2 = (ay - bx)^2 + cx^2.$$

b. Setting $\mu = \frac{4}{5}x$, and $\sigma = \frac{1}{10}$, show that

$$(\sqrt{50}y - \sqrt{32}x)^2 = \frac{1}{2}\left(\frac{y - \mu}{\sigma}\right)^2$$

and use this to show that

$$\int_{-\infty}^{\infty} e^{-(\sqrt{50}y - \sqrt{32}x)^2}\, dy = \frac{\sqrt{2\pi}}{10}.$$

c. Use the results from **b** to determine the probability density function f_X of X. What kind of distribution does X have?

9.20 Suppose we throw a needle on a large sheet of paper, on which horizontal lines are drawn, which are at needle-length apart (see also Exercise 21.16). Choose one of the horizontal lines as x-axis, and let (X, Y) be the center of the needle. Furthermore, let Z be the distance of this center (X, Y) to the nearest horizontal line under (X, Y), and let H be the angle between the needle and the positive x-axis.

a. Assuming that the length of the needle is equal to 1, argue that Z has a $U(0, 1)$ distribution. Also argue that H has a $U(0, \pi)$ distribution and that Z and H are independent.

b. Show that the needle hits a horizontal line when

$$Z \le \frac{1}{2}\sin H \quad \text{or} \quad 1 - Z \le \frac{1}{2}\sin H.$$

c. Show that the probability that the needle will hit one of the horizontal lines equals $2/\pi$.

10

Covariance and correlation

In this chapter we see how the joint distribution of two or more random variables is used to compute the expectation of a combination of these random variables. We discuss the expectation and variance of a sum of random variables and introduce the notions of *covariance* and *correlation*, which express to some extent the way two random variables influence each other.

10.1 Expectation and joint distributions

China vases of various shapes are produced in the Delftware factories in the old city of Delft. One particular simple cylindrical model has height H and radius R centimeters. Due to all kinds of circumstances—the place of the vase in the oven, the fact that the vases are handmade, etc.—H and R are not constants but are random variables. The volume of a vase is equal to the random variable $V = \pi H R^2$, and one is interested in its expected value $\mathrm{E}[V]$. When f_V denotes the probability density of V, then by definition

$$\mathrm{E}[V] = \int_{-\infty}^{\infty} v f_V(v)\,\mathrm{d}v.$$

However, to obtain $\mathrm{E}[V]$, we do not necessarily need to determine f_V from the joint probability density f of H and R! Since V is a function of H and R, we can use a rule similar to the change-of-variable formula from Chapter 7:

$$\mathrm{E}[V] = \mathrm{E}\left[\pi H R^2\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \pi h r^2 f(h, r)\,\mathrm{d}h\,\mathrm{d}r.$$

Suppose that H has a $U(25, 35)$ distribution and that R has a $U(7.5, 12.5)$ distribution. In the case that H and R are also independent, we have

$$E[V] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \pi h r^2 f_H(h) f_R(r) \, dh \, dr = \int_{25}^{35}\int_{7.5}^{12.5} \pi h r^2 \cdot \frac{1}{10} \cdot \frac{1}{5} \, dh \, dr$$

$$= \frac{\pi}{50}\int_{25}^{35} h \, dh \int_{7.5}^{12.5} r^2 \, dr = 9621.127 \text{ cm}^3.$$

This illustrates the following general rule.

TWO-DIMENSIONAL CHANGE-OF-VARIABLE FORMULA.   Let X and Y be random variables, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function. If X and Y are *discrete* random variables with values $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$, respectively, then

$$E[g(X,Y)] = \sum_i \sum_j g(a_i, b_j) P(X = a_i, Y = b_j).$$

If X and Y are *continuous* random variables with joint probability density function f, then

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f(x,y) \, dx \, dy.$$

As an example, take $g(x,y) = xy$ for discrete random variables X and Y with the joint probability distribution given in Table 10.1. The expectation of XY is computed as follows:

$$E[XY] = (0 \cdot 0) \cdot 0 + (1 \cdot 0) \cdot \frac{1}{4} + (2 \cdot 0) \cdot 0$$

$$+ (0 \cdot 1) \cdot \frac{1}{4} + (1 \cdot 1) \cdot 0 + (2 \cdot 1) \cdot \frac{1}{4}$$

$$+ (0 \cdot 2) \cdot 0 + (1 \cdot 2) \cdot \frac{1}{4} + (2 \cdot 2) \cdot 0 = 1.$$

A natural question is whether this value can also be obtained from $E[X] E[Y]$. We return to this question later in this chapter. First we address the expectation of the sum of two random variables.

Table 10.1. Joint probabilities $P(X = a, Y = b)$.

b	a 0	1	2
0	0	1/4	0
1	1/4	0	1/4
2	0	1/4	0

QUICK EXERCISE 10.1 Compute $E[X + Y]$ for the random variables with the joint distribution given in Table 10.1.

For discrete X and Y with values $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$, respectively, we see that

$$
\begin{aligned}
E[X + Y] &= \sum_i \sum_j (a_i + b_j)P(X = a_i, \, Y = b_j) \\
&= \sum_i \sum_j a_i P(X = a_i, \, Y = b_j) + \sum_i \sum_j b_j P(X = a_i, \, Y = b_j) \\
&= \sum_i a_i \left(\sum_j P(X = a_i, \, Y = b_j) \right) \\
&\qquad + \sum_j b_j \left(\sum_i P(X = a_i, \, Y = b_j) \right) \\
&= \sum_i a_i P(X = a_i) + \sum_j b_j P(Y = b_j) \\
&= E[X] + E[Y] .
\end{aligned}
$$

A similar line of reasoning applies in case X and Y are continuous random variables. The following general rule holds.

> LINEARITY OF EXPECTATIONS.   For all numbers r, s, and t and random variables X and Y, one has
>
> $$E[rX + sY + t] = rE[X] + sE[Y] + t.$$

QUICK EXERCISE 10.2 Determine the marginal distributions for the random variables X and Y with the joint distribution given in Table 10.1, and use them to compute $E[X]$ en $E[Y]$. Check that $E[X] + E[Y]$ is equal to $E[X + Y]$, which was computed in Quick exercise 10.1.

More generally, for random variables $X_1, \ldots, X_n$ and numbers $s_1, \ldots, s_n$ and t,

$$E[s_1 X_1 + \cdots + s_n X_n + t] = s_1 E[X_1] + \cdots + s_n E[X_n] + t.$$

This rule is a powerful instrument. For example, it provides an easy way to compute the expectation of a random variable X with a $Bin(n, p)$ distribution. If we would use the definition of expectation, we have to compute

$$E[X] = \sum_{k=0}^{n} kP(X = k) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1 - p)^{n-k}.$$

To determine this sum is not straightforward. However, there is a simple alternative. Recall the multiple-choice example from Section 4.3. We represented

the number of correct answers out of 10 multiple-choice questions as a sum of 10 Bernoulli random variables. More generally, any random variable X with a $Bin(n,p)$ distribution can be represented as

$$X = R_1 + R_2 + \cdots + R_n,$$

where $R_1, R_2, \ldots, R_n$ are independent $Ber(p)$ random variables, i.e.,

$$R_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases}$$

Since $\mathrm{E}[R_i] = 0 \cdot (1-p) + 1 \cdot p = p$, for every $i = 1, 2, \ldots, n$, the linearity-of-expectations rule yields

$$\mathrm{E}[X] = \mathrm{E}[R_1] + \mathrm{E}[R_2] + \cdots + \mathrm{E}[R_n] = np.$$

Hence we conclude that the expectation of a $Bin(n,p)$ distribution equals np.

Remark 10.1 (More than two random variables). In both the discrete and continuous cases, the change-of-variable formula for n random variables is a straightforward generalization of the change-of-variable formula for two random variables. For instance, if $X_1, X_2, \ldots, X_n$ are continuous random variables, with joint probability density function f, and g is a function from $\mathbb{R}^n$ to $\mathbb{R}$, then

$$\mathrm{E}[g(X_1, \ldots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) \, dx_1 \cdots dx_n.$$

10.2 Covariance

In the previous section we have seen that for two random variables X and Y always

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

Does such a simple relation also hold for the variance of the sum $\mathrm{Var}(X + Y)$ or for expectation of the product $\mathrm{E}[XY]$? We will investigate this in the current section.

For the variables X and Y from the example in Section 9.2 with joint probability density

$$f(x, y) = \frac{2}{75} \left(2x^2 y + xy^2 \right) \quad \text{for } 0 \le x \le 3 \text{ and } 1 \le y \le 2,$$

one can show that

$$\mathrm{Var}(X + Y) = \frac{939}{2000} \quad \text{and} \quad \mathrm{Var}(X) + \mathrm{Var}(Y) = \frac{989}{2500} + \frac{791}{10\,000} = \frac{4747}{10\,000}$$

(see Exercise 10.10). This shows, in contrast to the linearity-of-expectations rule, that $\mathrm{Var}(X+Y)$ is generally *not equal* to $\mathrm{Var}(X) + \mathrm{Var}(Y)$. To determine $\mathrm{Var}(X+Y)$, we exploit its definition:

$$\mathrm{Var}(X+Y) = \mathrm{E}\left[(X+Y-\mathrm{E}[X+Y])^2\right].$$

Now $X+Y - \mathrm{E}[X+Y] = (X - \mathrm{E}[X]) + (Y - \mathrm{E}[Y])$, so that

$$(X+Y - \mathrm{E}[X+Y])^2 = (X - \mathrm{E}[X])^2 + (Y - \mathrm{E}[Y])^2$$
$$+ 2\,(X - \mathrm{E}[X])\,(Y - \mathrm{E}[Y]).$$

Taking expectations on both sides, another application of the linearity-of-expectations rule gives

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right].$$

That is, the variance of the sum $X + Y$ equals the sum of the variances of X and Y, plus an extra term $2\mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$. To some extent this term expresses the way X and Y influence each other.

> DEFINITION. Let X and Y be two random variables. The *covariance* between X and Y is defined by
>
> $$\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right].$$

Loosely speaking, if the covariance of X and Y is positive, then if X has a realization larger than $\mathrm{E}[X]$, it is likely that Y will have a realization larger than $\mathrm{E}[Y]$, and the other way around. In this case we say that X and Y are *positively correlated*. In case the covariance is negative, the opposite effect occurs; X and Y are *negatively correlated*. In case $\mathrm{Cov}(X,Y) = 0$ we say that X and Y are *uncorrelated*. An easy consequence of the linearity-of-expectations property (see Exercise 10.19) is the following rule.

> AN ALTERNATIVE EXPRESSION FOR THE COVARIANCE. Let X and Y be two random variables, then
>
> $$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y].$$

For X and Y from the example in Section 9.2, we have $\mathrm{E}[X] = 109/50$, $\mathrm{E}[Y] = 157/100$, and $\mathrm{E}[XY] = 171/50$ (see Exercise 10.10). Thus we see that X and Y are negatively correlated:

$$\mathrm{Cov}(X,Y) = \frac{171}{50} - \frac{109}{50} \cdot \frac{157}{100} = -\frac{13}{5000} < 0.$$

Moreover, this also illustrates that, in contrast to the expectation of the sum, for the expectation of the product, in general $\mathrm{E}[XY]$ is *not equal* to $\mathrm{E}[X]\,\mathrm{E}[Y]$.

Independent versus uncorrelated

Now let X and Y be two *independent* random variables. One expects that X and Y are uncorrelated: they have nothing to do with one another! This is indeed the case, for instance, if X and Y are discrete; one finds that

$$
\begin{aligned}
\mathrm{E}[XY] &= \sum_i \sum_j a_i b_j \mathrm{P}(X = a_i, \, Y = b_j) \\
&= \sum_i \sum_j a_i b_j \mathrm{P}(X = a_i) \, \mathrm{P}(Y = b_j) \\
&= \left(\sum_i a_i \mathrm{P}(X = a_i) \right) \left(\sum_j b_j \mathrm{P}(Y = b_j) \right) \\
&= \mathrm{E}[X] \, \mathrm{E}[Y].
\end{aligned}
$$

A similar reasoning holds in case X and Y are continuous random variables. The alternative expression for the covariance leads to the following important observation.

> INDEPENDENT VERSUS UNCORRELATED.   If two random variables X and Y are independent, then X and Y are uncorrelated.

Note that the reverse is not necessarily true. If X and Y are uncorrelated, they need *not* be independent. This is illustrated in the next quick exercise.

QUICK EXERCISE 10.3 Consider the random variables X and Y with the joint distribution given in Table 10.1. Check that X and Y are dependent, but that also $\mathrm{E}[XY] = \mathrm{E}[X] \, \mathrm{E}[Y]$.

From the preceding we also deduce the following rule on the variance of the sum of two random variables.

> VARIANCE OF THE SUM.   Let X and Y be two random variables. Then always
>
> $$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).$$
>
> If X and Y are *uncorrelated*,
>
> $$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Hence, we always have that $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$, whereas $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ *only* holds for uncorrelated random variables (and hence for independent random variables!).

As with the linearity-of-expectations rule, the rule for the variance of the sum of uncorrelated random variables holds more generally. For uncorrelated random variables $X_1, X_2, \ldots, X_n$, we have

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n).$$

This rule provides an easy way to compute the variance of a random variable with a $Bin(n,p)$ distribution. Recall the representation for a $Bin(n,p)$ random variable X:

$$X = R_1 + R_2 + \cdots + R_n.$$

Each R_i has variance

$$\text{Var}(R_i) = \text{E}\left[R_i^2\right] - (\text{E}[R_i])^2 = 0^2 \cdot (1-p) + 1^2 \cdot p - (\text{E}[R_i])^2$$
$$= p - p^2 = p(1-p).$$

Using the independence of the R_i, the rule for the variance of the sum yields

$$\text{Var}(X) = \text{Var}(R_1) + \text{Var}(R_2) + \cdots + \text{Var}(R_n) = np(1-p).$$

10.3 The correlation coefficient

In the previous section we saw that the covariance between random variables gives an indication of how they influence one another. A disadvantage of the covariance is the fact that it depends on the units in which the random variables are represented. For instance, suppose that the length in inches and weight in kilograms of Dutch citizens are modeled by random variables L and W. Someone prefers to represent the length in centimeters. Since 1 inch $\equiv 2.53$ cm, one is dealing with a transformed random variable $2.53L$. The covariance between $2.53L$ and W is

$$\text{Cov}(2.53L, W) = \text{E}[(2.53L)W] - \text{E}[2.53L]\text{E}[W]$$
$$= 2.53\Big(\text{E}[LW] - \text{E}[L]\text{E}[W]\Big) = 2.53\,\text{Cov}(L, W).$$

That is, the covariance increases with a factor 2.53, which is somewhat disturbing since changing from inches to centimeters does not essentially alter the dependence between length and weight. This illustrates that the covariance changes under a change of units. The following rule provides the exact relationship.

COVARIANCE UNDER CHANGE OF UNITS.   Let X and Y be two random variables. Then

$$\text{Cov}(rX + s, tY + u) = rt\,\text{Cov}(X, Y)$$

for all numbers $r, s, t,$ and u.

See Exercise 10.14 for a derivation of this rule.

QUICK EXERCISE 10.4 For X and Y in the example in Section 9.2 (see also Section 10.2), show that $\mathrm{Cov}(-2X + 7, 5Y - 3) = 13/500$.

The preceding discussion indicates that the covariance $\mathrm{Cov}(X, Y)$ may not always be suitable to express the dependence between X and Y. For this reason there is a standardized version of the covariance called the correlation coefficient of X and Y.

> DEFINITION. Let X and Y be two random variables. The *correlation coefficient* $\rho(X, Y)$ is defined to be 0 if $\mathrm{Var}(X) = 0$ or $\mathrm{Var}(Y) = 0$, and otherwise
> $$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

Note that $\rho(X, Y)$ remains unaffected by a change of units, and therefore it is *dimensionless*. For instance, if X and Y are measured in kilometers, then $\mathrm{Cov}(X, Y)$, $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ are in km^2, so that the dimension of $\rho(X, Y)$ is in km$^2/(\sqrt{\mathrm{km}^2} \cdot \sqrt{\mathrm{km}^2})$.

For X and Y in the example in Section 9.2, recall that $\mathrm{Cov}(X, Y) = -13/5000$. We also have $\mathrm{Var}(X) = 989/2500$ and $\mathrm{Var}(Y) = 791/10\,000$ (see Exercise 10.10), so that

$$\rho(X, Y) = \frac{-\frac{13}{5000}}{\sqrt{\frac{989}{2500} \cdot \frac{791}{10\,000}}} = -0.0147.$$

QUICK EXERCISE 10.5 For X and Y in the example in Section 9.2, show that $\rho(-2X + 7, 5Y - 3) = 0.0147$.

The previous quick exercise illustrates the following linearity property for the correlation coefficient. For numbers r, s, t, and u fixed, $r, t \neq 0$, and random variables X and Y:

$$\rho(rX + s, tY + u) = \begin{cases} -\rho(X, Y) & \text{if } rt < 0, \\ \rho(X, Y) & \text{if } rt > 0. \end{cases}$$

Thus we see that the size of the correlation coefficient is unaffected by a change of units, but note the possibility of a change of sign.

Two random variables X and Y are "most correlated" if $X = Y$ or if $X = -Y$. As a matter of fact, in the former case $\rho(X, Y) = 1$, while in the latter case $\rho(X, Y) = -1$. In general—for nonconstant random variables X and Y—the following property holds:

$$-1 \leq \rho(X, Y) \leq 1.$$

For a formal derivation of this property, see the next remark.

Remark 10.2 (Correlations are between −1 and 1). Here we give a proof of the preceding formula. Since the variance of any random variable is nonnegative, we have that

$$0 \le \mathrm{Var}\left(\frac{X}{\sqrt{\mathrm{Var}(X)}} + \frac{Y}{\sqrt{\mathrm{Var}(Y)}}\right)$$

$$= \mathrm{Var}\left(\frac{X}{\sqrt{\mathrm{Var}(X)}}\right) + \mathrm{Var}\left(\frac{Y}{\sqrt{\mathrm{Var}(Y)}}\right)$$

$$+ 2\mathrm{Cov}\left(\frac{X}{\sqrt{\mathrm{Var}(X)}}, \frac{Y}{\sqrt{\mathrm{Var}(Y)}}\right)$$

$$= \frac{\mathrm{Var}(X)}{\mathrm{Var}(X)} + \frac{\mathrm{Var}(Y)}{\mathrm{Var}(Y)} + \frac{2\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = 2\left(1 + \rho(X,Y)\right).$$

This implies $\rho(X,Y) \ge -1$. Using the same argument but replacing X by $-X$ shows that $\rho(X,Y) \le 1$.

10.4 Solutions to the quick exercises

10.1 The expectation of $X + Y$ is computed as follows:

$$\mathrm{E}\left[X + Y\right] = (0+0)\cdot 0 + (1+0)\cdot\frac{1}{4} + (2+0)\cdot 0$$

$$+ (0+1)\cdot\frac{1}{4} + (1+1)\cdot 0 + (2+1)\cdot\frac{1}{4}$$

$$+ (0+2)\cdot 0 + (1+2)\cdot\frac{1}{4} + (2+2)\cdot 0 = 2.$$

10.2 First complete Table 10.1 with the marginal distributions:

		a		
b	0	1	2	$P(Y = b)$
0	0	1/4	0	1/4
1	1/4	0	1/4	1/2
2	0	1/4	0	1/4
$P(X = a)$	1/4	1/2	1/4	1

It follows that $\mathrm{E}[X] = 0\cdot\frac{1}{4} + 1\cdot\frac{1}{2} + 2\cdot\frac{1}{4} = 1$, and similarly $\mathrm{E}[Y] = 1$. Therefore $\mathrm{E}[X] + \mathrm{E}[Y] = 2$, which is equal to $\mathrm{E}[X + Y]$ as computed in Quick exercise 10.1.

10.3 From Table 10.1, as completed in Quick exercise 10.2, we see that X and Y are dependent. For instance, $P(X = 0, Y = 0) \neq P(X = 0) P(Y = 0)$. From Quick exercise 10.2 we know that $E[X] = E[Y] = 1$. Because we already computed $E[XY] = 1$, it follows that $E[XY] = E[X] E[Y]$. According to the alternative expression for the covariance this means that $\text{Cov}(X, Y) = 0$, i.e., X and Y are uncorrelated.

10.4 We already computed $\text{Cov}(X, Y) = -13/5000$ in Section 10.2. Hence, by the linearity-of-covariance rule $\text{Cov}(-2X + 7, 5Y - 3) = (-2) \cdot 5 \cdot (-13/5000) = 13/500$.

10.5 From Quick exercise 10.4 we have $\text{Cov}(-2X + 7, 5Y - 3) = 13/500$. Since $\text{Var}(X) = 989/2500$ and $\text{Var}(Y) = 791/10\,000$, by definition of the correlation coefficient and the rule for variances,

$$
\begin{aligned}
\rho(-2X + 7, 5Y - 3) &= \frac{\text{Cov}(-2X + 7, 5Y - 3)}{\sqrt{\text{Var}(-2X + 7) \cdot \text{Var}(5Y - 3)}} \\[2mm]
&= \frac{\frac{13}{500}}{\sqrt{4\text{Var}(X) \cdot 25\text{Var}(Y)}} = \frac{\frac{13}{500}}{\sqrt{\frac{3956}{2500} \cdot \frac{19775}{10\,000}}} = 0.0147.
\end{aligned}
$$

10.5 Exercises

10.1 ⊡ Consider the joint probability distribution of X and Y from Exercise 9.7, obtained from data on hair color and eye color, for which we already computed the expectations and variances of X and Y, as well as $E[XY]$.

a. Compute $\text{Cov}(X, Y)$. Are X and Y positively correlated, negative correlated, or uncorrelated?

b. Compute the correlation coefficient between X and Y.

10.2 ⊡ Consider the two discrete random variables X and Y with joint distribution derived in Exercise 9.2:

		a		
b	0	1	2	$P(Y = b)$
-1	1/6	1/6	1/6	1/2
1	0	1/2	0	1/2
$P(X = a)$	1/6	2/3	1/6	1

a. Determine $E[XY]$.

b. Note that X and Y are dependent. Show that X and Y are uncorrelated.

c. Determine Var$(X + Y)$.

d. Determine Var$(X - Y)$.

10.3 Let U and V be the two random variables from Exercise 9.6. We have seen that U and V are dependent with joint probability distribution

		a		
b	0	1	2	$P(V = b)$
0	1/4	0	1/4	1/2
1	0	1/2	0	1/2
$P(U = a)$	1/4	1/2	1/4	1

Determine the covariance $\text{Cov}(U, V)$ and the correlation coefficient $\rho(U, V)$.

10.4 Consider the joint probability distribution of the discrete random variables X and Y from the *Melencolia* Exercise 9.1. Compute $\text{Cov}(X, Y)$.

		a		
b	1	2	3	4
1	16/136	3/136	2/136	13/136
2	5/136	10/136	11/136	8/136
3	9/136	6/136	7/136	12/136
4	4/136	15/136	14/136	1/136

10.5 ⊡ Suppose X and Y are discrete random variables taking values 0,1, and 2. The following is given about the joint and marginal distributions:

		a		
b	0	1	2	$P(Y = b)$
0	8/72	...	10/72	1/3
1	12/72	9/72	...	1/2
2	...	3/72	...	...
$P(X = a)$	1/3	...	...	1

a. Complete the table.

b. Compute the expectation of X and of Y and the covariance between X and Y.

c. Are X and Y independent?

10.6 ⊞ Suppose X and Y are discrete random variables taking values $c-1$, c, and $c+1$. The following is given about the joint and marginal distributions:

b	$c-1$	c	$c+1$	$P(Y=b)$
$c-1$	2/45	9/45	4/45	1/3
c	7/45	5/45	3/45	1/3
$c+1$	6/45	1/45	8/45	1/3
$P(X=a)$	1/3	1/3	1/3	1

(The column group header above $c-1$, c, $c+1$ is a.)

a. Take $c=0$ and compute the expectation of X and of Y and the covariance between X and Y.

b. Show that X and Y are uncorrelated, no matter what the value of c is. *Hint:* one could compute $\mathrm{Cov}(X,Y)$, but there is a short solution using the rule on the covariance under change of units (see page 141) together with part **a**.

c. Are X and Y independent?

10.7 ⊡ Consider the joint distribution of Quick exercise 9.2 and take ε fixed between $-1/4$ and $1/4$:

a	0	1	$p_X(a)$
0	$1/4-\varepsilon$	$1/4+\varepsilon$	$1/2$
1	$1/4+\varepsilon$	$1/4-\varepsilon$	$1/2$
$p_Y(b)$	$1/2$	$1/2$	1

(The column group header above 0 and 1 is b.)

a. Take $\varepsilon=1/8$ and compute $\mathrm{Cov}(X,Y)$.

b. Take $\varepsilon=1/8$ and compute $\rho(X,Y)$.

c. For which values of ε is $\rho(X,Y)$ equal to -1, 0, or 1?

10.8 Let X and Y be random variables such that

$$E[X]=2,\quad E[Y]=3,\quad \text{and}\quad \mathrm{Var}(X)=4.$$

a. Show that $E[X^2]=8$.

b. Determine the expectation of $-2X^2+Y$.

10.9 ⊞ Suppose the blood of 1000 persons has to be tested to see which ones are infected by a (rare) disease. Suppose that the probability that the test

is positive is $p = 0.001$. The obvious way to proceed is to test each person, which results in a total of 1000 tests. An alternative procedure is the following. Distribute the blood of the 1000 persons over 25 groups of size 40, and mix half of the blood of each of the 40 persons with that of the others in each group. Now test the aggregated blood sample of each group: when the test is negative *no one* in that group has the disease; when the test is positive, at least one person in the group has the disease, and one will test the other half of the blood of all 40 persons of that group separately. In total, that gives 41 tests for that group. Let X_i be the total number of tests one has to perform for the ith group using this alternative procedure.

a. Describe the probability distribution of X_i, i.e., list the possible values it takes on and the corresponding probabilities.

b. What is the expected number of tests for the ith group? What is the expected total number of tests? What do you think of this alternative procedure for blood testing?

10.10 ⊞ Consider the variables X and Y from the example in Section 9.2 with joint probability density

$$f(x, y) = \frac{2}{75}(2x^2y + xy^2) \quad \text{for } 0 \le x \le 3 \text{ and } 1 \le y \le 2$$

and marginal probability densities

$$f_X(x) = \frac{2}{225}(9x^2 + 7x) \quad \text{for } 0 \le x \le 3$$

$$f_Y(y) = \frac{1}{25}(3y^2 + 12y) \quad \text{for } 1 \le y \le 2.$$

a. Compute $\mathrm{E}[X]$, $\mathrm{E}[Y]$, and $\mathrm{E}[X+Y]$.

b. Compute $\mathrm{E}\left[X^2\right]$, $\mathrm{E}\left[Y^2\right]$, $\mathrm{E}[XY]$, and $\mathrm{E}\left[(X+Y)^2\right]$,

c. Compute $\mathrm{Var}(X+Y)$, $\mathrm{Var}(X)$, and $\mathrm{Var}(Y)$ and check that $\mathrm{Var}(X+Y) \ne \mathrm{Var}(X) + \mathrm{Var}(Y)$.

10.11 Recall the relation between degrees Celsius and degrees Fahrenheit

$$\text{degrees Fahrenheit} = \frac{9}{5} \cdot \text{degrees Celsius} + 32.$$

Let X and Y be the average daily temperatures in degrees Celsius in Amsterdam and Antwerp. Suppose that $\mathrm{Cov}(X, Y) = 3$ and $\rho(X, Y) = 0.8$. Let T and S be the same temperatures in degrees Fahrenheit. Compute $\mathrm{Cov}(T, S)$ and $\rho(T, S)$.

10.12 Consider the independent random variables H and R from the vase example, with a $U(25, 35)$ and a $U(7.5, 12.5)$ distribution. Compute $\mathrm{E}[H]$ and $\mathrm{E}\left[R^2\right]$ and check that $\mathrm{E}[V] = \pi \mathrm{E}[H] \mathrm{E}\left[R^2\right]$.

10.13 Let X and Y be as in the triangle example in Exercise 9.15. Recall from Exercise 9.16 that X and Y represent the minimum and maximum coordinate of a point that is drawn from the unit square: $X = \min\{U, V\}$ and $Y = \max\{U, V\}$.

a. Show that $E[X] = 1/3$, $\mathrm{Var}(X) = 1/18$, $E[Y] = 2/3$, and $\mathrm{Var}(Y) = 1/18$. *Hint:* you might consult Exercise 8.15.

b. Check that $\mathrm{Var}(X + Y) = 1/6$, by using that U and V are independent and that $X + Y = U + V$.

c. Determine the covariance $\mathrm{Cov}(X, Y)$ using the results from **a** and **b**.

10.14 ⊞ Let X and Y be two random variables and let r, s, t, and u be arbitrary real numbers.

a. Derive from the definition that $\mathrm{Cov}(X + s, Y + u) = \mathrm{Cov}(X, Y)$.

b. Derive from the definition that $\mathrm{Cov}(rX, tY) = rt\mathrm{Cov}(X, Y)$.

c. Combine parts **a** and **b** to show $\mathrm{Cov}(rX + s, tY + u) = rt\mathrm{Cov}(X, Y)$.

10.15 In Figure 10.1 three plots are displayed. For each plot we carried out a simulation in which we generated 500 realizations of a pair of random variables (X, Y). We have chosen three different joint distributions of X and Y.
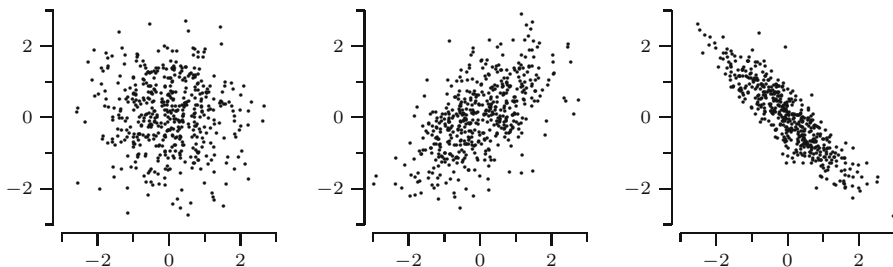
Fig. 10.1. Some scatterplots.

a. Indicate for each plot whether it corresponds to random variables X and Y that are positively correlated, negatively correlated, or uncorrelated.

b. Which plot corresponds to random variables X and Y for which $|\rho(X, Y)|$ is maximal?

10.16 ⊡ Let X and Y be random variables.

a. Express $\mathrm{Cov}(X, X + Y)$ in terms of $\mathrm{Var}(X)$ and $\mathrm{Cov}(X, Y)$.

b. Are X and $X + Y$ positively correlated, uncorrelated, or negatively correlated, or can anything happen?

c. Same question as in part **b**, but now assume that X and Y are uncorrelated.

10.17 Extending the variance of the sum rule. For mathematical convenience we first extend the sum rule to three random variables with zero expectation. Next we further extend the rule to three random variables with nonzero expectation. By the same line of reasoning we extend the rule to n random variables.

a. Let X, Y and Z be random variables with expectation 0. Show that

$$\text{Var}(X + Y + Z) = \text{Var}(X) + \text{Var}(Y) + \text{Var}(Z)$$
$$+ 2\text{Cov}(X, Y) + 2\text{Cov}(X, Z) + 2\text{Cov}(Y, Z).$$

Hint: directly apply that for real numbers $y_1, \ldots, y_n$

$$(y_1 + \cdots + y_n)^2 = y_1^2 + \cdots + y_n^2 + 2y_1y_2 + 2y_1y_3 + \cdots + 2y_{n-1}y_n.$$

b. Now show **a** for X, Y, and Z with nonzero expectation.
Hint: you might use the rules on pages 98 and 141 about variance and covariance under a change of units.

c. Derive a general variance of the sum rule, i.e., show that if $X_1, X_2, \ldots, X_n$ are random variables, then

$$\text{Var}(X_1 + X_2 + \cdots + X_n)$$
$$= \text{Var}(X_1) + \cdots + \text{Var}(X_n)$$
$$+ 2\text{Cov}(X_1, X_2) + 2\text{Cov}(X_1, X_3) + \cdots + 2\text{Cov}(X_1, X_n)$$
$$+ 2\text{Cov}(X_2, X_3) + \cdots + 2\text{Cov}(X_2, X_n)$$
$$\ddots$$
$$+ 2\text{Cov}(X_{n-1}, X_n).$$

d. Show that if the variances are all equal to σ^2 and the covariances are all equal to some constant γ, then

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = n\sigma^2 + n(n - 1)\gamma.$$

10.18 ⊞ Consider a vase containing balls numbered $1, 2, \ldots, N$. We draw n balls *without replacement* from the vase. Each ball is selected with equal probability, i.e., in the first draw each ball has probability $1/N$, in the second draw each of the $N - 1$ remaining balls has probability $1/(N - 1)$, and so on. For $i = 1, 2, \ldots, n$, let X_i denote the number on the ball in the ith draw. From Exercise 9.18 we know that the variance of X_i equals

$$\text{Var}(X_i) = \frac{1}{12}(N - 1)(N + 1).$$

Show that

$$\text{Cov}(X_1, X_2) = -\frac{1}{12}(N + 1).$$

Before you do the exercise: why do you think the covariance is negative?
Hint: use $\text{Var}(X_1 + X_2 + \cdots + X_N) = 0$ (why?), and apply Exercise 10.17.

10.19 Derive the alternative expression for the covariance: $\text{Cov}(X, Y) = \text{E}[XY] - \text{E}[X]\text{E}[Y]$.
Hint: work out $(X - \text{E}[X])(Y - \text{E}[Y])$ and use linearity of expectations.

10.20 Determine $\rho(U, U^2)$ when U has a $U(0, a)$ distribution. Here a is a positive number.

11

More computations with more random variables

Often one is interested in combining random variables, for instance, in taking the sum. In previous chapters, we have seen that it is fairly easy to describe the expected value and the variance of this new random variable. Often more details are needed, and one also would like to have its probability distribution. In this chapter we consider the probability distributions of the sum, the product, and the quotient of two random variables.

11.1 Sums of discrete random variables

In a solo race across the Pacific Ocean, a ship has one spare radio set for communications. Each of the two radios has probability p of failing each time it is switched on. The skipper uses the radio once every day. Let X be the number of days the radio is switched on until it fails (so if the radio can be used for two days and fails on the third day, X attains the value 3). Similarly, let Y be the number of days the spare radio is switched on until it fails. Note that these random variables are similar to the one discussed in Section 4.4, which modeled the number of cycles until pregnancy. Hence, X and Y are $Geo(p)$ distributed random variables. Suppose that $p = 1/75$ and that the trip will last 100 days. Then at first sight the skipper does not need to worry about radio contact: the number of days the first radio lasts is $X - 1$ days, and similarly the spare radio lasts $Y - 1$ days. Therefore the expected number of days he is able to have radio contact is

$$\mathrm{E}[X - 1 + Y - 1] = \mathrm{E}[X] + \mathrm{E}[Y] - 2 = \frac{1}{p} + \frac{1}{p} - 2 = 148 \text{ days!}$$

The skipper—who has some training in probability theory—still has some concerns about the risk he runs with these two radios. What if the probability $\mathrm{P}(X + Y - 2 \le 99)$ that his two radios break down before the end of the trip is large?

This example illustrates that it is important to study the probability distribution of the sum $Z = X + Y$ of two discrete random variables. The random variable Z takes on values $a_i + b_j$, where a_i is a possible value of X and b_j of Y. Hence, the probability mass function of Z is given by

$$p_Z(c) = \sum_{(i,j):a_i+b_j=c} P(X = a_i, Y = b_j),$$

where the sum runs over all possible values a_i of X and b_j of Y such that $a_i + b_j = c$. Because the sum only runs over values a_i that are equal to $c - b_j$, we simplify the summation and write

$$p_Z(c) = \sum_j P(X = c - b_j, Y = b_j),$$

where the sum runs over all possible values b_j of Y. When X and Y are *independent*, then $P(X = c - b_j, Y = b_j) = P(X = c - b_j) P(Y = b_j)$. This leads to the following rule.

ADDING TWO INDEPENDENT DISCRETE RANDOM VARIABLES. Let X and Y be two independent discrete random variables, with probability mass functions p_X and p_Y. Then the probability mass function p_Z of $Z = X + Y$ satisfies

$$p_Z(c) = \sum_j p_X(c - b_j) p_Y(b_j),$$

where the sum runs over all possible values b_j of Y.

QUICK EXERCISE 11.1 Let S be the sum of two independent throws with a die, so $S = X + Y$, where X and Y are independent, and $P(X = k) = P(Y = k) = 1/6$, for $k = 1, \ldots, 6$. Use the addition rule to compute $P(S = 3)$ and $P(S = 8)$, and compare your answers with Table 9.2.

In the solo race example, X and Y are independent $Geo(p)$ distributed random variables. Let $Z = X + Y$; then by the above rule for $k \geq 2$

$$P(X + Y = k) = p_Z(k) = \sum_{\ell=1}^{\infty} p_X(k - \ell) p_Y(\ell).$$

Because $p_X(a) = 0$ for $a \leq 0$, all terms in this sum with $\ell \geq k$ vanish, hence

$$P(X + Y = k) = \sum_{\ell=1}^{k-1} p_X(k - \ell) \cdot p_Y(\ell) = \sum_{\ell=1}^{k-1} (1 - p)^{k-\ell-1} p \cdot (1 - p)^{\ell-1} p$$

$$= \sum_{\ell=1}^{k-1} p^2 (1 - p)^{k-2} = (k - 1) p^2 (1 - p)^{k-2}.$$

Note that $X + Y$ does *not* have a geometric distribution.

Remark 11.1 (The expected value of a geometric distribution).
The preceding gives us the opportunity to calculate the expected value of
the geometric distribution in an easy way. Since the probabilities of Z add
up to one:

$$1 = \sum_{k=2}^{\infty} p_Z(k) = \sum_{k=2}^{\infty} (k-1)p^2(1-p)^{k-2} = p \sum_{\ell=1}^{\infty} \ell p(1-p)^{\ell-1};$$

it follows that

$$\mathrm{E}[X] = \sum_{\ell=1}^{\infty} \ell p(1-p)^{\ell-1} = \frac{1}{p}.$$

Returning to the solo race example, it is clear that the skipper does have
grounds to worry:

$$\mathrm{P}(X+Y-2 \le 99) = \mathrm{P}(X+Y \le 101) = \sum_{k=2}^{101} \mathrm{P}(X+Y=k)$$

$$= \sum_{k=2}^{101} (k-1)(\tfrac{1}{75})^2(1-\tfrac{1}{75})^{k-2} = 0.3904.$$

The sum of two binomial random variables

It is not always necessary to use the addition rule for two independent discrete
random variables to find the distribution of their sum. For example, let X and
Y be two independent random variables, where X has a $Bin(n, p)$ distribution
and Y has a $Bin(m, p)$ distribution. Since a $Bin(n, p)$ distribution models
the number of successes in n independent trials with success probability p,
heuristically, $X + Y$ represents the number of successes in $n + m$ trials with
success probability p and should therefore have a $Bin(n + m, p)$ distribution.

A more formal reasoning is the following. Let

$$R_1, R_2, \ldots, R_n, S_1, S_2, \ldots, S_m$$

be independent $Ber(p)$ distributed random variables. Recall that a $Bin(n, p)$
distributed random variable has the same distribution as the sum of n inde-
pendent $Ber(p)$ distributed random variables (see Section 4.3 or 10.2). Hence
X has the same distribution as $R_1 + R_2 + \cdots + R_n$ and Y has the same
distribution as $S_1 + S_2 + \cdots + S_m$. This means that $X + Y$ has the same dis-
tribution as the sum of $n+m$ independent $Ber(p)$ variables and therefore has
a $Bin(n + m, p)$ distribution. This can also be verified analytically by means
of the addition rule, using that X and Y are also independent.

QUICK EXERCISE 11.2 For $i = 1, 2, 3$, let X_i be a $Bin(n_i, p)$ distributed ran-
dom variable, and suppose that X_1, X_2, and X_3 are independent. Argue that
$Z = X_1 + X_2 + X_3$ is a $Bin(n_1 + n_2 + n_3, p)$ distributed random variable.

11.2 Sums of continuous random variables

Let X and Y be two continuous random variables. What can we say about the probability density function of $Z = X + Y$? We start with an example. Suppose that X and Y are two independent, $U(0,1)$ distributed random variables. One might be tempted to think that Z is also uniformly distributed.

Note that the joint probability density function f of X and Y is equal to the product of the marginal probability functions f_X and f_Y:

$$f(x,y) = f_X(x)f_Y(y) = 1 \quad \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1,$$

and $f(x,y) = 0$ otherwise. Let us compute the distribution function F_Z of Z. It is easy to see that $F_Z(a) = 0$ for $a \le 0$ and $F_Z(a) = 1$ for $a \ge 2$. For a between 0 and 1, let G be that part of the plane below the line $x + y = a$, and let Δ be the triangle with vertices $(0,0)$, $(a,0)$, and $(0,a)$; see Figure 11.1.
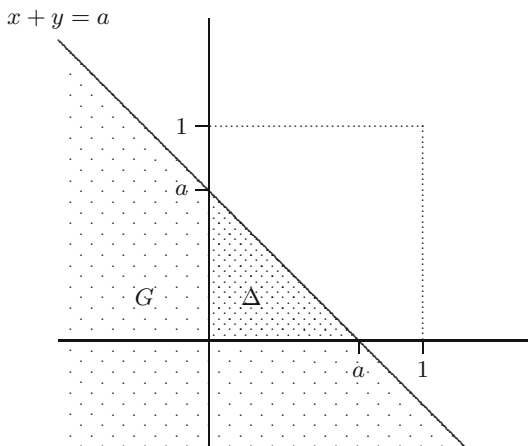
Fig. 11.1. The region G in the plane where $x + y \le a$ (with $0 < a < 1$) intersected with Δ.

Since $f(x,y) = 0$ outside $[0,1] \times [0,1]$, the distribution function of Z is given by

$$F_Z(a) = \mathrm{P}(Z \le a) = \mathrm{P}(X + Y \le a)$$
$$= \iint\limits_G f(x,y)\, \mathrm{d}x\, \mathrm{d}y = \iint\limits_\Delta 1\, \mathrm{d}x\, \mathrm{d}y = \text{area of } \Delta = \frac{1}{2}a^2$$

for $0 < a < 1$. For the case where $1 \le a < 2$ one can draw a similar figure (see Figure 11.2), from which one can find that

$$F_Z(a) = 1 - \frac{1}{2}(2-a)^2 \quad \text{for } 1 \le a < 2.$$
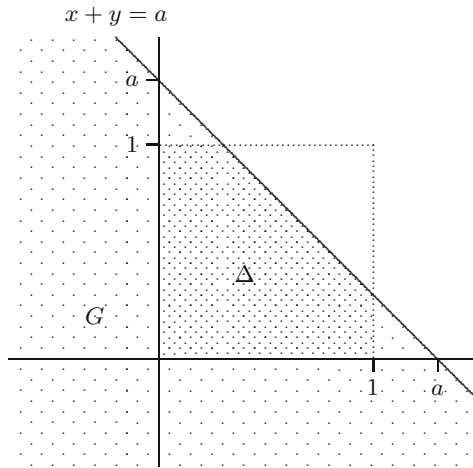
Fig. 11.2. The region G in the plane where $x + y \leq a$ (with $1 \leq a < 2$) intersected with Δ.

We see that Z is *not* uniformly distributed.

In general, the distribution function F_Z of the sum Z of two continuous random variables X and Y is given by

$$F_Z(a) = P(Z \leq a) = P(X + Y \leq a) = \iint\limits_{(x,y):x+y\leq a} f(x,y) \, dx \, dy.$$

The double integral on the right-hand side can be written as a repeated integral, first over x and then over y. Note that x and y are between minus and plus infinity and that they also have to satisfy $x + y \leq a$ or, equivalently, $x \leq a - y$. This means that the integral over x runs from minus infinity to $y - a$, and the integral over y runs from minus infinity to plus infinity. Hence

$$F_Z(a) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{a-y} f(x,y) \, dx \right) dy.$$

In case X and Y are independent, the last double integral can be written as

$$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{a-y} f_X(x) \, dx \right) f_Y(y) \, dy,$$

and we find that

$$F_Z(a) = \int_{-\infty}^{\infty} F_X(a - y) f_Y(y) \, dy$$

for $-\infty < a < \infty$. Differentiating F_Z we find the following rule.

ADDING TWO INDEPENDENT CONTINUOUS RANDOM VARIABLES.
Let X and Y be two independent continuous random variables, with
probability density functions f_X and f_Y. Then the probability den-
sity function f_Z of $Z = X + Y$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) \, dy$$

for $-\infty < z < \infty$.

The single-server queue revisited

In the single-server queue model from Section 6.4, T_1 is the time between
the start at time zero and the arrival of the first customer and T_i is the
time between the arrival of the $(i - 1)$th and ith customer at a well. We are
interested in the arrival time of the nth customer at the well. For $n \geq 1$, let
Z_n be the arrival time of the nth customer at the well: $Z_n = T_1 + \cdots + T_n$.
Since each T_i has an $Exp(0.5)$ distribution, it follows from the linearity-of-
expectations rule in Section 10.1 that the expected arrival time of the nth
customer is

$$\mathrm{E}[Z_n] = \mathrm{E}[T_1 + \cdots + T_n] = \mathrm{E}[T_1] + \cdots + \mathrm{E}[T_n] = 2n \text{ minutes.}$$

We would like to know whether the pump capacity is sufficient; for instance,
when the service times S_i are independent $U(2, 5)$ distributed random vari-
ables (this is the case when the pump capacity $v = 1$). In that case, *at most*
30 customers can pump water at the well in the first hour. If $\mathrm{P}(Z_{30} \leq 60)$ is
large, one might be tempted to increase the capacity of the well.

Recalling that the T_i are independent $Exp(\lambda)$ random variables, it follows
from the addition rule that $f_{T_1+T_2}(z) = 0$ if $z < 0$, and for $z \geq 0$ that

$$f_{Z_2}(z) = f_{T_1+T_2}(z) = \int_{-\infty}^{\infty} f_{T_1}(z - y) f_{T_2}(y) \, dy$$

$$= \int_{0}^{z} \lambda e^{-\lambda(z-y)} \cdot \lambda e^{-\lambda y} \, dy$$

$$= \lambda^2 e^{-\lambda z} \int_{0}^{z} dy = \lambda^2 z e^{-\lambda z}.$$

Viewing $T_1 + T_2 + T_3$ as the sum of T_1 and $T_2 + T_3$, we find, by applying the
addition rule again, that $f_{Z_3}(z) = 0$ if $z < 0$, and for $z \geq 0$ that

$$f_{Z_3}(z) = f_{T_1+T_2+T_3}(z) = \int_{-\infty}^{\infty} f_{T_1}(z - y) f_{T_2+T_3}(y) \, dy$$

$$= \int_{0}^{z} \lambda e^{-\lambda(z-y)} \cdot \lambda^2 y e^{-\lambda y} \, dy$$

$$= \lambda^3 e^{-\lambda z} \int_{0}^{z} y \, dy = \frac{1}{2} \lambda^3 z^2 e^{-\lambda z}.$$

Repeating this procedure, we find that $f_{Z_n}(z) = 0$ if $z < 0$, and

$$f_{Z_n}(z) = \frac{\lambda (\lambda z)^{n-1} e^{-\lambda z}}{(n-1)!}$$

for $z \geq 0$. Using integration by parts we find (see Exercise 11.13) that for $n \geq 1$ and $a \geq 0$:

$$P(Z_n \leq a) = 1 - e^{-\lambda a} \sum_{i=0}^{n-1} \frac{(\lambda a)^i}{i!}.$$

Since $\lambda = 1/2$, it follows that

$$P(Z_{30} \leq 60) = 0.524.$$

Even if each customer fills his jerrican in the minimum time of 2 minutes, we see that after an hour with probability 0.524, people will be waiting at the pump!

The random variable Z_n is an example of a *gamma random variable*, defined as follows.

> DEFINITION.   A continuous random variable X has a *gamma distribution* with parameters $\alpha > 0$ and $\lambda > 0$ if its probability density function f is given by $f(x) = 0$ for $x < 0$ and
>
> $$f(x) = \frac{\lambda (\lambda x)^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \qquad \text{for } x \geq 0,$$
>
> where the quantity $\Gamma(\alpha)$ is a normalizing constant such that f integrates to 1. We denote this distribution by $Gam(\alpha, \lambda)$.

The quantity $\Gamma(\alpha)$ is for $\alpha > 0$ defined by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} \, dt.$$

It satisfies for $\alpha > 0$ and $n = 1, 2, \ldots$

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \quad \text{and} \quad \Gamma(n) = (n-1)!$$

(see also Exercise 11.12). It follows from our example that the sum of n independent $Exp(\lambda)$ distributed random variables has a $Gam(n, \lambda)$ distribution, also known as the Erlang-n distribution with parameter λ.

The sum of independent normal random variables

Using the addition rule you can show that the sum of two independent normally distributed random variables is *again* a normally distributed random

variable. For instance, if X and Y are independent $N(0,1)$ distributed random variables, one has

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)\,dy$$

$$= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-y)^2}\right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\right) dy$$

$$= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}}\right)^2 e^{-\frac{1}{2}(2y^2-2yz+z^2)}\,dy.$$

To prepare a change of variables, we subtract the term $\frac{1}{2}z^2$ from $2y^2 - 2yz + z^2$ to complete the square in the exponent:

$$2y^2 - 2yz + \frac{1}{2}z^2 = \left[\sqrt{2}\left(y - \frac{z}{2}\right)\right]^2.$$

In this way we find with changing integration variables $t = \sqrt{2}(y - z/2)$:

$$f_{X+Y}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{4}z^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(2y^2-2yz+\frac{1}{2}z^2)}\,dy$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{4}z^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[\sqrt{2}(y-z/2)]^2}\,dy$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{4}z^2} \frac{1}{\sqrt{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\,dt$$

$$= \frac{1}{\sqrt{4\pi}} e^{-\frac{1}{4}z^2} \int_{-\infty}^{\infty} \phi(t)\,dt.$$

Since ϕ is the probability density of the standard normal distribution, it integrates to 1, so that

$$f_{X+Y}(z) = \frac{1}{\sqrt{4\pi}} e^{-\frac{1}{4}z^2},$$

which is the probability density of the $N(0,2)$ distribution. Thus, $X + Y$ also has a normal distribution. This is more generally true.

> THE SUM OF INDEPENDENT NORMAL RANDOM VARIABLES. If X and Y are independent random variables with a normal distribution, then $X + Y$ also has a normal distribution.

QUICK EXERCISE 11.3 Let X and Y be independent random variables, where X has an $N(3, 16)$ distribution, and Y an $N(5, 9)$ distribution. Then $X + Y$ is a normally distributed random variable. What are its parameters?

Rather surprisingly, independence of X and Y is not a prerequisite, as can be seen in the following remark.

Remark 11.2 (Sums of dependent normal random variables). We say the pair X, Y is has a bivariate normal distribution if their joint probability density equals

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\frac{1}{(1-\rho^2)}Q(x,y)\right),$$

where

$$Q(x,y) = \left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}.$$

Here μ_X and μ_Y are the expectations of X and Y, σ_X^2 and σ_Y^2 are their variances, and ρ is the correlation coefficient of X and Y. If X and Y have such a bivariate normal distribution, then X has an $N(\mu_X, \sigma_X^2)$ and Y has an $N(\mu_Y, \sigma_Y^2)$ distribution. Moreover, one can show that $X + Y$ has an $N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y)$ distribution. An example of a bivariate normal probability density is displayed in Figure 9.2. This probability density corresponds to parameters $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1/6$, and $\rho = 0.8$.

11.3 Product and quotient of two random variables

Recall from Chapter 7 the example of the architect who wants maximal variety in the sizes of buildings. The architect wants more variety and therefore replaces the square buildings by rectangular buildings: the buildings should be of width X and depth Y, where X and Y are independent and uniformly distributed between 0 and 10 meters. Since X and Y are independent, the expected area of a building equals $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y] = 5 \cdot 5 = 25\,\mathrm{m}^2$. But what can one say about the *distribution* of the area $Z = XY$ of an arbitrary building?

Let us calculate the distribution function of Z. Clearly $F_Z(a) = 0$ if $a < 0$ and $F_Z(a) = 1$ if $a > 100$. For a between 0 and 100 we can compute $F_Z(a)$ with the help of Figure 11.3.

We find

$$\begin{aligned} F_Z(a) &= \mathrm{P}(Z \le a) = \mathrm{P}(XY \le a) \\ &= \frac{\text{area of the shaded region in Figure 11.3}}{\text{area of }[0,10]\times[0,10]} \\ &= \frac{1}{100}\left(\frac{a}{10}\cdot 10 + \int_{a/10}^{10}\frac{a}{x}\,\mathrm{d}x\right) \\ &= \frac{1}{100}\left(a + [a\ln x]_{a/10}^{10}\right) = \frac{a(1 + 2\ln 10 - \ln a)}{100}. \end{aligned}$$

Hence the probability density function f_Z of Z is given by
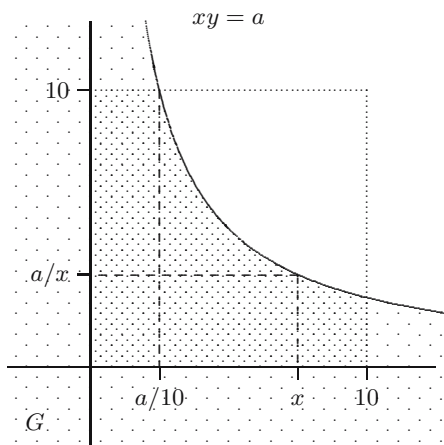
Fig. 11.3. The region G in the plane where $xy \leq a$ intersected with $[0, 10] \times [0, 10]$.

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{d}{dz} \frac{z(1 + 2\ln 10 - \ln z)}{100} = \frac{\ln 100 - \ln z}{100}$$

for $0 < z < 100\,\mathrm{m}^2$.

This computation can be generalized to arbitrary independent continuous random variables, and we obtain the following formula for the probability density function of the product of two random variables.

PRODUCT OF INDEPENDENT CONTINUOUS RANDOM VARIABLES. Let X and Y be two independent continuous random variables with probability densities f_X and f_Y. Then the probability density function f_Z of $Z = XY$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y \left(\frac{z}{x}\right) f_X(x) \frac{1}{|x|}\, dx$$

for $-\infty < z < \infty$.

For the quotient $Z = X/Y$ of two independent random variables X and Y it is now fairly easy to derive the probability density function. Since the independence of X and Y implies that X and $1/Y$ are independent, the preceding rule yields

$$f_Z(z) = \int_{-\infty}^{\infty} f_{1/Y} \left(\frac{z}{x}\right) f_X(x) \frac{1}{|x|}\, dx.$$

Recall from Section 8.2 that the probability density function of $1/Y$ is given by

$$f_{1/Y}(y) = \frac{1}{y^2} f_Y \left(\frac{1}{y}\right).$$

Substituting this in the integral, after changing the variable of integration, we find the following rule.

QUOTIENT OF INDEPENDENT CONTINUOUS RANDOM VARIABLES. Let X and Y be two independent continuous random variables with probability densities f_X and f_Y. Then the probability density function f_Z of $Z = X/Y$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(zx) f_Y(x) |x| \, dx$$

for $-\infty < z < \infty$.

The quotient of two independent normal random variables

Let X and Y be independent random variables, both having a standard normal distribution. When we compute the quotient Z of X and Y, we find a so-called *standard Cauchy distribution*:

$$
\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} |x| \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2 x^2} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \right) dx \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} |x| e^{-\frac{1}{2}(z^2+1)x^2} \, dx = 2 \cdot \frac{1}{2\pi} \int_0^{\infty} x e^{-\frac{1}{2}(z^2+1)x^2} \, dx \\
&= \frac{1}{\pi} \left[\frac{-1}{z^2+1} e^{-\frac{1}{2}(z^2+1)x^2} \right]_0^{\infty} = \frac{1}{\pi(z^2+1)}.
\end{aligned}
$$

This is the special case $\alpha = 0$, $\beta = 1$ of the following family of distributions.

DEFINITION. A continuous random variable has a *Cauchy distribution* with parameters α and $\beta > 0$ if its probability density function f is given by

$$f(x) = \frac{\beta}{\pi \left(\beta^2 + (x - \alpha)^2 \right)} \qquad \text{for } -\infty < x < \infty.$$

We denote this distribution by $Cau(\alpha, \beta)$.

By integrating, we find that the distribution function F of a Cauchy distribution is given by

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan \left(\frac{x - \alpha}{\beta} \right).$$

The parameter α is the point of symmetry of the probability density function f. Note that α is *not* the expected value of Z. As a matter of fact, it was shown in Remark 7.1 that the expected value does not exist! The probability density f is shown together with the distribution function F for the case $\alpha = 2$, $\beta = 5$ in Figure 11.4.
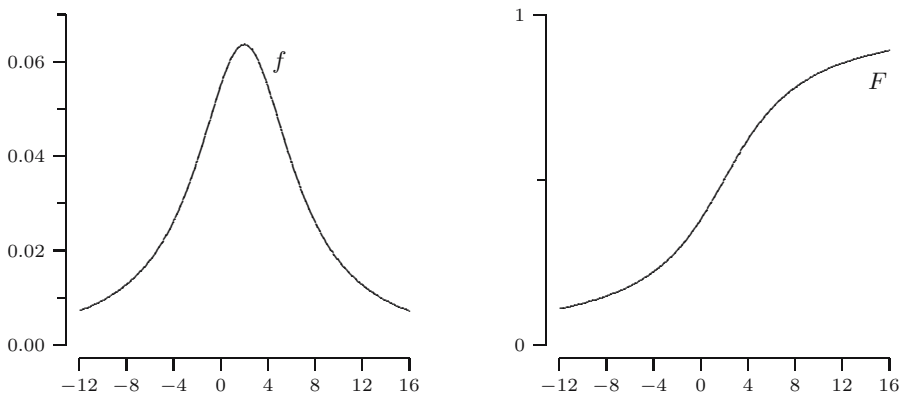
Fig. 11.4. The graphs of f and F of the $Cau(2,5)$ distribution.

QUICK EXERCISE 11.4 Argue—without doing *any* calculations—that if Z has a standard Cauchy distribution, $1/Z$ also has a standard Cauchy distribution.

11.4 Solutions to the quick exercises

11.1 Using the addition rule we find

$$P(S = 3) = \sum_{j=1}^{6} p_X(3 - j)p_Y(j)$$
$$= p_X(2)p_Y(1) + p_X(1)p_Y(2) + p_X(0)p_Y(3)$$
$$\quad + p_X(-1)p_Y(4) + p_X(-2)p_Y(5) + p_X(-3)p_Y(6)$$
$$= \frac{1}{36} + \frac{1}{36} + 0 + 0 + 0 + 0 = \frac{1}{18}$$

and

$$P(S = 8) = \sum_{j=1}^{6} p_X(8 - j)p_Y(j)$$
$$= p_X(7)p_Y(1) + p_X(6)p_Y(2) + p_X(5)p_Y(3)$$
$$\quad + p_X(4)p_Y(4) + p_X(3)p_Y(5) + p_X(2)p_Y(6)$$
$$= 0 + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{5}{36}.$$

11.2 We have seen that $X_1 + X_2$ is a $Bin(n_1 + n_2, p)$ distributed random variable. Viewing $X_1 + X_2 + X_3$ as the sum of $X_1 + X_2$ and X_3, it follows that $X_1 + X_2 + X_3$ is a $Bin(n_1 + n_2 + n_3, p)$ distributed random variable.

11.3 The sum rule for two normal random variables tells us that $X + Y$ is a normally distributed random variable. Its parameters are expectation and variance of $X + Y$. Hence by linearity of expectations

$$\mu_{X+Y} = \text{E}[X + Y] = \text{E}[X] + \text{E}[Y] = \mu_X + \mu_Y = 3 + 5 = 8,$$

and by the rule for the variance of the sum

$$\sigma_{X+Y}^2 = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) = \sigma_X^2 + \sigma_Y^2 = 16 + 9 = 25,$$

using that $\text{Cov}(X, Y) = 0$ due to independence of X and Y.

11.4 In the examples we have seen that the quotient X/Y of two independent standard normal random variables has a standard Cauchy distribution. Since $Z = X/Y$, the random variable $1/Z = Y/X$. This is *also* the quotient of two independent standard normal random variables, and it has a standard Cauchy distribution.

11.5 Exercises

11.1 ⊡ Let X and Y be independent random variables with a *discrete uniform* distribution, i.e., with probability mass functions

$$p_X(k) = p_Y(k) = \frac{1}{N}, \quad \text{for } k = 1, \ldots, N.$$

Use the addition rule for discrete random variables on page 152 to determine the probability mass function of $Z = X + Y$ for the following two cases.

a. Suppose $N = 6$, so that X and Y represent two throws with a die. Show that

$$p_Z(k) = \text{P}(X + Y = k) = \begin{cases} \dfrac{k-1}{36} & \text{for } k = 2, \ldots, 6, \\ \dfrac{13-k}{36} & \text{for } k = 7, \ldots, 12. \end{cases}$$

You may check this with Quick exercise 11.1.

b. Determine the expression for $p_Z(k)$ for general N.

11.2 ⊞ Consider a discrete random variable X taking values $k = 0, 1, 2, \ldots$ with probabilities

$$\text{P}(X = k) = \frac{\mu^k}{k!}\text{e}^{-\mu},$$

where $\mu > 0$. This is the *Poisson* distribution with parameter μ. We will learn more about this distribution in Chapter 12. This exercise illustrates that the sum of independent Poisson variables again has a Poisson distribution.

a. Let X and Y be independent random variables, each having a Poisson distribution with $\mu = 1$. Show that for $k = 0, 1, 2, \ldots$

$$P(X + Y = k) = \frac{2^k}{k!}e^{-2},$$

by using $\sum_{\ell=0}^{k} \binom{k}{\ell} = 2^k$.

b. Let X and Y be independent random variables, each having a Poisson distribution with parameters λ and μ. Show that for $k = 0, 1, 2, \ldots$

$$P(X + Y = k) = \frac{(\lambda + \mu)^k}{k!}e^{-(\lambda+\mu)},$$

by using $\sum_{\ell=0}^{k} \binom{k}{\ell}p^\ell(1-p)^{k-\ell} = 1$ for $p = \mu/(\lambda + \mu)$.
We conclude that $X + Y$ has a Poisson distribution with parameter $\lambda + \mu$.

11.3 Let X and Y be two independent random variables, where X has a $Ber(p)$ distribution, and Y has a $Ber(q)$ distribution. When $p = q = r$, we know that $X + Y$ has a $Bin(2, r)$ distribution. Suppose that $p = 1/2$ and $q = 1/4$. Determine $P(X + Y = k)$, for $k = 0, 1, 2$, and conclude that $X + Y$ does not have a binomial distribution.

11.4 ⊞ Let X and Y be two independent random variables, where X has an $N(2, 5)$ distribution and Y has an $N(5, 9)$ distribution. Define $Z = 3X - 2Y + 1$.

a. Compute $E[Z]$ and $Var(Z)$.

b. What is the distribution of Z?

c. Compute $P(Z \leq 6)$.

11.5 ⊡ Let X and Y be two independent, $U(0, 1)$ distributed random variables. Use the rule on addition of independent continuous random variables on page 156 to show that the probability density function of $X + Y$ is given by

$$f_Z(z) = \begin{cases} z & \text{for } 0 \leq z < 1, \\ 2 - z & \text{for } 1 \leq z \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

11.6 ⊡ Let X and Y be independent random variables with probability densities

$$f_X(x) = \frac{1}{4}xe^{-x/2} \quad \text{and} \quad f_Y(y) = \frac{1}{4}ye^{-y/2}.$$

Use the rule on addition of independent continuous random variables to determine the probability density of $Z = X + Y$.

11.7 ⊡ The two random variables in Exercise 11.6 are special cases of $Gam(\alpha, \lambda)$ variables, namely with $\alpha = 2$ and $\lambda = 1/2$. More generally, let

$X_1, \ldots, X_n$ be independent $Gam(k, \lambda)$ distributed random variables, where $\lambda > 0$ and k is a positive integer. Argue—without doing any calculations— that $X_1 + \cdots + X_n$ has a $Gam(nk, \lambda)$ distribution.

11.8 We investigate the effect on the Cauchy distribution under a change of units.

a. Let X have a standard Cauchy distribution. What is the distribution of $Y = rX + s$?

b. Let X have a $Cau(\alpha, \beta)$ distribution. What is the distribution of the random variable $(X - \alpha)/\beta$?

11.9 ⊞ Let X and Y be independent random variables with a $Par(\alpha)$ and $Par(\beta)$ distribution.

a. Take $\alpha = 3$ and $\beta = 1$ and determine the probability density of $Z = XY$.

b. Determine the probability density of $Z = XY$ for general α and β.

11.10 Let X and Y be independent random variables with a $Par(\alpha)$ and $Par(\beta)$ distribution.

a. Take $\alpha = \beta = 2$. Show that $Z = X/Y$ has probability density

$$f_Z(z) = \begin{cases} z & \text{for } 0 < z < 1, \\ 1/z^3 & \text{for } 1 \le z < \infty. \end{cases}$$

b. For general $\alpha, \beta > 0$, show that $Z = X/Y$ has probability density

$$f_Z(z) = \begin{cases} \dfrac{\alpha\beta}{\alpha + \beta} z^{\beta-1} & \text{for } 0 < z < 1, \\[2ex] \dfrac{\alpha\beta}{\alpha + \beta} \dfrac{1}{z^{\alpha+1}} & \text{for } 1 \le z < \infty. \end{cases}$$

11.11 Let X_1, X_2, and X_3 be three independent $Geo(p)$ distributed random variables, and let $Z = X_1 + X_2 + X_3$.

a. Show for $k \ge 3$ that the probability mass function p_Z of Z is given by

$$p_Z(k) = P(X_1 + X_2 + X_3 = k) = \frac{1}{2}(k - 2)(k - 1)p^3(1 - p)^{k-3}.$$

b. Use the fact that $\sum_{k=3}^{\infty} p_Z(k) = 1$ to show that

$$p^2 \left(E\left[X_1^2\right] + E[X_1] \right) = 2.$$

c. Use $E[X_1] = 1/p$ and part **b** to conclude that

$$E\left[X_1^2\right] = \frac{2 - p}{p^2} \quad \text{and} \quad \text{Var}(X_1) = \frac{1 - p}{p^2}.$$

11.12 Show that $\Gamma(1) = 1$, and use integration by parts to show that

$$\Gamma(x+1) = x\Gamma(x) \quad \text{for } x > 0.$$

Use this last expression to show for $n = 1, 2, \ldots$ that

$$\Gamma(n) = (n-1)!$$

11.13 Let Z_n have an Erlang-n distribution with parameter λ.

a. Use integration by parts to show that for $a \geq 0$ and $n \geq 2$:

$$P(Z_n \leq a) = \int_0^a \frac{\lambda^n z^{n-1} e^{-\lambda z}}{(n-1)!} \, dz = -\frac{(\lambda a)^{n-1}}{(n-1)!} e^{-\lambda a} + P(Z_{n-1} \leq a).$$

b. Use **a** to show that for $a \geq 0$:

$$P(Z_n \leq a) = -\sum_{i=1}^{n-1} \frac{(\lambda a)^i}{i!} e^{-\lambda a} + P(Z_1 \leq a).$$

c. Conclude that for $a \geq 0$:

$$P(Z_n \leq a) = 1 - e^{-\lambda a} \sum_{i=0}^{n-1} \frac{(\lambda a)^i}{i!}.$$

12

The Poisson process

In many random phenomena we encounter, it is not just one or two random variables that play a role but a whole collection. In that case one often speaks of a random *process*. The Poisson process is a simple kind of random process, which models the occurrence of random points in time or space. There are numerous ways in which processes of random points arise: some examples are presented in the first section. The Poisson process describes in a certain sense the *most random way* to distribute points in time or space. This is made more precise with the notions of homogeneity and independence.

12.1 Random points

Typical examples of the occurrence of random time points are: arrival times of email messages at a server, the times at which asteroids hit the earth, arrival times of radioactive particles at a Geiger counter, times at which your computer crashes, the times at which electronic components fail, and arrival times of people at a pump in an oasis.

Examples of the occurrence of random points in space are: the locations of asteroid impacts with earth (2-dimensional), the locations of imperfections in a material (3-dimensional), and the locations of trees in a forest (2-dimensional).

Some of these phenomena are better modeled by the Poisson process than others. Loosely speaking, one might say that the Poisson process model often applies in situations where there is a very large population, and each member of the population has a very small probability to produce a point of the process. This is, for instance, well fulfilled in the Geiger counter example where, in a huge collection of atoms, just a few will emit a radioactive particle (see [28]). A property of the Poisson process—as we will see shortly—is that points may lie arbitrarily close together. Therefore the tree locations are not so well modeled by the Poisson process.

12.2 Taking a closer look at random arrivals

A well-known example that is usually modeled by the Poisson process is that of calls arriving at a telephone exchange—the exchange is connected to a large number of people who make phone calls now and then. This will be our leading example in this section.

Telephone calls arrive at random times $X_1, X_2, \ldots$ at the telephone exchange during a time interval $[0, t]$.
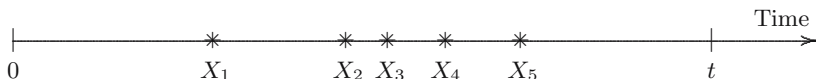
The two basic assumptions we make on these random arrivals are

1. (*Homogeneity*) The rate λ at which arrivals occur is constant over time: in a subinterval of length u the expectation of the number of telephone calls is λu.
2. (*Independence*) The numbers of arrivals in disjoint time intervals are independent random variables.

Homogeneity is also called weak stationarity. We denote the total number of calls in an interval I by $N(I)$, abbreviating $N([0, t])$ to N_t. Homogeneity then implies that we require

$$\mathrm{E}[N_t] = \lambda t.$$

To get hold of the *distribution* of N_t we divide the interval $[0, t]$ into n intervals of length t/n. When n is large enough, every interval $I_{j,n} = ((j-1)\, t/n, j\, t/n]$ will contain either 0 or 1 arrival: For such a large n (which also satisfies
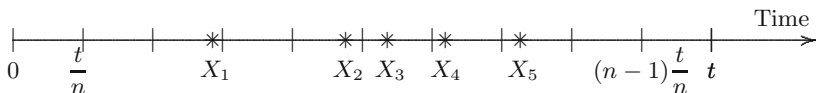
$n > \lambda t$), let R_j be the number of arrivals in the time interval $I_{j,n}$. Since R_j is 0 or 1, R_j has a $Ber(p_j)$ distribution for some p_j. Recall that for a Bernoulli random variable $\mathrm{E}[R_j] = 0 \cdot (1 - p_j) + 1 \cdot p_j = p_j$. By the homogeneity assumption, for each j

$$p_j = \lambda \cdot \text{length of } I_{j,n} = \frac{\lambda t}{n}.$$

Summing the number of calls in the intervals gives the total number of calls, hence

$$N_t = R_1 + R_2 + \cdots + R_n.$$

By the independence assumption, the R_j are independent random variables, therefore N_t has a $Bin(n,p)$ distribution, with $p = \lambda t/n$.

Remark 12.1 (About this approximation). The argument just given seems pretty convincing, but actually R_j does *not* have a Bernoulli distribution, whatever the value of n. A way to see this is the following. Every interval $I_{j,n}$ is a union of the two intervals $I_{2j-1,2n}$ and $I_{2j,2n}$. Hence the probability that $I_{j,n}$ contains two calls is at least $(\lambda t/2n)^2 = \lambda^2 t^2/4n^2$, which is larger than zero.

Note however, that the probability of having two arrivals is of smaller order than the probability that R_j takes the value 1. If we add a third assumption, namely that the probability of two or more calls arriving in an interval $I_{j,n}$ tends to zero faster than $1/n$, then the conclusion below on the distribution of N_t is valid.

We have found that (at least in first approximation)

$$P(N_t = k) = \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} \qquad \text{for } k = 0, \ldots, n.$$

In this analysis n is a rather artificial parameter, of which we only know that it should not be "too small." It therefore seems a good idea to get rid of n by letting n go to infinity, hoping that the probability distribution of N_t will settle down. Note that

$$\lim_{n\to\infty} \binom{n}{k} \frac{1}{n^k} = \lim_{n\to\infty} \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{(n-k+1)}{n} \cdot \frac{1}{k!} = \frac{1}{k!},$$

and from calculus we know that

$$\lim_{n\to\infty} \left(1 - \frac{\lambda t}{n}\right)^n = e^{-\lambda t}.$$

Since certainly

$$\lim_{n\to\infty} \left(1 - \frac{\lambda t}{n}\right)^{-k} = 1,$$

we obtain, combining these three limits, that

$$\lim_{n\to\infty} P(N_t = k) = \lim_{n\to\infty} \binom{n}{k} \frac{1}{n^k} \cdot (\lambda t)^k \cdot \left(1 - \frac{\lambda t}{n}\right)^n \cdot \left(1 - \frac{\lambda t}{n}\right)^{-k} = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

Since

$$e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} = e^{-\lambda t} e^{\lambda t} = 1,$$

we have indeed run into a probability distribution on the numbers $0, 1, 2, \ldots$. Note that all these probabilities are determined by the single value λt. This motivates the following definition.

> DEFINITION. A discrete random variable X has a *Poisson distribution* with parameter μ, where $\mu > 0$ if its probability mass function p is given by
>
> $$p(k) \;=\; \mathrm{P}(X = k) \;=\; \frac{\mu^k}{k!}\,\mathrm{e}^{-\mu} \quad \text{for } k = 0, 1, 2, \ldots.$$
>
> We denote this distribution by $Pois(\mu)$.

Figure 12.1 displays the graphs of the probability mass functions of the Poisson distribution with $\mu = 0.9$ (left) and the Poisson distribution with $\mu = 5$ (right).
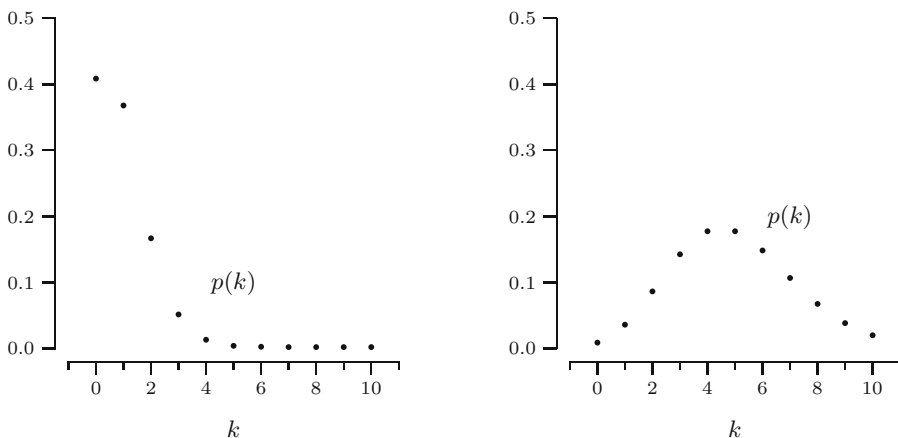
Fig. 12.1. The probability mass functions of the $Pois(0.9)$ and the $Pois(5)$ distributions.

QUICK EXERCISE 12.1 Consider the event "exactly one call arrives in the interval $[0, 2s]$." The probability of this event is $\mathrm{P}(N_{2s} = 1) = \lambda \cdot 2s \cdot \mathrm{e}^{-\lambda \cdot 2s}$. But note that this event is the same as "there is exactly one call in the interval $[0, s)$ and no calls in the interval $[s, 2s]$, *or* no calls in $[0, s)$ and exactly one call in $[s, 2s]$." Verify (using assumptions 1 and 2) that you get the same answer if you compute the probability of the event in this way.

We do have a hint[1] about what the expectation and variance of a Poisson random variable might be: since $\mathrm{E}[N_t] = \lambda t$ for all n, we anticipate that the limiting Poisson distribution will have expectation λt. Similarly, since N_t has a $Bin(n, \frac{\lambda t}{n})$ distribution, we anticipate that the variance will be

---

[1] This is really not more than a hint: there are simple examples where the distributions of random variables converge to a distribution whose expectation is different from the limit of the expectations of the distributions! (cf. Exercise 12.14).

$$\lim_{n\to\infty} \mathrm{Var}(N_t) = \lim_{n\to\infty} n \cdot \frac{\lambda t}{n} \cdot \left(1 - \frac{\lambda t}{n}\right) = \lambda t.$$

Actually, the expectation of a Poisson random variable X with parameter μ is easy to compute:

$$\mathrm{E}[X] = \sum_{k=0}^{\infty} k\frac{\mu^k}{k!}\mathrm{e}^{-\mu} = \mathrm{e}^{-\mu}\sum_{k=1}^{\infty} \frac{\mu^k}{(k-1)!}$$

$$= \mu\mathrm{e}^{-\mu}\sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} = \mu\mathrm{e}^{-\mu}\sum_{j=0}^{\infty} \frac{\mu^j}{j!} = \mu.$$

In a similar way the variance can be determined (see Exercise 12.8), and we arrive at the following rule.

THE EXPECTATION AND VARIANCE OF A POISSON DISTRIBUTION. Let X have a Poisson distribution with parameter μ; then

$$\mathrm{E}[X] = \mu \quad \text{and} \quad \mathrm{Var}(X) = \mu.$$

12.3 The one-dimensional Poisson process

We will derive some properties of the sequence of random points $X_1, X_2, \ldots$ that we considered in the previous section. What we derived so far is that for any interval $(s, s+t]$ the number $N((s, s+t])$ of points X_i in that interval is a random variable with a $Pois(\lambda t)$ distribution.

Interarrival times

The differences

$$T_i = X_i - X_{i-1}$$

are called interarrival times. Here we define $T_1 = X_1$, the time of the *first* arrival. To determine the probability distribution of T_1, we observe that the event $\{T_1 > t\}$ that the first call arrives after time t is the same as the event $\{N_t = 0\}$ that no calls have been made in $[0, t]$. But this implies that

$$\mathrm{P}(T_1 \leq t) = 1 - \mathrm{P}(T_1 > t) = 1 - \mathrm{P}(N_t = 0) = 1 - \mathrm{e}^{-\lambda t}.$$

Therefore T_1 has an exponential distribution with parameter λ.

To compute the joint distribution of T_1 and T_2, we consider the conditional probability that $T_2 > t$, given that $T_1 = s$, and use the property that arrivals in different intervals are independent:

$$P(T_2 > t \,|\, T_1 = s) = P(\text{no arrivals in } (s, s+t] \,|\, T_1 = s)$$
$$= P(\text{no arrivals in } (s, s+t])$$
$$= P(N((s, s+t]) = 0) = e^{-\lambda t}.$$

Since this answer does not depend on s, we conclude that T_1 and T_2 are independent, and

$$P(T_2 > t) = e^{-\lambda t},$$

i.e., T_2 also has an exponential distribution with parameter λ. Actually, although the conclusion is correct, the method to derive it is not, because we conditioned on the event $\{T_1 = s\}$, which has zero probability. This problem could be circumvented by conditioning on the event that T_1 lies in some small interval, but that will not be done here. Analogously, one can show that the T_i are independent and have an $Exp(\lambda)$ distribution. This nice property allows us to give a simple definition of the one-dimensional Poisson process.

> DEFINITION. The one-dimensional *Poisson process* with intensity λ is a sequence $X_1, X_2, X_3, \ldots$ of random variables having the property that the interarrival times $X_1, X_2 - X_1, X_3 - X_2, \ldots$ are independent random variables, each with an $Exp(\lambda)$ distribution.

Note that the connection with N_t is as follows: N_t is equal to the number of X_i that are smaller than (or equal to) t.

QUICK EXERCISE 12.2 We model the arrivals of email messages at a server as a Poisson process. Suppose that on average 330 messages arrive per minute. What would you choose for the intensity λ in messages per second? What is the expectation of the interarrival time?

An obvious question is: what is the distribution of X_i? This has already been answered in Chapter 11: since X_i is a sum of i independent exponentially distributed random variables, we have the following.

> THE POINTS OF THE POISSON PROCESS. For $i = 1, 2, \ldots$ the random variable X_i has a $Gam(i, \lambda)$ distribution.

The distribution of points

Another interesting question is: if we know that n points are generated in an interval, where do these points lie? Since the distribution of the number of points only depends on the length of the interval, and not on its location, it suffices to determine this for an interval starting at 0. Let this interval be $[0, a]$. We start with the simplest case, where there is one point in $[0, a]$: suppose that $N([0, a]) = 1$. Then, for $0 < s < a$:

$$P(X_1 \le s \mid N([0, a]) = 1) = \frac{P(X_1 \le s, N([0, a]) = 1)}{P(N([0, a]) = 1)}$$

$$= \frac{P(N([0, s]) = 1, N((s, a]) = 0)}{P(N([0, a]) = 1)}$$

$$= \frac{\lambda s e^{-\lambda s} e^{-\lambda(a-s)}}{\lambda a e^{-\lambda a}}$$

$$= \frac{s}{a}.$$

We find that conditional on the event $\{N([0, a]) = 1\}$, the random variable X_1 is uniformly distributed over the interval $[0, a]$.

Now suppose that it is given that there are two points in $[0, a]$: $N([0, a]) = 2$. In a way similar to what we did for *one* point, we can show that (see Exercise 12.12)

$$P(X_1 \le s, X_2 \le t \mid N([0, a]) = 2) = \frac{t^2 - (t - s)^2}{a^2}.$$

Now recall the result of Exercise 9.17: if U_1 and U_2 are two independent random variables, both uniformly distributed over $[0, a]$, then the joint distribution function of $V = \min(U_1, U_2)$ and $Z = \max(U_1, U_2)$ is given by

$$P(V \le s, Z \le t) = \frac{t^2 - (t - s)^2}{a^2} \quad \text{for } 0 \le s \le t \le a.$$

Thus we have found that, if we forget about their order, the two points in $[0, a]$ are independent and uniformly distributed over $[0, a]$. With somewhat more work, this generalizes to an arbitrary number of points, and we arrive at the following property.

> LOCATION OF THE POINTS, GIVEN THEIR NUMBER. Given that the Poisson process has n points in the interval $[a, b]$, the locations of these points are independently distributed, each with a uniform distribution on $[a, b]$.

12.4 Higher-dimensional Poisson processes

Our definition of the one-dimensional Poisson process, starting with the interarrival times, does not generalize easily, because it is based on the ordering of the real numbers. However, we can easily extend the assumptions of independence, homogeneity, and the Poisson distribution property. To do this we need a higher-dimensional version of the concept of length. We denote the k-dimensional volume of a set A in k-dimensional space by $m(A)$. For instance, in the plane $m(A)$ is the area of A, and in space $m(A)$ is the volume of A.

> DEFINITION. The k-dimensional *Poisson process* with intensity λ is a collection $X_1, X_2, X_3, \ldots$ of random points having the property that if $N(A)$ denotes the number of points in the set A, then
> 1. (*Homogeneity*) The random variable $N(A)$ has a Poisson distribution with parameter $\lambda m(A)$.
> 2. (*Independence*) For disjoint sets $A_1, A_2, \ldots, A_n$ the random variables $N(A_1), N(A_2), \ldots, N(A_n)$ are independent.

QUICK EXERCISE 12.3 Suppose that the locations of defects in a certain type of material follow the two-dimensional Poisson process model. For this material it is known that it contains on average five defects per square meter. What is the probability that a strip of length 2 meters and width 5 cm will be without defects?

In Figure 7.4 the locations of the buildings the architect wanted to distribute over a 100-by-300-m terrain have been generated by a two-dimensional Poisson process. This has been done in the following way. One can again show that given the total number of points in a set, these points are uniformly distributed over the set. This leads to the following procedure: first one generates a value n from a Poisson distribution with the appropriate parameter (λ times the area), then one generates n times a point uniformly distributed over the 100-by-300 rectangle.

Actually one can generate a higher-dimensional Poisson process in a way that is very similar to the natural way this can be done for the one-dimensional process. Directly from the definition of the one-dimensional process we see that it can be obtained by consecutively generating points with exponentially distributed gaps. We will explain a similar procedure for dimension two. For $s > 0$, let

$$M_s = N(C_s),$$

where C_s is the circular region of radius s, centered at the origin. Since C_s has area πs^2, M_s has a Poisson distribution with parameter $\lambda \pi s^2$. Let R_i denote the distance of the ith closest point to the origin. This is illustrated in Figure 12.2.

Note that R_i is the analogue of the ith arrival time for the one-dimensional Poisson process: we have in fact that

$$R_i \leq s \qquad \text{if and only if} \qquad M_s \geq i.$$

In particular, with $i = 1$ and $s = \sqrt{t}$,

$$P\left(R_1^2 \leq t\right) = P\left(R_1 \leq \sqrt{t}\right) = P\left(M_{\sqrt{t}} > 0\right) = 1 - e^{-\lambda \pi t}.$$

In other words: R_1^2 is $Exp(\lambda \pi)$ distributed. For general i, we can similarly write

$$P\left(R_i^2 \leq t\right) = P\left(R_i \leq \sqrt{t}\right) = P\left(M_{\sqrt{t}} \geq i\right).$$
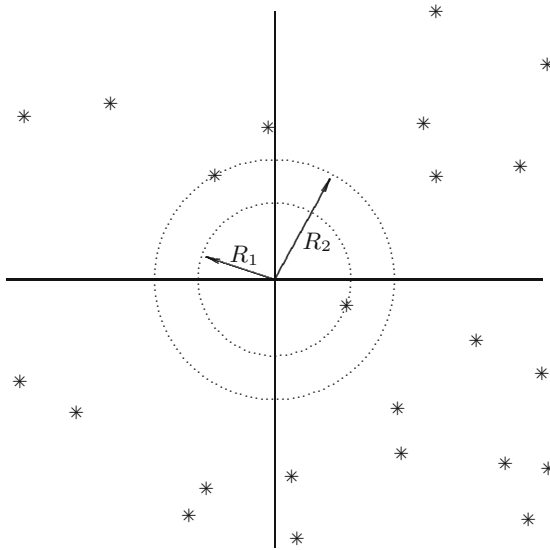
Fig. 12.2. The Poisson process in the plane, with the two circles of the two points closest to the origin.

So

$$\mathrm{P}\big(R_i^2 \leq t\big) = 1 - \mathrm{e}^{-\lambda\pi t} \sum_{j=0}^{i-1} \frac{(\lambda\pi t)^j}{j!},$$

which means that R_i^2 has a $Gam(i, \lambda\pi)$ distribution—as we saw on page 157. Since gamma distributions arise as sums of independent exponential distributions, we can also write

$$R_i^2 = R_{i-1}^2 + T_i,$$

where the T_i are independent $Exp(\lambda\pi)$ random variables (and where $R_0 = 0$). Note that this is quite similar to the one-dimensional case. To simulate the two-dimensional Poisson process from a sequence $U_1, U_2, \ldots$ of independent $U(0,1)$ random variables, one can therefore proceed as follows (recall from Section 6.2 that $-(1/\lambda)\ln(U_i)$ has an $Exp(\lambda)$ distribution): for $i = 1, 2, \ldots$ put

$$R_i = \sqrt{R_{i-1}^2 - \frac{1}{\lambda\pi} \ln(U_{2i})};$$

this gives the distance of the ith point to the origin, and then put the point on this circle according to an angle value generated by $2\pi U_{2i-1}$. This is the correct way to do it, because one can show that in polar coordinates the radius and the angle of a Poisson process point are independent of each other, and the angle is uniformly distributed over $[0, 2\pi]$. The latter is called the *isotropy* property of the Poisson process.

12.5 Solutions to the quick exercises

12.1 The probability of exactly one call in $[0, s)$ and no calls in $[s, 2s]$ equals

$$P(N([0, s)) = 1, N([s, 2s]) = 0) = P(N([0, s)) = 1) P(N([s, 2s]) = 0)$$
$$= P(N([0, s)) = 1) P(N([0, s]) = 0)$$
$$= \lambda s e^{-\lambda s} \cdot e^{-\lambda s},$$

because of independence and homogeneity. In the same way, the probability of exactly one call in $[s, 2s]$ and no calls in $[0, s)$ is equal to $e^{-\lambda s} \cdot \lambda s e^{-\lambda s}$. And indeed: $\lambda s e^{-\lambda s} \cdot e^{-\lambda s} + e^{-\lambda s} \cdot \lambda s e^{-\lambda s} = 2\lambda s e^{-\lambda \cdot 2s}$.

12.2 Because there are 60 seconds in a minute, we have $60\lambda = 330$. It follows that $\lambda = 5\frac{1}{2}$. Since the interarrival times have an $Exp(\lambda)$ distribution, the expected time between messages is $1/\lambda = 0.18$ second.

12.3 The intensity of this process is $\lambda = 5$ per m^2. The area of the strip is $2 \cdot (1/20) = 1/10 \, \text{m}^2$. Hence the probability that no defects occur in the strip is $e^{-\lambda \cdot (\text{area of strip})} = e^{-5 \cdot (1/10)} = e^{-1/2} = 0.60$.

12.6 Exercises

12.1 ⊞ In each of the following examples, try to indicate whether the Poisson process would be a good model.

  a. The times of bankruptcy of enterprises in the United States.
  b. The times a chicken lays its eggs.
  c. The times of airplane crashes in a worldwide registration.
  d. The locations of worngly spelled words in a book.
  e. The times of traffic accidents at a crossroad.

12.2 The number of customers that visit a bank on a day is modeled by a Poisson distribution. It is known that the probability of no customers at all is 0.00001. What is the expected number of customers?

12.3 Let N have a $Pois(4)$ distribution. What is $P(N = 4)$?

12.4 Let X have a $Pois(2)$ distribution. What is $P(X \leq 1)$?

12.5 ⊡ The number of errors on a hard disk is modeled as a Poisson random variable with expectation one error in every Mb, that is, in every 2^{20} bytes.

  a. What is the probability of at least one error in a sector of 512 bytes?
  b. The hard disk is an 18.62-Gb disk drive with $39\,054\,015$ sectors. What is the probability of at least one error on the hard disk?

12.6 ⊞ A certain brand of copper wire has flaws about every 40 centimeters. Model the locations of the flaws as a Poisson process. What is the probability of two flaws in 1 meter of wire?

12.7 ⊞ The Poisson model is sometimes used to study the flow of traffic ([15]). If the traffic can flow freely, it behaves like a Poisson process. A 20-minute time interval is divided into 10-second time slots. At a certain point along the highway the number of passing cars is registered for each 10-second time slot. Let n_j be the number of slots in which j cars have passed for $j = 0, \ldots, 9$. Suppose that one finds

j	0	1	2	3	4	5	6	7	8	9
n_j	19	38	28	20	7	3	4	0	0	1

Note that the total number of cars passing in these 20 minutes is 230.

   a. What would you choose for the intensity parameter λ?

   b. Suppose one estimates the probability of 0 cars passing in a 10-second time slot by n_0 divided by the total number of time slots. Does that (reasonably) agree with the value that follows from your answer in **a**?

   c. What would you take for the probability that 10 cars pass in a 10-second time slot?

12.8 ⊡ Let X be a Poisson random variable with parameter μ.

   a. Compute $\mathrm{E}[X(X-1)]$.

   b. Compute $\mathrm{Var}(X)$, using that $\mathrm{Var}(X) = \mathrm{E}[X(X-1)] + \mathrm{E}[X] - (\mathrm{E}[X])^2$.

12.9 Let Y_1 and Y_2 be independent Poisson random variables with parameter μ_1, respectively μ_2. Show that $Y = Y_1 + Y_2$ also has a Poisson distribution. Instead of using the addition rule in Section 11.1 as in Exercise 11.2, you can prove this without doing any computations by considering the number of points of a Poisson process (with intensity 1) in two disjoint intervals of length μ_1 and μ_2.

12.10 Let X be a random variable with a $Pois(\mu)$ distribution. Show the following. If $\mu < 1$, then the probabilities $\mathrm{P}(X = k)$ are strictly decreasing in k. If $\mu > 1$, then the probabilities $\mathrm{P}(X = k)$ are first increasing, then decreasing (cf. Figure 12.1). What happens if $\mu = 1$?

12.11 ⊞ Consider the one-dimensional Poisson process with intensity λ. Show that the number of points in $[0, t]$, *given* that the number of points in $[0, 2t]$ is equal to n, has a $Bin(n, \frac{1}{2})$ distribution.
Hint: write the event $\{N([0, s]) = k, N([0, 2s]) = n\}$ as the intersection of the (independent!) events $\{N([0, s]) = k\}$ and $\{N((s, 2s]) = n - k\}$.

12.12 We consider the one-dimensional Poisson process. Suppose for some $a > 0$ it is given that there are exactly two points in $[0, a]$, or in other words: $N_a = 2$. The goal of this exercise is to determine the joint distribution of X_1 and X_2, the locations of the two points, conditional on $N_a = 2$.

a. Prove that for $0 < s < t < a$

$$P(X_1 \le s, X_2 \le t, N_a = 2)$$
$$= P(X_2 \le t, N_a = 2) - P(X_1 > s, X_2 \le t, N_a = 2).$$

b. Deduce from **a** that

$$P(X_1 \le s, X_2 \le t, N_a = 2) = e^{-\lambda a} \left(\frac{\lambda^2 t^2}{2!} - \frac{\lambda^2 (t-s)^2}{2!} \right).$$

c. Deduce from **b** that for $0 < s < t < a$

$$P(X_1 \le s, X_2 \le t \mid N_a = 2) = \frac{t^2 - (t-s)^2}{a^2}.$$

12.13 Walking through a meadow we encounter two kinds of flowers, daisies and dandelions. As we walk in a straight line, we model the positions of the flowers we encounter with a one-dimensional Poisson process with intensity λ. It appears that about one in every four flowers is a daisy. Forgetting about the dandelions, what does the process of the *daisies* look like? This question will be answered with the following steps.

a. Let N_t be the total number of flowers, X_t the number of daisies, and Y_t be the number of dandelions we encounter during the first t minutes of our walk. Note that $X_t + Y_t = N_t$. Suppose that each flower is a daisy with probability $1/4$, independent of the other flowers. Argue that

$$P(X_t = n, Y_t = m \mid N_t = n + m) = \binom{n+m}{n} \left(\frac{1}{4} \right)^n \left(\frac{3}{4} \right)^m.$$

b. Show that

$$P(X_t = n, Y_t = m) = \frac{1}{n!} \frac{1}{m!} \left(\frac{1}{4} \right)^n \left(\frac{3}{4} \right)^m e^{-\lambda t} (\lambda t)^{n+m},$$

by conditioning on N_t and using **a**.

c. By writing $e^{-\lambda t} = e^{-(\lambda/4)t} e^{-(3\lambda/4)t}$ and summing over m, show that

$$P(X_t = n) = \frac{1}{n!} e^{-(\lambda/4)t} \left(\frac{\lambda t}{4} \right)^n.$$

Since it is clear that the numbers of daisies that we encounter in disjoint time intervals are independent, we may conclude from **c** that the process (X_t) is again a Poisson process, with intensity $\lambda/4$. One often says that the process (X_t) is obtained by *thinning* the process (N_t). In our example this corresponds to picking all the dandelions.

12.14 ⊡ In this exercise we look at a simple example of random variables X_n that have the property that their distributions converge to the distribution of a random variable X as $n \to \infty$, while it is *not* true that their expectations converge to the expectation of X. Let for $n = 1, 2, \ldots$ the random variables X_n be defined by

$$P(X_n = 0) = 1 - \frac{1}{n} \quad \text{and} \quad P(X_n = 7n) = \frac{1}{n}.$$

a. Let X be the random variable that is equal to 0 with probability 1. Show that for all a the probability mass functions $p_{X_n}(a)$ of the X_n converge to the probability mass function $p_X(a)$ of X as $n \to \infty$. Note that $\mathrm{E}[X]=0$.

b. Show that nonetheless $\mathrm{E}[X_n] = 7$ for all n.

13

The law of large numbers

For many experiments and observations concerning natural phenomena—such as measuring the speed of light—one finds that performing the procedure twice under (what seem) identical conditions results in two different outcomes. Uncontrollable factors cause "random" variation. In practice one tries to overcome this as follows: the experiment is repeated a number of times and the results are averaged in some way. In this chapter we will see why this works so well, using a model for repeated measurements. We view them as a sequence of independent random variables, each with the same unknown distribution. It is a probabilistic fact that from such a sequence—in principle—any feature of the distribution can be recovered. This is a consequence of the law of large numbers.

13.1 Averages vary less

Scientists and engineers involved in experimental work have known for centuries that more accurate answers are obtained when measurements or experiments are repeated a number of times and one averages the individual outcomes.[1] For example, if you read a description of A.A. Michelson's work done in 1879 to determine the speed of light, you would find that for each value he collected, repeated measurements at several levels were performed. In an article in *Statistical Science* describing his work ([18]), R.J. MacKay and R.W. Oldford state: "It is clear that Michelson appreciated the power of averaging to reduce variability in measurement." We shall see that we can understand this reduction using only what we have learned so far about probability in combination with a simple inequality called Chebyshev's inequality.

Throughout this chapter we consider a sequence of random variables X_1, X_2, X_3, ... You should think of X_i as the result of the ith repetition of a particular measurement or experiment. We confine ourselves to the situation where

---

[1] We leave the problem of systematic errors aside but will return to it in Chapter 19.

experimental conditions of subsequent experiments are identical, and the outcome of any one experiment does not influence the outcomes of others. Under those circumstances, the random variables of the sequence are independent, and all have the same distribution, and we therefore call $X_1, X_2, X_3, \ldots$ an *independent and identically distributed sequence*. We shall denote the distribution function of each random variable X_i by F, its expectation by μ, and the standard deviation by σ.

The average of the first n random variables in the sequence is

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

and using linearity of expectations we find:

$$\mathrm{E}\left[\bar{X}_n\right] = \frac{1}{n}\mathrm{E}[X_1 + X_2 + \cdots + X_n] = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu.$$

By the variance-of-the-sum rule, using the independence of $X_1, \ldots, X_n$,

$$\mathrm{Var}\left(\bar{X}_n\right) = \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}.$$

This establishes the following rule.

> EXPECTATION AND VARIANCE OF AN AVERAGE. If $\bar{X}_n$ is the average of n independent random variables with the same expectation μ and variance σ^2, then
>
> $$\mathrm{E}\left[\bar{X}_n\right] = \mu \quad \text{and} \quad \mathrm{Var}\left(\bar{X}_n\right) = \frac{\sigma^2}{n}.$$

The expectation of $\bar{X}_n$ is again μ, and its standard deviation is less than that of a single X_i by a factor $\sqrt{n}$; the "typical distance" from μ is $\sqrt{n}$ smaller. The latter property is what Michelson used to gain accuracy. To illustrate this, we analyze an example.

Suppose the random variables $X_1, X_2, \ldots$ are continuous with a $Gam(2,1)$ distribution, so with probability density:

$$f(x) = xe^{-x} \qquad \text{for } x \geq 0.$$

Recall from Section 11.2 that this means that each X_i is distributed as the sum of two independent $Exp(1)$ random variables. Hence, $S_n = X_1 + \cdots + X_n$ is distributed as the sum of $2n$ independent $Exp(1)$ random variables, which has a $Gam(2n, 1)$ distribution, with probability density

$$f_{S_n}(x) = \frac{x^{2n-1}e^{-x}}{(2n-1)!} \qquad \text{for } x \geq 0.$$

Because $\bar{X}_n = S_n/n$, we find by applying the change-of-units rule (page 106):

$$f_{\bar{X}_n}(x) = nf_{S_n}(nx) = \frac{n\,(nx)^{2n-1}\,\mathrm{e}^{-nx}}{(2n-1)!} \qquad \text{for } x \geq 0.$$

This is the probability density of the $Gam(2n, n)$ distribution.

So we have determined the distribution of $\bar{X}_n$ explicitly and we can investigate what happens as n increases, for example, by plotting probability densities. In the left-hand column of Figure 13.1 you see plots of $f_{\bar{X}_n}$ for $n = 1, 2, 4, 9$, 16, and 400 (note that for $n = 1$ this is just f itself). For comparison, we take as a second example a so-called *bimodal* density function: a density with two bumps, formally called *modes*. For the same values of n we determined the probability density function of $\bar{X}_n$ (unlike the previous example, we are not concerned with the computations, just with the results). The graphs of these densities are given side by side with the gamma densities in Figure 13.1.

The graphs clearly show that, as n increases, there is "contraction" of the probability mass near the expected value μ (for the gamma densities this is 2, for the bimodal densities 2.625).

QUICK EXERCISE 13.1 Compare the probabilities that $\bar{X}_n$ is within 0.5 of its expected value for $n = 1, 4, 16$, and 400. Do this for the gamma case only by estimating the probabilities from the graphs in the left-hand column of Figure 13.1.

13.2 Chebyshev's inequality

The contraction of probability mass near the expectation is a consequence of the fact that, for any probability distribution, most probability mass is within a few standard deviations from the expectation. To show this we will employ the following tool, which provides a bound for the probability that the random variable Y is outside the interval $(\mathrm{E}[Y] - a, \mathrm{E}[Y] + a)$.

> CHEBYSHEV'S INEQUALITY.   For an arbitrary random variable Y and any $a > 0$:
> $$\mathrm{P}(|Y - \mathrm{E}[Y]| \geq a) \leq \frac{1}{a^2}\mathrm{Var}(Y).$$

We shall derive this inequality for continuous Y (the discrete case is similar). Let f_Y be the probability density function of Y. Let μ denote $\mathrm{E}[Y]$. Then:

$$\mathrm{Var}(Y) = \int_{-\infty}^{\infty} (y - \mu)^2 f_Y(y)\,\mathrm{d}y \geq \int_{|y-\mu|\geq a} (y - \mu)^2 f_Y(y)\,\mathrm{d}y$$

$$\geq \int_{|y-\mu|\geq a} a^2 f_Y(y)\,\mathrm{d}y = a^2\mathrm{P}(|Y - \mu| \geq a).$$
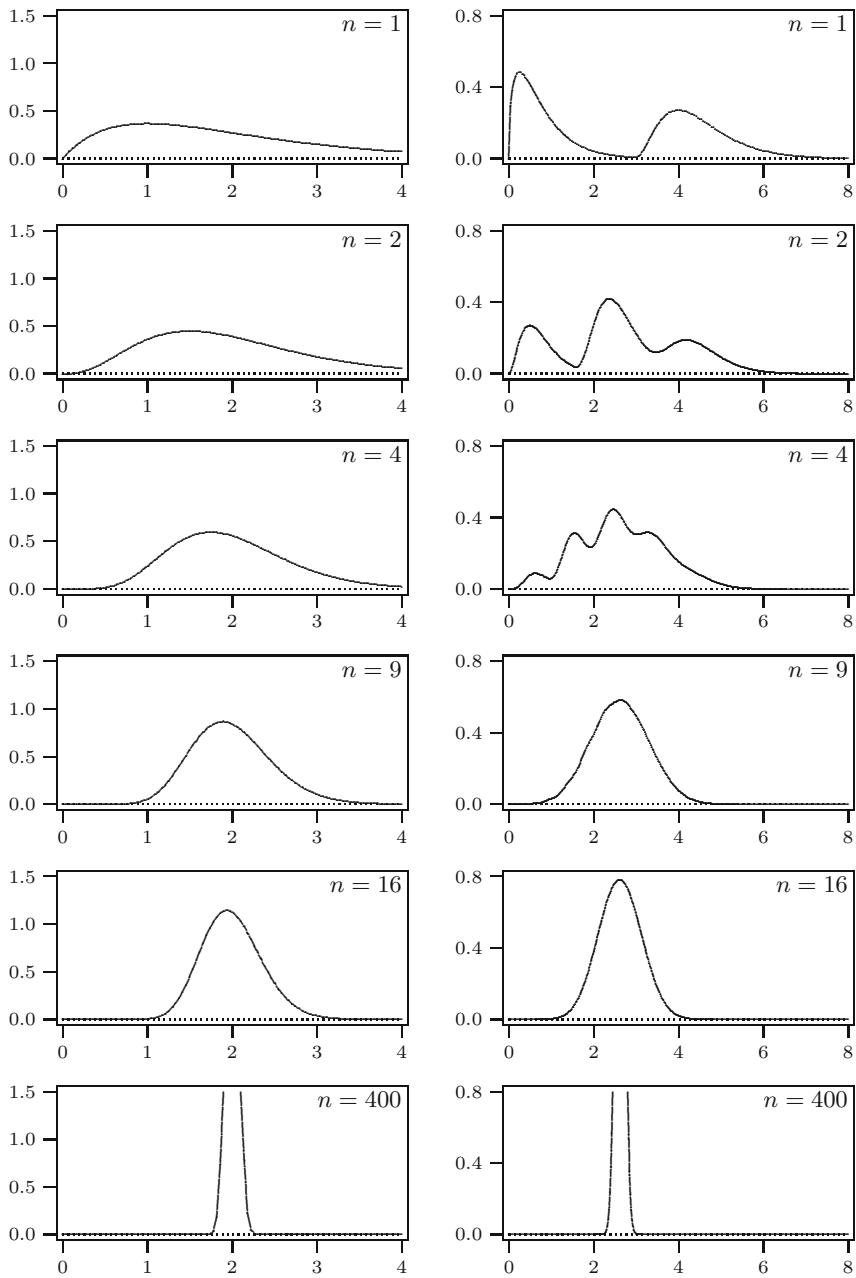
Fig. 13.1. Densities of averages. Left column: from a gamma density; right column: from a bimodal density.

Dividing both sides of the resulting inequality by a^2, we obtain Chebyshev's inequality.

Denote $\text{Var}(Y)$ by σ^2 and consider the probability that Y is within a few standard deviations from its expectation μ:

$$\text{P}(|Y - \mu| < k\sigma) = 1 - \text{P}(|Y - \mu| \geq k\sigma),$$

where k is a small integer. Setting $a = k\sigma$ in Chebyshev's inequality, we find

$$\text{P}(|Y - \mu| < k\sigma) \geq 1 - \frac{\text{Var}(Y)}{k^2\sigma^2} = 1 - \frac{1}{k^2}. \tag{13.1}$$

For $k = 2, 3, 4$ the right-hand side is 3/4, 8/9, and 15/16, respectively. This suggests that with Chebyshev's inequality we can make very strong statements. For most distributions, however, the actual value of $\text{P}(|Y - \mu| < k\sigma)$ is even *higher* than the lower bound (13.1). We summarize this as a somewhat loose rule.

> THE "μ ± A FEW σ" RULE. Most of the probability mass of a random variable is within a few standard deviations from its expectation.

QUICK EXERCISE 13.2 Calculate $\text{P}(|Y - \mu| < k\sigma)$ exactly for $k = 1, 2, 3, 4$ when Y has an $Exp(1)$ distribution and compare this with the bounds from Chebyshev's inequality.

13.3 The law of large numbers

We return to the independent and identically distributed sequence of random variables $X_1, X_2, \ldots$ with expectation μ and variance σ^2. We apply Chebyshev's inequality to the average $\bar{X}_n$, where we use $\text{E}[\bar{X}_n] = \mu$ and $\text{Var}(\bar{X}_n) = \sigma^2/n$, and where $\varepsilon > 0$:

$$\text{P}(|\bar{X}_n - \mu| > \varepsilon) = \text{P}(|\bar{X}_n - \text{E}[\bar{X}_n]| > \varepsilon) \leq \frac{1}{\varepsilon^2}\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n\varepsilon^2}.$$

The right-hand side vanishes as n goes to infinity, no matter how small ε is. This proves the following law.

> THE LAW OF LARGE NUMBERS. If $\bar{X}_n$ is the average of n independent random variables with expectation μ and variance σ^2, then for any $\varepsilon > 0$:
> $$\lim_{n\to\infty} \text{P}(|\bar{X}_n - \mu| > \varepsilon) = 0.$$

A connection with experimental work

Let us try to interpret the law of large numbers from an experimenter's perspective. Imagine you conduct a series of experiments. The experimental setup is complicated and your measurements vary quite a bit around the "true" value you are after. Suppose (unknown to you) your measurements have a gamma distribution, and its expectation is what you want to determine. You decide to do a certain number of measurements, say n, and to use their average as your estimate of the expectation.

We can simulate all this, and Figure 13.2 shows the results of a simulation, where we chose the same $Gam(2, 1)$ distribution, i.e., with expectation $\mu = 2$. We anticipated that you might want to do as many as 500 measurements, so we generated realizations for X_1, X_2, ..., X_{500}. For each n we computed the average of the first n values and plotted these averages against n in Figure 13.2.
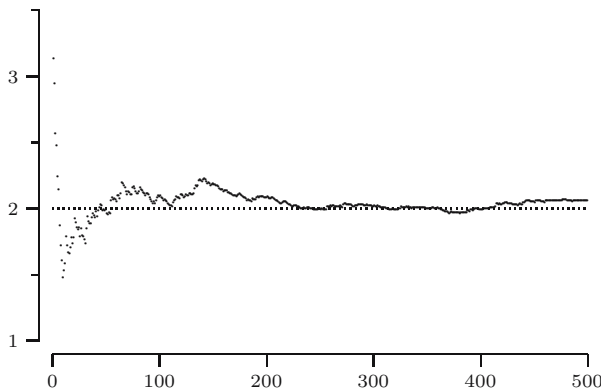
Fig. 13.2. Averages of realizations of a sequence of gamma distributed random variables.

If your decision is to do 200 repetitions, you would find (in this simulation) a value of about 2.09 (slightly too high, but you wouldn't know!), whereas with $n = 400$ you would be almost exactly correct with 1.99, and with $n = 500$ again a little farther away with 2.06. For another sequence of realizations, the details in the pattern that you see in Figure 13.2 would be different, but the general dampening of the oscillations would still be present. This follows from what we saw earlier, that as n is larger, the probability for the average to be within a certain distance of the expectation increases, in the limit even to 1. In practice it *may* happen that with a large number of repetitions your average is farther from the "true" value than with a smaller number of repetitions—if it is, then you had bad luck, because the odds are in your favor.

The averages may fail to converge

The law of large numbers is valid if the expectation of the distribution F is finite. This is not always the case. For example, the Cauchy and some Pareto distributions have heavy tails: their probability densities do go to 0 as x becomes large, but (too) slowly.[2] On the left in Figure 13.3 you see the result of a simulation with $Cau(2,1)$ random variables. As in the gamma case, the averages tend to go toward 2 (which is the point of symmetry of the $Cau(2,1)$ density), but once in a while a very large (positive or negative) realization of an X_i throws off the average.
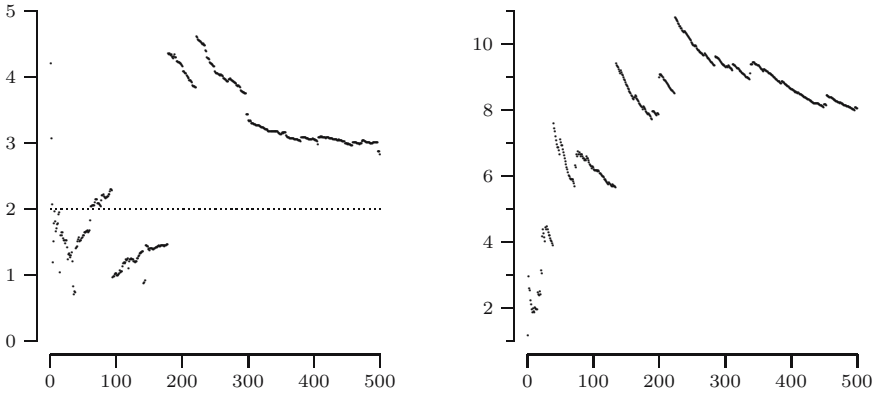
Fig. 13.3. Averages of realizations of a sequence of Cauchy (at left) and Pareto (at right) distributed random variables.

On the right in Figure 13.3 the result of a simulation with a $Par(0.99)$ distribution is shown. Its expectation is infinite. In the plot we see segments where the average "drifts downward," separated by upward jumps, which correspond to X_i with extremely large values. The effect of the jumps dominates: it can be shown that $\bar{X}_n$ grows beyond any level.

You might think that these patterns are phenomena that occur because of the short length of the simulation and that in longer simulations they would disappear after some value of n. However, the patterns as described will continue to occur and the results of a longer simulation, say to $n = 5000$, would not look any "better."

> **Remark 13.1 (There is a stronger law of large numbers).** Even though it is a strong statement, the law of large numbers in this paragraph is more accurately known as the *weak* law of large numbers. A stronger result holds, the *strong* law of large numbers, which says that:

---

[2] They represent two separate cases: the Cauchy expectation does not exist (see Remark 7.1) and the $Par(\alpha)$'s expectation is $+\infty$ if $\alpha \leq 1$ (see Section 7.2).

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

This is also expressed as "as n goes to infinity, $\bar{X}_n$ converges to μ with probability 1." It is not easy to see, but it is true that the strong law is actually stronger. The conditions for the law of large numbers, as stated in this section, could be relaxed. They suffice for both versions of the law. The conditions can be weakened to a point where the weak law still follows from them, but the strong law does not anymore; the strong law requires the stronger conditions.

13.4 Consequences of the law of large numbers

We continue with the sequence X_1, X_2, ... of independent random variables with distribution function F. In the previous section we saw how we could recover the (unknown) expectation μ from a realization of the sequence. We shall see that in fact we can recover any feature of the probability distribution. In order to avoid unnecessary indices, as in $\mathrm{E}[X_1]$ and $\mathrm{P}(X_1 \in C)$, we introduce an additional random variable X that also has F as its distribution function.

Recovering the probability of an event

Suppose that, rather than being interested in $\mu = \mathrm{E}[X]$, we want to know the probability of an event, for example,

$$p = \mathrm{P}(X \in C), \quad \text{where } C = (a, b] \text{ for some } a < b.$$

If you do not know this probability p, you would probably estimate it from how often the event $\{X_i \in C\}$ occurs in the sequence. You would use the relative frequency of $X_i \in C$ among X_1, ..., X_n: the number of times the set C was hit divided by n. Define for each i:

$$Y_i = \begin{cases} 1 & \text{if } X_i \in C, \\ 0 & \text{if } X_i \notin C. \end{cases}$$

The random variable Y_i indicates whether the corresponding X_i hits the set C; it is called an *indicator random variable*. In general, an indicator random variable for an event A is a random variable that is 1 when A occurs and 0 when A^c occurs. Using this terminology, Y_i is the indicator random variable of the event $X_i \in C$. Its expectation is given by

$$\mathrm{E}[Y_i] = 1 \cdot \mathrm{P}(X_i \in C) + 0 \cdot \mathrm{P}(X_i \notin C) = \mathrm{P}(X_i \in C) = \mathrm{P}(X \in C) = p.$$

Using the Y_i, the relative frequency is expressed as $(Y_1 + Y_2 + \cdots + Y_n)/n = \bar{Y}_n$. Note that the random variables $Y_1, Y_2, \ldots$ are independent; the X_i form an independent sequence, and Y_i is determined from X_i only (this is an application of the rule about propagation of independence; see page 126).

The law of large numbers, with p in the role of μ, can now be applied to $\bar{Y}_n$; it is the average of n independent random variables with expectation p and variance $p(1-p)$, so

$$\lim_{n\to\infty} \mathrm{P}\big(|\bar{Y}_n - p| > \varepsilon\big) = 0 \qquad (13.2)$$

for any $\varepsilon > 0$. By reasoning along the same lines as in the previous section, we see that from a long sequence of realizations we can get an accurate estimate of the probability p.

Recovering the probability density function

Consider the continuous case, where f is the probability density function corresponding with F, and now choose $C = (a - h, a + h]$, for some (small) positive h. By equation (13.2), for large n:

$$\bar{Y}_n \approx p = \mathrm{P}(X \in C) = \int_{a-h}^{a+h} f(x)\,\mathrm{d}x \approx 2hf(a). \qquad (13.3)$$

This relationship suggests to estimate the probability density in a as follows:

$$f(a) \approx \frac{\bar{Y}_n}{2h} = \frac{\text{the number of times } X_i \in C \text{ for } i \le n}{n \cdot \text{the length of } C}.$$

In Figure 13.4 we have done so for $h = 0.25$ and two values of a: 2 and 4. Rather than plotting the estimate in just one point, we use the same value for the whole interval $(a - h, a + h]$. This results in a vertical bar, whose area corresponds to $\bar{Y}_n$:

$$\text{height} \cdot \text{width} = \frac{\bar{Y}_n}{2h} \cdot 2h = \bar{Y}_n.$$

These estimates are based on the realizations of 500 independent $Gam(2,1)$ distributed random variables. In order to be able to see how well things came
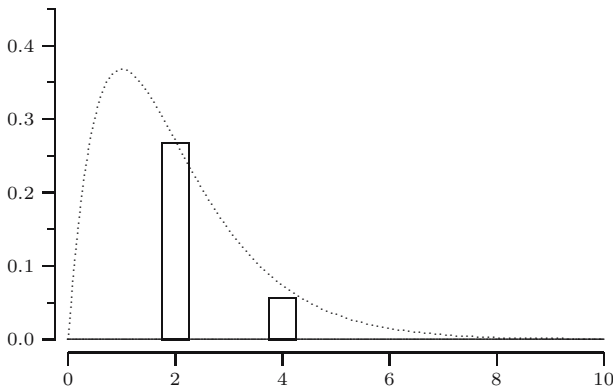
Fig. 13.4. Estimating the density at two points.

out, the $Gam(2, 1)$ density function is shown as well; near $a = 2$ the estimate is very accurate, but around $a = 4$ it is a little too low.

There really is no reason to derive estimated values around just a few points, as is done in Figure 13.4. We might as well cover the whole x-axis with a grid (with grid size $2h$) and do the computation for each point in the grid, thus covering the axis with a series of bars. The resulting bar graph is called a *histogram*. Figure 13.5 shows the result for two sets of realizations.
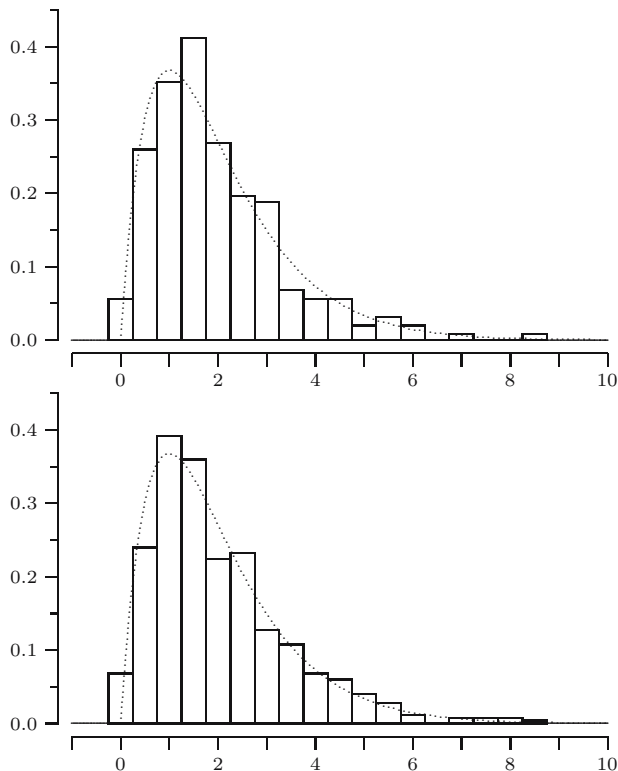
Fig. 13.5. Recovering the density function by way of histograms.

The top graph is constructed from the same realizations as Figure 13.4 and the bottom graph is constructed from a new set of realizations. Both graphs match the general shape of the density, with some bumps and valleys that are particular for the corresponding set of realizations. In Chapters 15 and 17 we shall return to histograms and treat them more elaborately.

QUICK EXERCISE 13.3 The height of the bar at $x = 2$ in the first histogram is 0.26. How many of the 500 realizations were between 1.75 and 2.25?

13.5 Solutions to the quick exercises

13.1 The answers you have found should be in the neighborhood of the following exact values:

n	1	4	16	400
$P\left(\lvert \bar{X}_n - \mu\rvert < 0.5\right)$	0.27	0.52	0.85	1.00

13.2 Because Y has an $Exp(1)$ distribution $\mu = 1$ and $\text{Var}(Y) = \sigma^2 = 1$; we find for $k \geq 1$:

$$P(\lvert Y - \mu\rvert < k\sigma) = P(\lvert Y - 1\rvert < k)$$
$$= P(1 - k < Y < k + 1) = P(Y < k + 1) = 1 - e^{-k-1}.$$

Using this formula and (13.1) we obtain the following numbers:

k	1	2	3	4
Lower bound from Chebyshev	0	0.750	0.889	0.938
$P(\lvert Y - 1\rvert < k)$	0.865	0.950	0.982	0.993

13.3 The value of $\bar{Y}_n$ for this bar equals its area $0.26 \cdot 0.5 = 0.13$. The bar represents 13% of the values, or $0.13 \cdot 500 = 65$ realizations.

13.6 Exercises

13.1 Verify the "$\mu \pm$ a few σ" rule as you did in Quick exercise 13.2 for the following distributions: $U(-1, 1)$, $U(-a, a)$, $N(0, 1)$, $N(\mu, \sigma^2)$, $Par(3)$, $Geo(1/2)$. Construct a table as in the answer to the quick exercise and enter a line for each distribution.

13.2 ⊞ An accountant wants to simplify his bookkeeping by rounding amounts to the nearest integer, for example, rounding € 99.53 and € 100.46 both to € 100. What is the cumulative effect of this if there are, say, 100 amounts? To study this we model the rounding errors by 100 independent $U(-0.5, 0.5)$ random variables $X_1, X_2, \ldots, X_{100}$.

  a. Compute the expectation and the variance of the X_i.

  b. Use Chebyshev's inequality to compute an upper bound for the probability $P(\lvert X_1 + X_2 + \cdots + X_{100}\rvert > 10)$ that the cumulative rounding error $X_1 + X_2 + \cdots + X_{100}$ exceeds € 10.

13.3 Consider the situation of the previous exercise. A manager wants to know what happens to the mean absolute error $\frac{1}{n}\sum_{i=1}^{n}|X_i|$ as n becomes large. What can you say about this, applying the law of large numbers?

13.4 ⊞ Of the voters in Florida, a proportion p will vote for candidate G, and a proportion $1-p$ will vote for candidate B. In an election poll a number of voters are asked for whom they will vote. Let X_i be the indicator random variable for the event "the ith person interviewed will vote for G." A model for the election poll is that the people to be interviewed are selected in such a way that the indicator random variables X_1, X_2,... are independent and have a $Ber(p)$ distribution.

a. Suppose we use $\bar{X}_n$ to predict p. According to Chebyshev's inequality, how large should n be (how many people should be interviewed) such that the probability that $\bar{X}_n$ is within 0.2 of the "true" p is at least 0.9?
Hint: solve this first for $p = 1/2$, and use that $p(1-p) \le 1/4$ for all $0 \le p \le 1$.

b. Answer the same question, but now $\bar{X}_n$ should be within 0.1 of p.

c. Answer the question from part **a**, but now the probability should be at least 0.95.

d. If $p > 1/2$ candidate G wins; if $\bar{X}_n > 1/2$ you predict that G will win. Find an n (as small as you can) such that the probability that you predict correctly is at least 0.9, if in fact $p = 0.6$.

13.5 You are trying to determine the melting point of a new material, of which you have a large number of samples. For each sample that you measure you find a value close to the actual melting point c but corrupted with a measurement error. We model this with random variables:

$$M_i = c + U_i$$

where M_i is the measured value in degree Kelvin, and U_i is the occurring random error. It is known that $E[U_i] = 0$ and $Var(U_i) = 3$, for each i, and that we may consider the random variables M_1, M_2, ... independent. According to Chebyshev's inequality, how many samples do you need to measure to be 90% sure that the average of the measurements is within half a degree of c?

13.6 ⊡ The casino La bella Fortuna is for sale and you think you might want to buy it, but you want to know how much money you are going to make. All the present owner can tell you is that the roulette game Red or Black is played about 1000 times a night, 365 days a year. Each time it is played you have probability 19/37 of winning the player's bet of €1 and probability 18/37 of having to pay the player €1.

Explain in detail why the law of large numbers can be used to determine the income of the casino, and determine how much it is.

13.7 Let X_1, X_2, ... be a sequence of independent and identically distributed random variables with distributions function F. Define F_n as follows: for any a

$$F_n(a) = \frac{\text{number of } X_i \text{ in } (-\infty, a]}{n}.$$

Consider a fixed and introduce the appropriate indicator random variables (as in Section 13.4). Compute their expectation and variance and show that the law of large numbers tells us that

$$\lim_{n\to\infty} \mathrm{P}(|F_n(a) - F(a)| > \varepsilon) = 0.$$

13.8 □ In Section 13.4 we described how the probability density function could be recovered from a sequence X_1, X_2, X_3, ... We consider the $Gam(2, 1)$ probability density discussed in the main text and a histogram bar around the point $a = 2$. Then $f(a) = f(2) = 2\mathrm{e}^{-2} = 0.27$ and the estimate for $f(2)$ is $\bar{Y}_n/2h$, where $\bar{Y}_n$ as in (13.3).

a. Express the standard deviation of $\bar{Y}_n/2h$ in terms of n and h.

b. Choose $h = 0.25$. How large should n be (according to Chebyshev's inequality) so that the estimate is within 20% of the "true value", with probability 80%?

13.9 ⊞ Let X_1, X_2, ... be an independent sequence of $U(-1, 1)$ random variables and let $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$. It is claimed that for some a and any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}(|T_n - a| > \varepsilon) = 0.$$

a. Explain how this could be true.

b. Determine a.

13.10 □ Let M_n be the maximum of n independent $U(0, 1)$ random variables.

a. Derive the exact expression for $\mathrm{P}(|M_n - 1| > \varepsilon)$.
   Hint: see Section 8.4.

b. Show that $\lim_{n\to\infty} \mathrm{P}(|M_n - 1| > \varepsilon) = 0$. Can this be derived from Chebyshev's inequality or the law of large numbers?

13.11 For some $t > 1$, let X be a random variable taking the values 0 and t, with probabilities

$$\mathrm{P}(X = 0) = 1 - \frac{1}{t} \quad \text{and} \quad \mathrm{P}(X = t) = \frac{1}{t}.$$

Then $\mathrm{E}[X] = 1$ and $\mathrm{Var}(X) = t - 1$. Consider the probability $\mathrm{P}(|X - 1| > a)$.

a. Verify the following: if $t = 10$ and $a = 8$ then $\mathrm{P}(|X - 1| > a) = 1/10$ and Chebyshev's inequality gives an upper bound for this probability of 9/64. The difference is $9/64 - 1/10 \approx 0.04$. We will say that for $t = 10$ the Chebyshev gap for X at $a = 8$ is 0.04.

b. Compute the Chebyshev gap for $t = 10$ at $a = 5$ and at $a = 10$.

c. Can you find a gap smaller than 0.01, smaller than 0.001, smaller than 0.0001?

d. Do you think one could improve Chebyshev's inequality, i.e., find an upper bound closer to the true probabilities?

13.12 (A more general law of large numbers). Let $X_1, X_2, \ldots$ be a sequence of independent random variables, with $\mathrm{E}[X_i] = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$, for $i = 1, 2, \ldots$. Suppose that $0 < \sigma_i^2 \le M$, for all i. Let a be an arbitrary positive number.

a. Apply Chebyshev's inequality to show that

$$\mathrm{P}\left(\left|\bar{X}_n - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right| > a\right) \le \frac{\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)}{n^2 a^2}.$$

b. Conclude from **a** that

$$\lim_{n\to\infty} \mathrm{P}\left(\left|\bar{X}_n - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right| > a\right) = 0.$$

Check that the law of large numbers is a special case of this result.

14

The central limit theorem

The central limit theorem is a refinement of the law of large numbers. For a large number of independent identically distributed random variables $X_1, \ldots, X_n$, with finite variance, the average $\bar{X}_n$ approximately has a normal distribution, no matter what the distribution of the X_i is. In the first section we discuss the proper normalization of $\bar{X}_n$ to obtain a normal distribution in the limit. In the second section we will use the central limit theorem to approximate probabilities of averages and sums of random variables.

14.1 Standardizing averages

In the previous chapter we saw that the law of large numbers guarantees the convergence to μ of the average $\bar{X}_n$ of n independent random variables $X_1, \ldots, X_n$, all having the same expectation μ and variance σ^2. This convergence was illustrated by Figure 13.1. Closer examination of this figure suggests another phenomenon: for the two distributions considered (i.e., the $Gam(2,1)$ distribution and a bimodal distribution), the probability density function of $\bar{X}_n$ seems to become symmetrical and bell shaped around the expected value μ as n becomes larger and larger. However, the bell collapses into a single spike at μ. Nevertheless, by a proper normalization it is possible to stabilize the bell shape, as we will see.

In order to let the distribution of $\bar{X}_n$ settle down it seems to be a good idea to stabilize the expectation and variance. Since $\mathrm{E}\left[\bar{X}_n\right] = \mu$ for all n, only the variance needs some special attention. In Figure 14.1 we depict the probability density function of the centered average $\bar{X}_n - \mu$ of $Gam(2,1)$ random variables, multiplied by three different powers of n. In the left column we display the density of $n^{\frac{1}{4}}(\bar{X}_n - \mu)$, in the middle column the density of $n^{\frac{1}{2}}(\bar{X}_n - \mu)$, and in the right column the density of $n(\bar{X}_n - \mu)$. These figures suggest that $\sqrt{n}$ is the right factor to stabilize the bell shape.
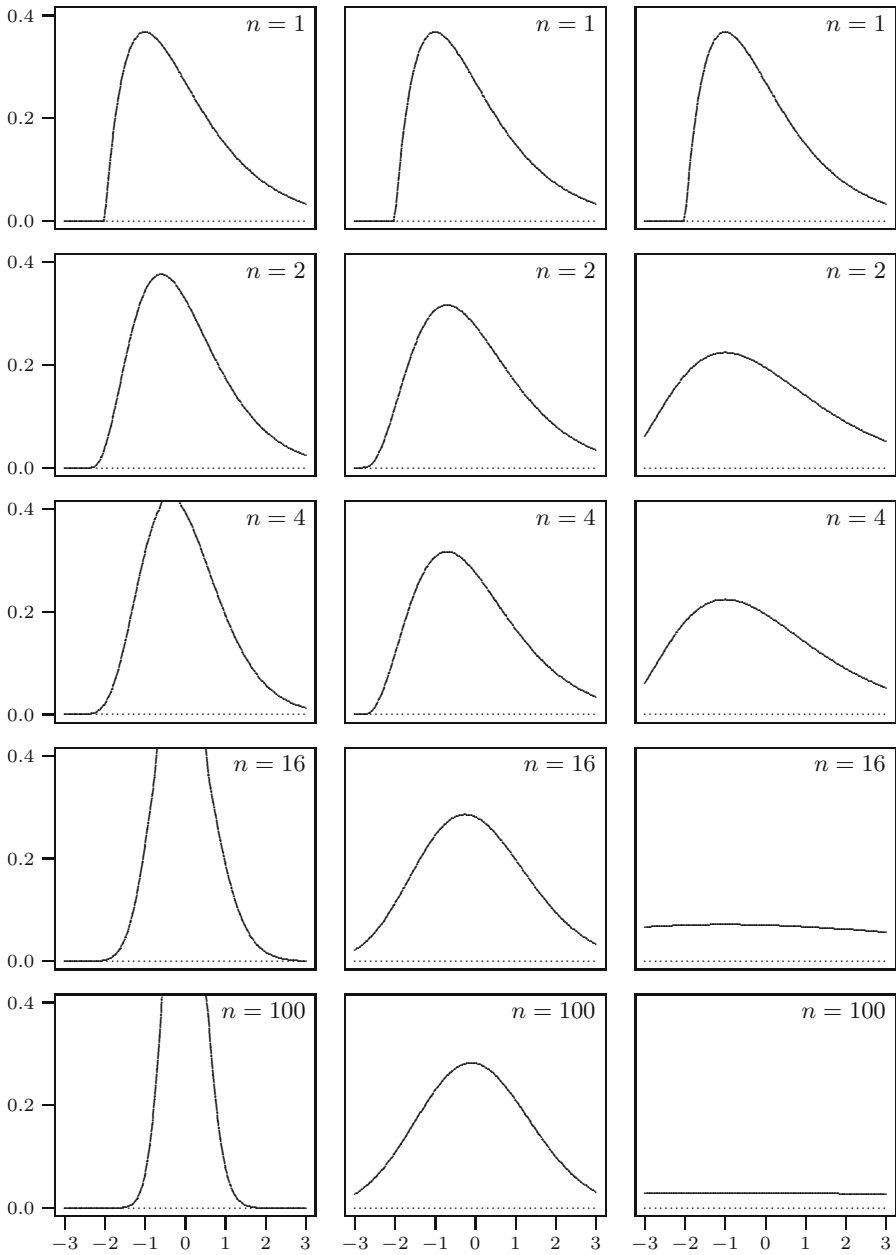
Fig. 14.1. Multiplying the difference $\bar{X}_n - \mu$ of n $Gam(2,1)$ random variables. Left column: $n^{\frac{1}{4}}(\bar{X}_n - \mu)$; middle column: $\sqrt{n}(\bar{X}_n - \mu)$; right column: $n(\bar{X}_n - \mu)$.

Indeed, according to the rule for the variance of an average (see page 182), we have $\text{Var}(\bar{X}_n) = \sigma^2/n$, and therefore for any number C:

$$\text{Var}\big(C(\bar{X}_n - \mu)\big) = \text{Var}\big(C\bar{X}_n\big) = C^2\text{Var}\big(\bar{X}_n\big) = C^2\frac{\sigma^2}{n}.$$

To stabilize the variance we therefore must choose $C = \sqrt{n}$. In fact, by choosing $C = \sqrt{n}/\sigma$, one *standardizes* the averages, i.e., the resulting random variable Z_n, defined by

$$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma}, \quad n = 1, 2, \ldots,$$

has expected value 0 and variance 1. What more can we say about the distribution of the random variables Z_n?

In case $X_1, X_2, \ldots$ are independent $N(\mu, \sigma^2)$ distributed random variables, we know from Section 11.2 and the rule on expectation and variance under change of units (see page 98), that Z_n has an $N(0, 1)$ distribution for all n. For the gamma and bimodal random variables from Section 13.1 we depicted the probability density function of Z_n in Figure 14.2. For both examples we see that the probability density functions of the Z_n seem to converge to the probability density function of the $N(0, 1)$ distribution, indicated by the dotted line. The following amazing result states that this behavior generally occurs no matter what distribution we start with.

> THE CENTRAL LIMIT THEOREM. Let $X_1, X_2, \ldots$ be any sequence of independent identically distributed random variables with finite positive variance. Let μ be the expected value and σ^2 the variance of each of the X_i. For $n \geq 1$, let Z_n be defined by
>
> $$Z_n = \sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma};$$
>
> then for any number a
>
> $$\lim_{n \to \infty} F_{Z_n}(a) = \Phi(a),$$
>
> where Φ is the distribution function of the $N(0, 1)$ distribution. In words: the distribution function of Z_n converges to the distribution function Φ of the standard normal distribution.

Note that

$$Z_n = \frac{\bar{X}_n - \text{E}\big[\bar{X}_n\big]}{\sqrt{\text{Var}\big(\bar{X}_n\big)}},$$

which is a more direct way to see that Z_n is the average $\bar{X}_n$ standardized.
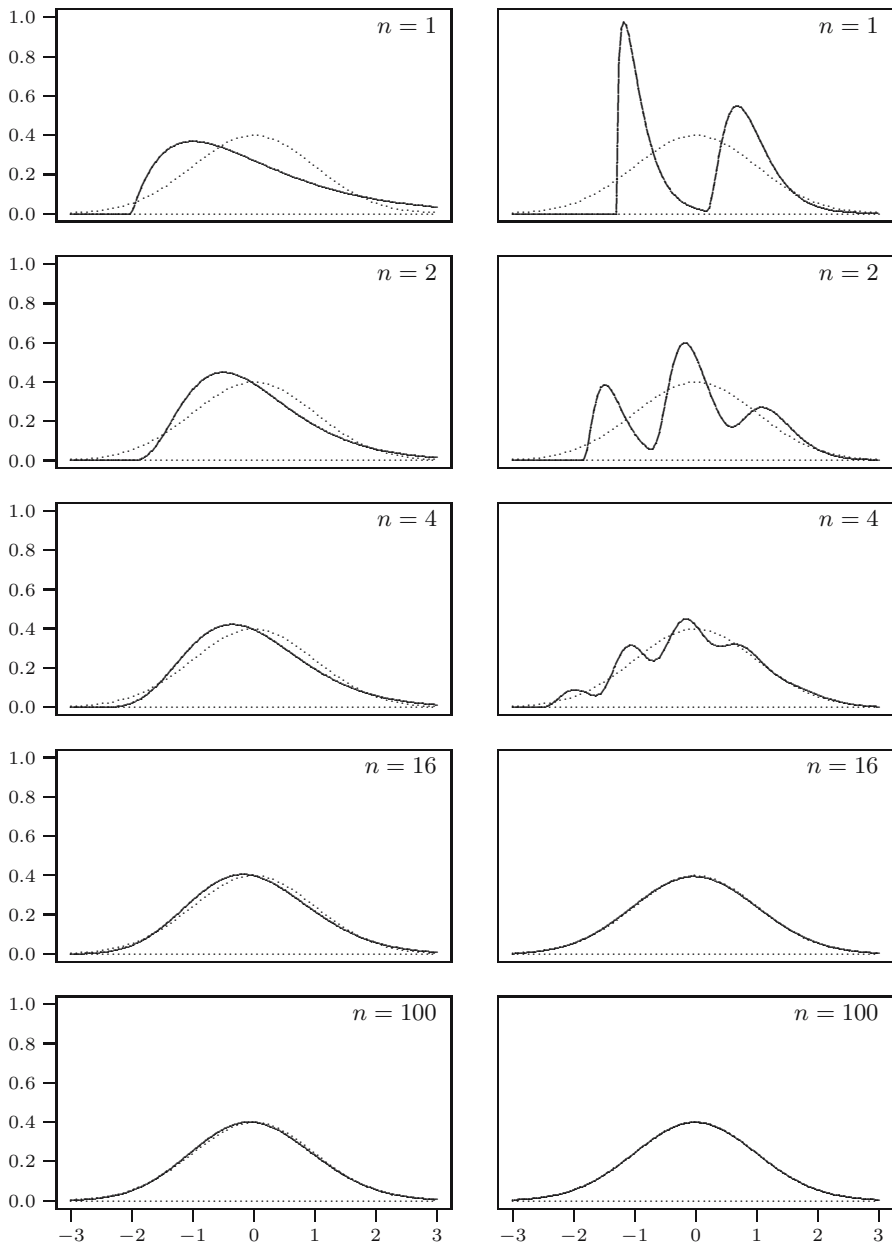
Fig. 14.2. Densities of standardized averages Z_n. Left column: from a gamma density; right column: from a bimodal density. Dotted line: $N(0,1)$ probability density.

One can also write Z_n as a standardized sum

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}. \tag{14.1}$$

In the next section we will see that this last representation of Z_n is very helpful when one wants to approximate probabilities of sums of independent identically distributed random variables.

Since

$$\bar{X}_n = \frac{\sigma}{\sqrt{n}}Z_n + \mu,$$

it follows that $\bar{X}_n$ approximately has an $N(\mu, \sigma^2/n)$ distribution; see the change-of-units rule for normal random variables on page 106. This explains the symmetrical bell shape of the probability densities in Figure 13.1.

> **Remark 14.1 (Some history).** Originally, the central limit theorem was proved in 1733 by De Moivre for independent $Ber(\frac{1}{2})$ distributed random variables. Lagrange extended De Moivre's result to $Ber(p)$ random variables and later formulated the central limit theorem as stated above. Around 1901 a first rigorous proof of this result was given by Lyapunov. Several versions of the central limit theorem exist with weaker conditions than those presented here. For example, for applications it is interesting that it is not necessary that all X_i have the same distribution; see Ross [26], Section 8.3, or Feller [8], Section 8.4, and Billingsley [3], Section 27.

14.2 Applications of the central limit theorem

The central limit theorem provides a tool to approximate the probability distribution of the average or the sum of independent identically distributed random variables. This plays an important role in applications, for instance, see Sections 23.4, 24.1, 26.2, and 27.2. Here we will illustrate the use of the central limit theorem to approximate probabilities of averages and sums of random variables in three examples. The first example deals with an average; the other two concern sums of random variables.

Did we have bad luck?

In the example in Section 13.3 averages of independent $Gam(2, 1)$ distributed random variables were simulated for $n = 1, \ldots, 500$. In Figure 13.2 the realization of $\bar{X}_n$ for $n = 400$ is 1.99, which is almost exactly equal to the expected value 2. For $n = 500$ the simulation was 2.06, a little bit farther away. Did we have bad luck, or is a value 2.06 or higher not unusual? To answer this question we want to compute $P(\bar{X}_n \geq 2.06)$. We will find an approximation of this probability using the central limit theorem.

Note that

$$P\left(\bar{X}_n \geq 2.06\right) = P\left(\bar{X}_n - \mu \geq 2.06 - \mu\right)$$
$$= P\left(\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \geq \sqrt{n}\frac{2.06 - \mu}{\sigma}\right)$$
$$= P\left(Z_n \geq \sqrt{n}\frac{2.06 - \mu}{\sigma}\right).$$

Since the X_i are $Gam(2,1)$ random variables, $\mu = E[X_i] = 2$ and $\sigma^2 = \text{Var}(X_i) = 2$. We find for $n = 500$ that

$$P\left(\bar{X}_{500} \geq 2.06\right) = P\left(Z_{500} \geq \sqrt{500}\frac{2.06 - 2}{\sqrt{2}}\right)$$
$$= P(Z_{500} \geq 0.95)$$
$$= 1 - P(Z_{500} < 0.95).$$

It now follows from the central limit theorem that

$$P\left(\bar{X}_{500} \geq 2.06\right) \approx 1 - \Phi(0.95) = 0.1711.$$

This is close to the exact answer 0.1710881, which was obtained using the probability density of $\bar{X}_n$ as given in Section 13.1.

Thus we see that there is about a 17% probability that the average $\bar{X}_{500}$ is at least 0.06 above 2. Since 17% is quite large, we conclude that the value 2.06 is not unusual. In other words, we did not have bad luck; $n = 500$ is simply not large enough to be that close. Would 2.06 be unusual if $n = 5000$?

QUICK EXERCISE 14.1 Show that $P\left(\bar{X}_{5000} \geq 2.06\right) \approx 0.0013$, using the central limit theorem.

Rounding amounts to the nearest integer

In Exercise 13.2 an accountant wanted to simplify his bookkeeping by rounding amounts to the nearest integer, and you were asked to use Chebyshev's inequality to compute an upper bound for the probability

$$p = P(|X_1 + X_2 + \cdots + X_{100}| > 10)$$

that the cumulative rounding error $X_1 + X_2 + \cdots + X_{100}$ exceeds € 10. This upper bound equals $1/12$. In order to know the exact value of p one has to determine the distribution of the sum $X_1 + \cdots + X_{100}$. This is difficult, but the central limit theorem is a handy tool to get an approximation of p. Clearly,

$$p = P(X_1 + \cdots + X_{100} < -10) + P(X_1 + \cdots + X_{100} > 10).$$

Standardizing as in (14.1), for the second probability we write, with $n = 100$

$$P(X_1 + \cdots + X_n > 10) = P(X_1 + \cdots + X_n - n\mu > 10 - n\mu)$$
$$= P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} > \frac{10 - n\mu}{\sigma\sqrt{n}}\right)$$
$$= P\left(Z_n > \frac{10 - n\mu}{\sigma\sqrt{n}}\right).$$

The X_i are $U(-0.5, 0.5)$ random variables, $\mu = E[X_i] = 0$, and $\sigma^2 = \text{Var}(X_i) = 1/12$, so that

$$P(X_1 + \cdots + X_{100} > 10) = P\left(Z_{100} > \frac{10 - 100 \cdot 0}{\sqrt{1/12}\sqrt{100}}\right) = P(Z_{100} > 3.46).$$

It follows from the central limit theorem that

$$P(Z_{100} > 3.46) \approx 1 - \Phi(3.46) = 0.0003.$$

Similarly,

$$P(X_1 + \cdots + X_{100} < -10) \approx \Phi(-3.46) = 0.0003.$$

Thus we find that $p = 0.0006$.

Normal approximation of the binomial distribution

In Section 4.3 we considered the (fictitious) situation that you attend, completely unprepared, a multiple-choice exam consisting of 10 questions. We saw that the probability you will pass equals

$$P(X \geq 6) = 0.0197,$$

where X—being the sum of 10 independent $Ber(\frac{1}{4})$ random variables—has a $Bin(10, \frac{1}{4})$ distribution. As we saw in Chapter 4 it is rather easy, but tedious, to calculate $P(X \geq 6)$. Although n is small, we investigate what the central limit theorem will yield as an approximation of $P(X \geq 6)$. Recall that a random variable with a $Bin(n, p)$ distribution can be written as the sum of n independent $Ber(p)$ distributed random variables $R_1, \ldots, R_n$. Substituting $n = 10$, $\mu = p = 1/4$, and $\sigma^2 = p(1 - p) = 3/16$, it follows from the central limit theorem that

$$P(X \geq 6) = P(R_1 + \cdots + R_n \geq 6)$$
$$= P\left(\frac{R_1 + \cdots + R_n - n\mu}{\sigma\sqrt{n}} \geq \frac{6 - n\mu}{\sigma\sqrt{n}}\right)$$
$$= P\left(Z_{10} \geq \frac{6 - 2\frac{1}{2}}{\sqrt{\frac{3}{16}}\sqrt{10}}\right)$$
$$\approx 1 - \Phi(2.56) = 0.0052.$$

The number 0.0052 is quite a poor approximation for the true value 0.0197. Note however, that we could also argue that

$$
\begin{aligned}
P(X \geq 6) &= P(X > 5) \\
&= P(R_1 + \cdots + R_n > 5) \\
&= P\left(Z_{10} \geq \frac{5 - 2\frac{1}{2}}{\sqrt{\frac{3}{16}}\sqrt{10}}\right) \\
&\approx 1 - \Phi(1.83) = 0.0336,
\end{aligned}
$$

which gives an approximation that is too large! A better approach lies somewhere in the middle, as the following quick exercise illustrates.

QUICK EXERCISE 14.2 Apply the central limit theorem to find 0.0143 as an approximation to $P\left(X \geq 5\frac{1}{2}\right)$. Since $P(X \geq 6) = P\left(X \geq 5\frac{1}{2}\right)$, this also provides an approximation of $P(X \geq 6)$.

How large should n be?

In view of the previous examples one might raise the question of how large n should be to have a good approximation when using the central limit theorem. In other words, how fast is the convergence to the normal distribution? This is a difficult question to answer in general. For instance, in the third example one might initially be tempted to think that the approximation was quite poor, but after taking the fact into account that we approximate a discrete distribution by a continuous one we obtain a considerable improvement of the approximation, as was illustrated in Quick exercise 14.2. For another example, see Figure 14.2. Here we see that the convergence is slightly faster for the bimodal distribution than for the $Gam(2,1)$ distribution, which is due to the fact that the $Gam(2,1)$ is rather asymmetric.

In general the approximation might be poor when n is small, when the distribution of the X_i is asymmetric, bimodal, or discrete, or when the value a in

$$
P(\bar{X}_n > a)
$$

is far from the center of the distribution of the X_i.

14.3 Solutions to the quick exercises

14.1 In the same way we approximated $P(\bar{X}_n \geq 2.06)$ using the central limit theorem, we have that

$$
P(\bar{X}_n \geq 2.06) = P\left(Z_n \geq \sqrt{n}\frac{2.06 - \mu}{\sigma}\right).
$$

With $\mu = 2$ and $\sigma = \sqrt{2}$, we find for $n = 5000$ that

$$P\big(\bar{X}_{5000} \geq 2.06\big) = P(Z_{5000} \geq 3)\,,$$

which is approximately equal to $1 - \Phi(3) = 0.0013$, thanks to the central limit theorem. Because we think that 0.13% is a small probability, to find 2.06 as a value for $\bar{X}_{5000}$ would mean that you really had bad luck!

14.2 Similar to the computation $P(X \geq 6)$, we have

$$P\left(X \geq 5\frac{1}{2}\right) = P\left(R_1 + \cdots + R_{10} \geq 5\frac{1}{2}\right)$$

$$= P\left(Z_{10} \geq \frac{5\frac{1}{2} - 2\frac{1}{2}}{\sqrt{\frac{3}{16}}\sqrt{10}}\right)$$

$$\approx 1 - \Phi(2.19) = 0.0143.$$

We have seen that using the central limit theorem to approximate $P(X \geq 6)$ gives an underestimate of this probability, while using the central limit theorem to $P(X > 5)$ gives an overestimation. Since $5\frac{1}{2}$ is "in the middle," the approximation will be better.

14.4 Exercises

14.1 Let $X_1, X_2, \ldots, X_{144}$ be independent identically distributed random variables, each with expected value $\mu = \mathrm{E}[X_i] = 2$, and variance $\sigma^2 = \mathrm{Var}(X_i) = 4$. Approximate $P(X_1 + X_2 + \cdots + X_{144} > 144)$, using the central limit theorem.

14.2 ☐ Let $X_1, X_2, \ldots, X_{625}$ be independent identically distributed random variables, with probability density function f given by

$$f(x) = \begin{cases} 3(1-x)^2 & \text{for } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Use the central limit theorem to approximate $P(X_1 + X_2 + \cdots + X_{625} < 170)$.

14.3 ⊞ In Exercise 13.4 **a** you were asked to use Chebyshev's inequality to determine how large n should be (how many people should be interviewed) so that the probability that $\bar{X}_n$ is within 0.2 of the "true" p is at least 0.9. Here p is the proportion of the voters in Florida who will vote for G (and $1 - p$ is the proportion of the voters who will vote for B). How large should n at least be according to the central limit theorem?

14.4 ⊡ In the single-server queue model from Section 6.4, T_i is the time between the arrival of the $(i-1)$th and ith customers. Furthermore, one of the model assumptions is that the T_i are independent, $Exp(0.5)$ distributed random variables. In Section 11.2 we saw that the probability $P(T_1 + \cdots + T_{30} \le 60)$ of the 30th customer arriving within an hour at the well is equal to 0.542. Find the normal approximation of this probability.

14.5 ⊞ Let X be a $Bin(n,p)$ distributed random variable. Show that the random variable
$$\frac{X - np}{\sqrt{np(1-p)}}$$
has a distribution that is approximately standard normal.

14.6 ⊡ Again, as in the previous exercise, let X be a $Bin(n,p)$ distributed random variable.

a. An exact computation yields that $P(X \le 25) = 0.55347$, when $n = 100$ and $p = 1/4$. Use the central limit theorem to give an approximation of $P(X \le 25)$ and $P(X < 26)$.

b. When $n = 100$ and $p = 1/4$, then $P(X \le 2) = 1.87 \cdot 10^{-10}$. Use the central limit theorem to give an approximation of this probability.

14.7 Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each with expected value μ and finite positive variance σ^2. Use Chebyshev's inequality to show that for any $a > 0$ one has
$$P\left(\left| n^{\frac{1}{4}} \frac{\bar{X}_n - \mu}{\sigma} \right| \ge a \right) \le \frac{1}{a^2 \sqrt{n}}.$$

Use this fact to explain the occurrence of a single spike in the left column of Figure 14.1.

14.8 Let $X_1, X_2, \ldots$ be a sequence of independent $N(0,1)$ distributed random variables. For $n = 1, 2, \ldots$, let Y_n be the random variable, defined by
$$Y_n = X_1^2 + \cdots + X_n^2.$$

a. Show that $E\left[X_i^2\right] = 1$.

b. One can show—using integration by parts—that $E\left[X_i^4\right] = 3$. Deduce from this that $Var(X_i^2) = 2$.

c. Use the central limit theorem to approximate $P(Y_{100} > 110)$.

14.9 ⊞ A factory produces links for heavy metal chains. The research lab of the factory models the length (in cm) of a link by the random variable X, with expected value $E[X] = 5$ and variance $Var(X) = 0.04$. The length of a link is defined in such a way that the length of a chain is equal to the sum of

the lengths of its links. The factory sells chains of 50 meters; to be on the safe side 1002 links are used for such chains. The factory guarantees that the chain is not shorter than 50 meters. If by chance a chain is too short, the customer is reimbursed, and a new chain is given for free.

a. Give an estimate of the probability that for a chain of at least 50 meters more than 1002 links are needed. For what percentage of the chains does the factory have to reimburse clients and provide free chains?

b. The sales department of the factory notices that it has to hand out a lot of free chains and asks the research lab what is wrong. After further investigations the research lab reports to the sales department that the expectation value 5 is incorrect, and that the correct value is 4.99 (cm). Do you think that it was necessary to report such a minor change of this value?

14.10 Chebyshev's inequality was used in Exercise 13.5 to determine how many times n one needs to measure a sample to be 90% sure that the average of the measurements is within half a degree of the actual melting point c of a new material.

a. Use the normal approximation to find a less conservative value for n.

b. Only in case the random errors U_i in the measurements have a normal distribution the value of n from **a** is "exact," in all other cases an approximation. Explain this.

15

Exploratory data analysis: graphical summaries

In the previous chapters we focused on probability models to describe random phenomena. Confronted with a new phenomenon, we want to learn about the randomness that is associated with it. It is common to conduct an experiment for this purpose and record observations concerning the phenomenon. The set of observations is called a *dataset*. By exploring the dataset we can gain insight into what probability model suits the phenomenon.

Frequently you will have to deal with a dataset that contains so many elements that it is necessary to condense the data for easy visual comprehension of general characteristics. In this chapter we present several graphical methods to do so. To graphically represent univariate datasets, consisting of repeated measurements of one particular quantity, we discuss the classical *histogram*, the more recently introduced *kernel density estimates* and the *empirical distribution function*. To represent a bivariate dataset, which consists of repeated measurements of two quantities, we use the *scatterplot*.

15.1 Example: the Old Faithful data

The Old Faithful geyser at Yellowstone National Park, Wyoming, USA, was observed from August 1st to August 15th, 1985. During that time, data were collected on the duration of eruptions. There were 272 eruptions observed, of which the recorded durations are listed in Table 15.1. The data are given in seconds.

The variety in the lengths of the eruptions indicates that randomness is involved. By exploring the dataset we might learn about this randomness. For instance: we like to know which durations are more likely to occur than others; is there something like "the typical duration of an eruption"; do the durations vary symmetrically around the center of the dataset; and so on. In order to retrieve this type of information, just listing the observed durations does not help us very much. Somehow we must summarize the observed data. We could

Table 15.1. Duration in seconds of 272 eruptions of the Old Faithful geyser.

216	108	200	137	272	173	282	216	117	261
110	235	252	105	282	130	105	288	96	255
108	105	207	184	272	216	118	245	231	266
258	268	202	242	230	121	112	290	110	287
261	113	274	105	272	199	230	126	278	120
288	283	110	290	104	293	223	100	274	259
134	270	105	288	109	264	250	282	124	282
242	118	270	240	119	304	121	274	233	216
248	260	246	158	244	296	237	271	130	240
132	260	112	289	110	258	280	225	112	294
149	262	126	270	243	112	282	107	291	221
284	138	294	265	102	278	139	276	109	265
157	244	255	118	276	226	115	270	136	279
112	250	168	260	110	263	113	296	122	224
254	134	272	289	260	119	278	121	306	108
302	240	144	276	214	240	270	245	108	238
132	249	120	230	210	275	142	300	116	277
115	125	275	200	250	260	270	145	240	250
113	275	255	226	122	266	245	110	265	131
288	110	288	246	238	254	210	262	135	280
126	261	248	112	276	107	262	231	116	270
143	282	112	230	205	254	144	288	120	249
112	256	105	269	240	247	245	256	235	273
245	145	251	133	267	113	111	257	237	140
249	141	296	174	275	230	125	262	128	261
132	267	214	270	249	229	235	267	120	257
286	272	111	255	119	135	285	247	129	265
109	268								

start by computing the mean of the data, which is 209.3 for the Old Faithful data. However, this is a poor summary of the dataset, because there is a lot more information in the observed durations. How do we get hold of this?

Just staring at the dataset for a while tells us very little. To see something, we have to rearrange the data somehow. The first thing we could do is order the data. The result is shown in Table 15.2. Putting the elements in order already provides more information. For instance, it is now immediately clear that all elements lie between 96 and 306.

QUICK EXERCISE 15.1 Which two elements of the Old Faithful dataset split the dataset in three groups of equal size?

A closer look at the ordered data shows that the two middle elements (the 136th and 137th elements in ascending order) are equal to 240, which is much closer to the maximum value 306 than to the minimum value 96. This seems to

Table 15.2. Ordered durations of eruptions of the Old Faithful geyser.

96	100	102	104	105	105	105	105	105	105
107	107	108	108	108	108	109	109	109	110
110	110	110	110	110	110	111	111	112	112
112	112	112	112	112	112	113	113	113	113
115	115	116	116	117	118	118	118	119	119
119	120	120	120	120	121	121	121	122	122
124	125	125	126	126	126	128	129	130	130
131	132	132	132	133	134	134	135	135	136
137	138	139	140	141	142	143	144	144	145
145	149	157	158	168	173	174	184	199	200
200	202	205	207	210	210	214	214	216	216
216	216	221	223	224	225	226	226	229	230
230	230	230	230	231	231	233	235	235	235
237	237	238	238	240	240	240	240	240	240
242	242	243	244	244	245	245	245	245	245
246	246	247	247	248	248	249	249	249	249
250	250	250	250	251	252	254	254	254	255
255	255	255	256	256	257	257	258	258	259
260	260	260	260	260	261	261	261	261	262
262	262	262	263	264	265	265	265	265	266
266	267	267	267	268	268	269	270	270	270
270	270	270	270	270	271	272	272	272	272
272	273	274	274	274	275	275	275	275	276
276	276	276	277	278	278	278	279	280	280
282	282	282	282	282	282	283	284	285	286
287	288	288	288	288	288	288	289	289	290
290	291	293	294	294	296	296	296	300	302
304	306								

indicate that the dataset is somewhat asymmetric, but even from the ordered dataset we cannot get a clear picture of this asymmetry. Also, geologists believe that there are two different kinds of eruptions that play a role. Hence one would expect two separate values around which the elements of the dataset would accumulate, corresponding to the typical durations of the two types of eruptions. Again it is not clear, not even from the ordered dataset, what these two typical values are. It would be better to have a plot of the dataset that reflects symmetry or asymmetry of the data and from which we can easily see where the elements accumulate. In the following sections we will discuss two such methods.

15.2 Histograms

The classical method to graphically represent data is the histogram, which probably dates from the mortality studies of John Graunt in 1662 (see West-

ergaard [39], p.22). The term *histogram* appears to have been used first by Karl Pearson ([22]). Figure 15.1 displays a histogram of the Old Faithful data. The picture immediately reveals the asymmetry of the dataset and the fact that the elements accumulate somewhere near 120 and 270, which was not clear from Tables 15.1 and 15.2.
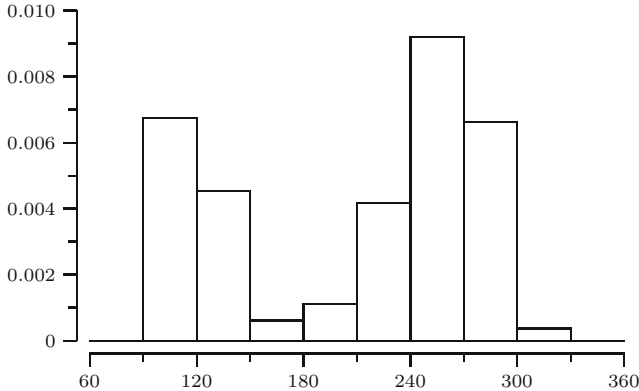
Fig. 15.1. Histogram of the Old Faithful data.

The construction of the histogram is as follows. Let us denote a generic (univariate) dataset of size n by

$$x_1, x_2, \ldots, x_n$$

and suppose we want to construct a histogram. We use the version of the histogram that is scaled in such a way that the total area under the curve is equal to one.[1]

First we divide the range of the data into intervals. These intervals are called *bins* and are denoted by

$$B_1, B_2, \ldots, B_m.$$

The length of an interval B_i is denoted by $|B_i|$ and is called the *bin width*. The bins do not necessarily have the same width. In Figure 15.1 we have eight bins of equal bin width. We want the area under the histogram on each bin B_i to reflect the number of elements in B_i. Since the total area 1 under the histogram then corresponds to the total number of elements n in the dataset, the area under the histogram on a bin B_i is equal to the proportion of elements in B_i:

$$\frac{\text{the number of } x_j \text{ in } B_i}{n}.$$

---

[1] The reason to scale the histogram so that the total area under the curve is equal to one is that if we view the data as being generated from some unknown probability density f (see Chapter 17), such a histogram can be used as a crude estimate of f.

The *height* of the histogram on bin B_i must then be equal to

$$\frac{\text{the number of } x_j \text{ in } B_i}{n|B_i|}.$$

QUICK EXERCISE 15.2 Use Table 15.2 to count how many elements fall into each of the bins $(90, 120], (120, 150], \ldots, (300, 330]$ in Figure 15.1 and compute the height on each bin.

Choice of the bin width

Consider a histogram with bins of equal width. In that case the bins are of the form

$$B_i = (r + (i-1)b, r + ib] \quad \text{for } i = 1, 2, \ldots, m,$$

where r is some reference point smaller than the minimum of the dataset, and b denotes the bin width. In Figure 15.2, three histograms of the Old Faithful data of Table 15.2 are displayed with bin widths equal to 2, 30, and 90, respectively. Clearly, the choice of the bin width b, or the corresponding choice of the number of bins m, will determine what the resulting histogram will look like. Choosing the bin width too small will result in a chaotic figure with many isolated peaks. Choosing the bin width too large will result in a figure without much detail, at the risk of losing information about general characteristics. In Figure 15.2, bin width $b = 2$ is somewhat too small. Bin width $b = 90$ is clearly too large and produces a histogram that no longer captures the fact that the data show two separate modes near 120 and 270.

How does one go about choosing the bin width? In practice, this might boil down to picking the bin width by trial and error, continuing until the figure looks reasonable. Mathematical research, however, has provided some guidelines for a data-based choice for b or m. Formulas that may effectively be used are $m = 1 + 3.3 \log_{10}(n)$ (see [34]) or $b = 3.49 \, sn^{-1/3}$ (see [29]; see also Remark 15.1), where s is the sample standard deviation (see Section 16.2 for the definition of the sample standard deviation).
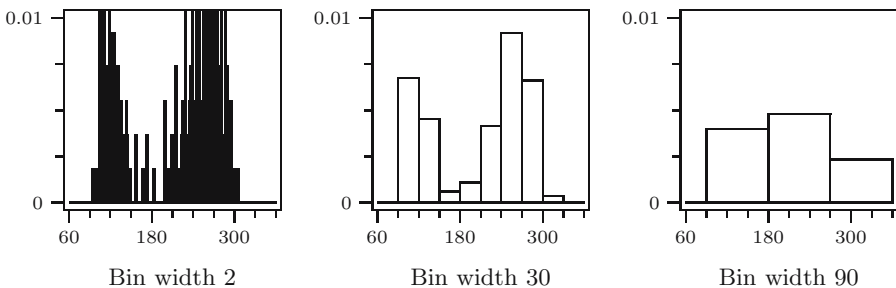
Fig. 15.2. Histograms of the Old Faithful data with different bin widths.

Remark 15.1 (Normal reference method for histograms). Let $H_n(x)$ denote the height of the histogram at x and suppose that we view our dataset as being generated from a probability distribution with density f. We would like to find the bin width that minimizes the difference between H_n and f, measured by the so-called mean integrated squared error (MISE)

$$\mathrm{E}\left[\int_{-\infty}^{\infty}(H_n(x) - f(x))^2\,\mathrm{d}x\right].$$

Under suitable smoothness conditions on f, the value of b that minimizes the MISE as n goes to infinity is given by

$$b = C(f)n^{-1/3} \quad \text{where } C(f) = 6^{1/3}\left(\int_{-\infty}^{\infty} f'(x)^2\,\mathrm{d}x\right)^{-1/3}$$

(see for instance [29] or [12]). A simple data-based choice for b is obtained by estimating the constant $C(f)$. The normal reference method takes f to be the density of an $N(\mu, \sigma^2)$ distribution, in which case $C(f) = (24\sqrt{\pi})^{1/3}\sigma$. Estimating σ by the sample standard deviation s (see Chapter 16 for a definition of s) would result in bin width

$$b = (24\sqrt{\pi})^{1/3}sn^{-1/3}.$$

For the Old Faithful data this would give $b = 36.89$.

QUICK EXERCISE 15.3 If we construct a histogram for the Old Faithful data with equal bin width $b = 3.49\,sn^{-1/3}$, how may bins will we need to cover the data if $s = 68.48$?

The main advantage of the histogram is that it is simple. Its disadvantage is the discrete character of the plot. In Figure 15.1 it is still somewhat unclear which two values correspond to the typical durations of the two types of eruptions. Another well-known artifact is that changing the bin width slightly or keeping the bin width fixed and shifting the bins slightly may result in a figure of a different nature. A method that produces a smoother figure and is less sensitive to these kinds of changes will be discussed in the next section.

15.3 Kernel density estimates

We can graphically represent data in a more variegated plot by a so-called kernel density estimate. The basic ideas of kernel density estimation first appeared in the early 1950s. Rosenblatt [25] and Parzen [21] provided the stimulus for further research on this topic. Although the method was introduced in the middle of the last century, until recently it remained unpopular as a tool for practitioners because of its computationally intensive nature.

Figure 15.3 displays a kernel density estimate of the Old Faithful data. Again the picture immediately reveals the asymmetry of the dataset, but it is much
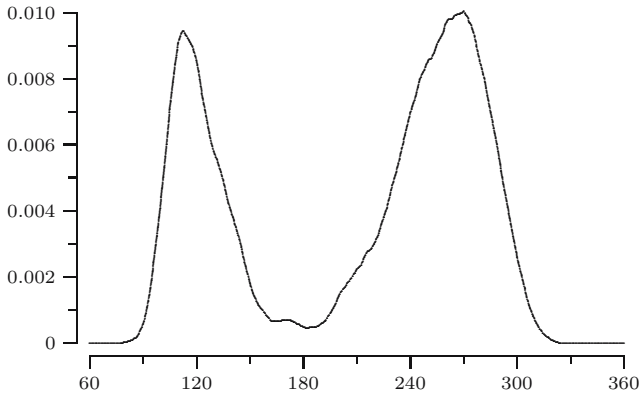
Fig. 15.3. Kernel density estimate of the Old Faithful data.

smoother than the histogram in Figure 15.1. Note that it is now easier to detect the two typical values around which the elements accumulate.

The idea behind the construction of the plot is to "put a pile of sand" around each element of the dataset. At places where the elements accumulate, the sand will pile up. The actual plot is constructed by choosing a *kernel K* and a *bandwidth h*. The kernel K reflects the shape of the piles of sand, whereas the bandwidth is a tuning parameter that determines how wide the piles of sand will be. Formally, a kernel K is a function $K : \mathbb{R} \to \mathbb{R}$. Figure 15.4 displays several well-known kernels. A kernel K typically satisfies the following conditions:

(K1) K is a probability density, i.e., $K(u) \geq 0$ and $\int_{-\infty}^{\infty} K(u)\,du = 1$;

(K2) K is symmetric around zero, i.e., $K(u) = K(-u)$;

(K3) $K(u) = 0$ for $|u| > 1$.

Examples are the *Epanechnikov kernel*:

$$K(u) = \frac{3}{4}\left(1 - u^2\right) \quad \text{for } -1 \leq u \leq 1$$

and $K(u) = 0$ elsewhere, and the *triweight kernel*

$$K(u) = \frac{35}{32}\left(1 - x^2\right)^3 \quad \text{for } -1 \leq u \leq 1$$

and $K(u) = 0$ elsewhere. Sometimes one uses kernels that do not satisfy condition (K3), for example, the *normal kernel*

$$K(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2} \quad \text{for } -\infty < u < \infty.$$

Let us denote a kernel density estimate by $f_{n,h}$, and suppose that we want to construct $f_{n,h}$ for a dataset $x_1, x_2, \ldots, x_n$. In Figure 15.5 the construction is
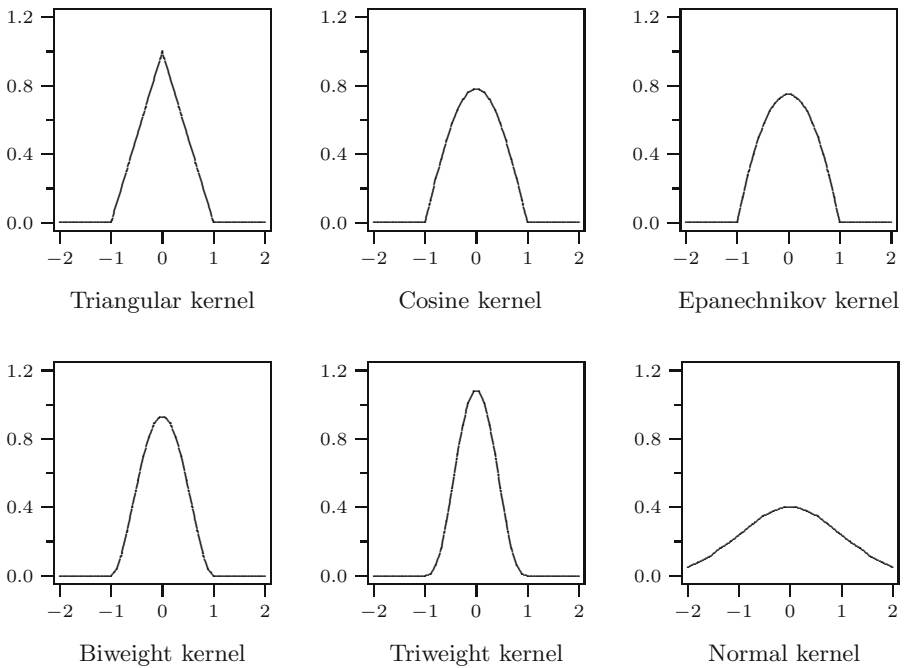
Fig. 15.4. Examples of well-known kernels K.

illustrated for a dataset containing five elements, where we use the Epanechnikov kernel and bandwidth $h = 0.5$. First we scale the kernel K (solid line) into the function

$$t \mapsto \frac{1}{h}K\left(\frac{t}{h}\right).$$

The scaled kernel (dotted line) is of the same type as the original kernel, with area 1 under the curve but is positive on the interval $[-h, h]$ instead of $[-1, 1]$ and higher (lower) when h is smaller (larger) than 1. Next, we put a scaled kernel around each element x_i in the dataset. This results in functions of the type

$$t \mapsto \frac{1}{h}K\left(\frac{t - x_i}{h}\right).$$

These shifted kernels (dotted lines) have the same shape as the transformed kernel, all with area 1 under the curve, but they are now symmetric around x_i and positive on the interval $[x_i - h, x_i + h]$. We see that the graphs of the shifted kernels will overlap whenever x_i and x_j are close to each other, so that things will pile up more at places where more elements accumulate. The kernel density estimate $f_{n,h}$ is constructed by summing the scaled kernels and dividing them by n, in order to obtain area 1 under the curve:

| Kernel and scaled kernel | Shifted kernel | Kernel density estimate |
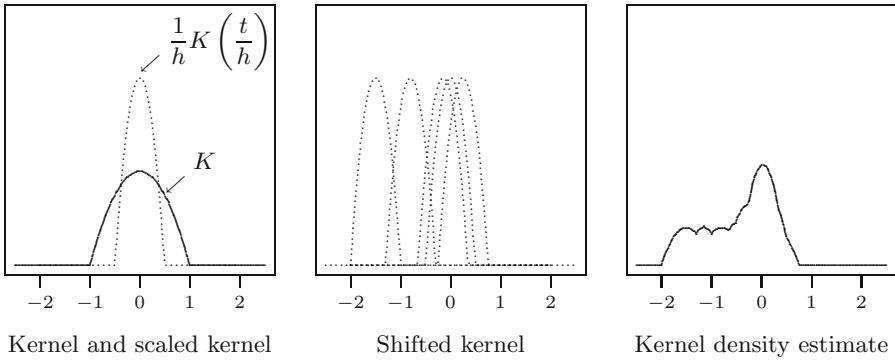
Fig. 15.5. Construction of a kernel density estimate $f_{n,h}$.

$$f_{n,h}(t) = \frac{1}{n} \left\{ \frac{1}{h} K \left(\frac{t - x_1}{h} \right) + \frac{1}{h} K \left(\frac{t - x_2}{h} \right) + \cdots + \frac{1}{h} K \left(\frac{t - x_n}{h} \right) \right\}$$

or briefly,

$$f_{n,h}(t) = \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{t - x_i}{h} \right). \tag{15.1}$$

When computing $f_{n,h}(t)$, we assign higher weights to observations x_i closer to t, in contrast to the histogram where we simply count the number of observations in the bin that contains t. Note that as a consequence of condition (K1), $f_{n,h}$ itself is a probability density:

$$f_{n,h}(t) \geq 0 \text{ and } \int_{-\infty}^{\infty} f_{n,h}(t) \, dt = 1.$$

QUICK EXERCISE 15.4 Check that the total area under the kernel density estimate is equal to one, i.e., show that $\int_{-\infty}^{\infty} f_{n,h}(t) \, dt = 1$.

Note that computing $f_{n,h}$ is very computationally intensive. Its common use nowadays is therefore a typical product of the recent developments in computer hardware, despite the fact that the method was introduced much earlier.

Choice of the bandwidth

The bandwidth h plays the same role for kernel density estimates as the bin width b does for histograms. In Figure 15.6 three kernel density estimates of the Old Faithful data are plotted with the triweight kernel and bandwidths 1.8, 18, and 180. It is clear that the choice of the bandwidth h determines largely what the resulting kernel density estimate will look like. Choosing the bandwidth too small will produce a curve with many isolated peaks. Choosing the bandwidth too large will produce a very smooth curve, at the risk of smoothing away important features of the data. In Figure 15.6 bandwidth

$h = 1.8$ is somewhat too small. Bandwidth $h = 180$ is clearly too large and produces an oversmoothed kernel density estimate that no longer captures the fact that the data show two separate modes.
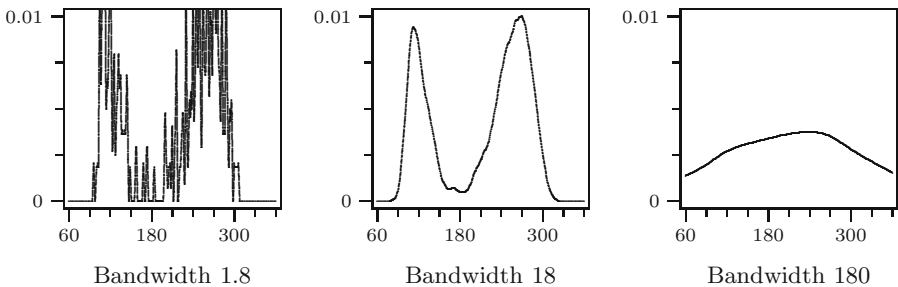
Fig. 15.6. Kernel estimates of the Old Faithful data.

How does one go about choosing the bandwidth? Similar to histograms, in practice one could do this by trial and error and continue until one obtains a reasonable picture. Recent research, however, has provided some guidelines for a data-based choice of h. A formula that may effectively be used is $h = 1.06\,s n^{-1/5}$, where s denotes the sample standard deviation (see, for instance, [31]; see also Remark 15.2).

Remark 15.2 (Normal reference method for kernel estimates). Suppose we view our dataset as being generated from a probability distribution with density f. Let K be a fixed chosen kernel and let $f_{n,h}$ be the kernel density estimate. We would like to take the bandwidth that minimizes the difference between $f_{n,h}$ and f, measured by the so-called mean integrated squared error (MISE)

$$\mathrm{E}\left[\int_{-\infty}^{\infty} (f_{n,h}(x) - f(x))^2 \, \mathrm{d}x\right].$$

Under suitable smoothness conditions on f, the value of h that minimizes the MISE, as n goes to infinity, is given by

$$h = C_1(f)C_2(K)n^{-1/5},$$

where the constants $C_1(f)$ and $C_2(K)$ are given by

$$C_1(f) = \left(\frac{1}{\int_{-\infty}^{\infty} f''(x)^2 \, \mathrm{d}x}\right)^{1/5} \quad \text{and} \quad C_2(K) = \frac{\left(\int_{-\infty}^{\infty} K(u)^2 \, \mathrm{d}u\right)^{1/5}}{\left(\int_{-\infty}^{\infty} u^2 K(u) \, \mathrm{d}u\right)^{2/5}}.$$

After choosing the kernel K, one can compute the constant $C_2(K)$ to obtain a simple data-based choice for h by estimating the constant $C_1(f)$. For instance, for the normal kernel one finds $C_2(K) = (2\sqrt{\pi})^{-1/5}$. As with

histograms (see Remark 15.1), the normal reference method takes f to be the density of an $N(\mu, \sigma^2)$ distribution, in which case $C_1(f) = (8\sqrt{\pi}/3)^{1/5}\sigma$. Estimating σ by the sample standard deviation s (see Chapter 16 for a definition of s) would result in bandwidth

$$h = \left(\tfrac{4}{3}\right)^{1/5} sn^{-1/5}.$$

For the Old Faithful data, this would give $h = 23.64$.

QUICK EXERCISE 15.5 If we construct a kernel density estimate for the Old Faithful data with bandwidth $h = 1.06sn^{-1/5}$, then on what interval is $f_{n,h}$ strictly positive if $s = 68.48$?

Choice of the kernel

To construct a kernel density estimate, one has to choose a kernel K and a bandwidth h. The choice of kernel is less important. In Figure 15.7 we have plotted two kernel density estimates for the Old Faithful data of Table 15.1: one is constructed with the triweight kernel (solid line), and one with the Epanechnikov kernel (dotted line), both with the same bandwidth $h = 24$. As one can see, the graphs are very similar. If one wants to compare with the normal kernel, one should set the bandwidth of the normal kernel at about $h/4$. This has to do with the fact that the normal kernel is much more spread out than the two kernels mentioned here, which are zero outside $[-1, 1]$.
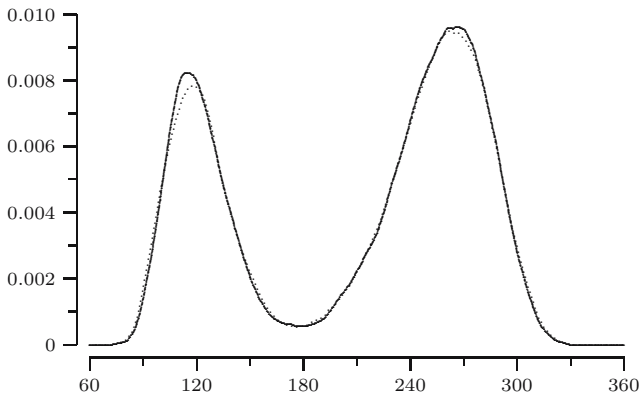
Fig. 15.7. Kernel estimates of the Old Faithful data with different kernels: triweight (solid line) and Epanechnikov kernel (dotted), both with bandwidth $h = 24$.

Boundary kernels

In order to estimate the parameters of a software reliability model, failure data are collected. Usually the most desirable type of failure data results when the

Table 15.3. Interfailure times between successive failures.

30	113	81	115	9	2	91	112	15	138
50	77	24	108	88	670	120	26	114	325
55	242	68	422	180	10	1146	600	15	36
4	0	8	227	65	176	58	457	300	97
263	452	255	197	193	6	79	816	1351	148
21	233	134	357	193	236	31	369	748	0
232	330	365	1222	543	10	16	529	379	44
129	810	290	300	529	281	160	828	1011	445
296	1755	1064	1783	860	983	707	33	868	724
2323	2930	1461	843	12	261	1800	865	1435	30
143	108	0	3110	1247	943	700	875	245	729
1897	447	386	446	122	990	948	1082	22	75
482	5509	100	10	1071	371	790	6150	3321	1045
648	5485	1160	1864	4116					

failure times are recorded, or equivalently, the length of an interval between successive failures. The data in Table 15.3 are observed interfailure times in CPU seconds for a certain control software system. On the left in Figure 15.8 a kernel density estimate of the observed interfailure times is plotted. Note that to the left of the origin, $f_{n,h}$ is positive. This is absurd, since it suggests that there are negative interfailure times.

This phenomenon is a consequence of the fact that one uses a symmetric kernel. In that case, the resulting kernel density estimate will always be positive on the interval $[x_i - h, x_i + h]$ for every element x_i in the dataset. Hence, obser-
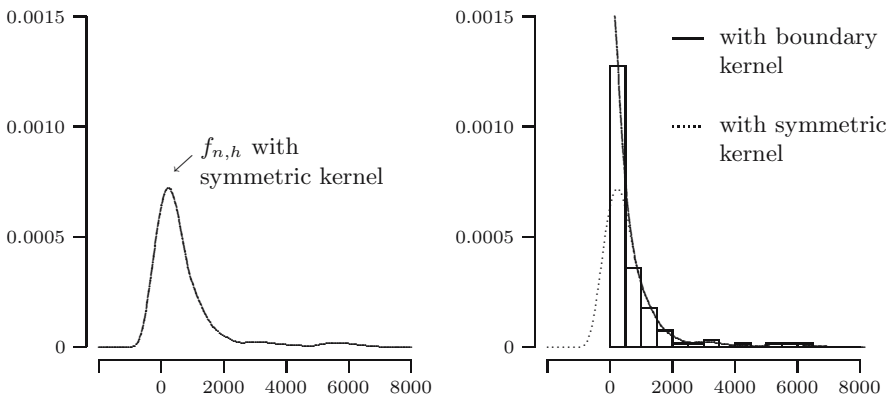
Fig. 15.8. Kernel density estimate of the software reliability data with symmetric and boundary kernel.

vations close to zero will cause the kernel density estimate $f_{n,h}$ to be positive to the left of zero. It is possible to improve the kernel density estimate in a neighborhood of zero by means of a so-called boundary kernel. Without going into detail about the construction of such an improvement, we will only show the result of this. On the right in Figure 15.8 the histogram of the interfailure times is plotted together with the kernel density estimate constructed with a symmetric kernel (dotted line) and with the boundary kernel density estimate (solid line). The boundary kernel density estimate is 0 to the left of the origin and is adjusted on the interval $[0, h)$. On the interval $[h, \infty)$ both kernel density estimates are the same.

15.4 The empirical distribution function

Another way to graphically represent a dataset is to plot the data in a cumulative manner. This can be done using the *empirical cumulative distribution function* of the data. It is denoted by F_n and is defined at a point x as the proportion of elements in the dataset that are less than or equal to x:

$$F_n(x) = \frac{\text{number of elements in the dataset} \leq x}{n}.$$

To illustrate the construction of F_n, consider the dataset consisting of the elements

$$4 \quad 3 \quad 9 \quad 1 \quad 7.$$

The corresponding empirical distribution function is displayed in Figure 15.9. For $x < 1$, there are no elements less than or equal to x, so that $F_n(x) = 0$. For $1 \leq x < 3$, only the element 1 is less than or equal to x, so that $F_n(x) = 1/5$. For $3 \leq x < 4$, the elements 1 and 3 are less than or equal to x, so that $F_n(x) = 2/5$, and so on.

In general, the graph of F_n has the form of a staircase, with $F_n(x) = 0$ for all x smaller than the minimum of the dataset and $F_n(x) = 1$ for all x greater than the maximum of the dataset. Between the minimum and maximum, F_n has a jump of size $1/n$ at each element of the dataset and is constant between successive elements. In Figure 15.9, the marks $\bullet$ and $\circ$ are added to the graph to emphasize the fact that, for instance, the value of $F_n(x)$ at $x = 3$ is 0.4, not 0.2. Usually, we leave these out, and one might also connect the horizontal segments by vertical lines.

In Figure 15.10 the empirical distribution functions are plotted for the Old Faithful data and the software reliability data. The fact that the Old Faithful data accumulate in the neighborhood of 120 and 270 is reflected in the graph of F_n by the fact that it is steeper at these places: the jumps of F_n succeed each other faster. In regions where the elements of the dataset are more stretched
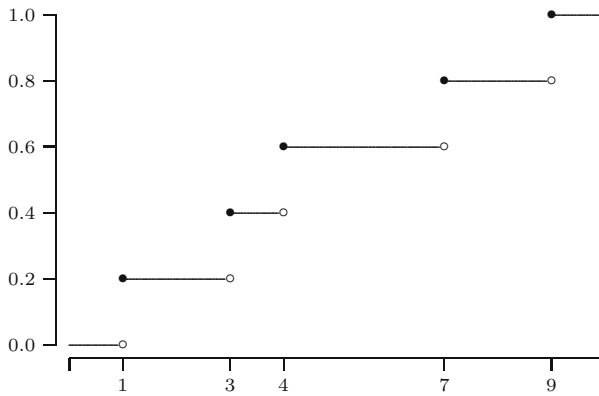
Fig. 15.9. Empirical distribution function.

out, the graph of F_n is flatter. Similar behavior can be seen for the software reliability data in the neighborhood of zero. The elements accumulate more close to zero, less as we move to the right. This is reflected by the empirical distribution function, which is very steep near zero and flattens out if we move to the right.

The graph of the empirical distribution function for the Old Faithful data agrees with the histogram in Figure 15.1 whose height is the largest on the bins (90, 120] and (240, 270]. In fact, there is a one-to-one relation between the two graphical summaries of the data: the area under the histogram on a single bin is equal to the relative frequency of elements that lie in that bin, which is also equal to the increase of F_n on that bin. For instance, the area under the histogram on bin (240, 270] for the Old Faithful data is equal to $30 \cdot 0.0092 =$
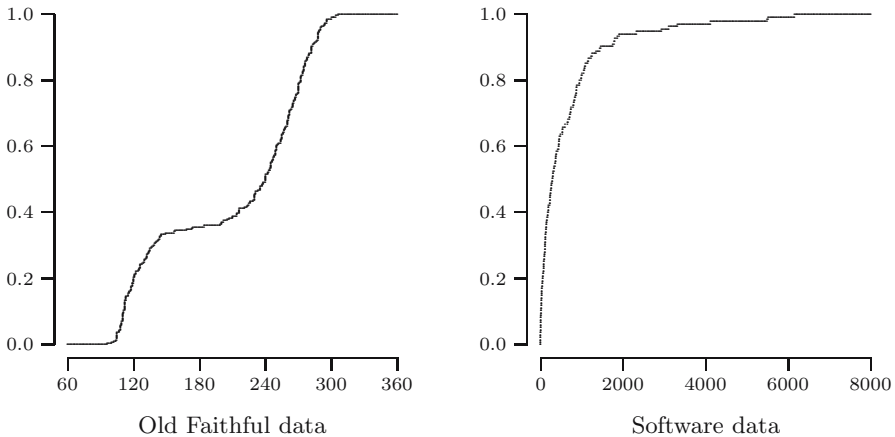
Fig. 15.10. Empirical distribution function of the Old Faithful data and the software reliability data.

0.276 (see Quick exercise 15.2). On the other hand, $F_n(270) = 215/272 = 0.7904$ and $F_n(240) = 140/272 = 0.5147$, whose difference $F_n(270) - F_n(240)$ is also equal to 0.276.

QUICK EXERCISE 15.6 Suppose that for a dataset consisting of 300 elements, the value of the empirical distribution function in the point 1.5 is equal to 0.7. How many elements in the dataset are strictly greater than 1.5?

> **Remark 15.3 (F_n as a discrete distribution function).** Note that F_n satisfies the four properties of a distribution function: it is continuous from the right, $F_n(x) \to 0$ as $x \to -\infty$, $F_n(x) \to 1$ as $x \to \infty$ and F_n is nondecreasing. This means that F_n itself is a distribution function of some random variable. Indeed, F_n is the distribution function of the discrete random variable that attains values $x_1, x_2, \ldots, x_n$ with equal probability $1/n$.

15.5 Scatterplot

In some situations one wants to investigate the relationship between two or more variables. In the case of two variables x and y, the dataset consists of *pairs of observations*:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

We call such a dataset a *bivariate* dataset in contrast to the *univariate* dataset, which consists of observations of one particular quantity. We often like to investigate whether the value of variable y depends on the value of the variable x, and if so, whether we can describe the relation between the two variables. A first step is to take a look at the data, i.e., to plot the points (x_i, y_i) for $i = 1, 2 \ldots, n$. Such a plot is called a *scatterplot*.

Drilling in rock

During a study about "dry" and "wet" drilling in rock, six holes were drilled, three corresponding to each process. In a dry hole one forces compressed air down the drill rods to flush the cutting and the drive hammer, whereas in a wet hole one forces water. As the hole gets deeper, one has to add a rod of 5 feet length to the drill. In each hole the time was recorded to advance 5 feet to a total depth of 400 feet. The data in Table 15.4 are in 1/100 minute and are derived from the original data in [23]. The original data consisted of drill times for each of the six holes and contained missing observations and observations that were known to be too large. The data in Table 15.4 are the mean drill times of the bona fide observations at each depth for dry and wet drilling.

One of the questions of interest is whether drill time depends on depth. To investigate this, we plot the mean drill time against depth. Figure 15.11 displays

Table 15.4. Mean drill time.

Depth	Dry	Wet	Depth	Dry	Wet
5	640.67	830.00	205	803.33	962.33
10	674.67	800.00	210	794.33	864.67
15	708.00	711.33	215	760.67	805.67
20	735.67	867.67	220	789.50	966.00
25	754.33	940.67	225	904.50	1010.33
30	723.33	941.33	230	940.50	936.33
35	664.33	924.33	235	882.00	915.67
40	727.67	873.00	240	783.50	956.33
45	658.67	874.67	245	843.50	936.00
50	658.00	843.33	250	813.50	803.67
55	705.67	885.67	255	658.00	697.33
60	700.00	881.67	260	702.50	795.67
65	720.67	822.00	265	623.50	1045.33
70	701.33	886.33	270	739.00	1029.67
75	716.67	842.50	275	907.50	977.00
80	649.67	874.67	280	846.00	1054.33
85	667.33	889.33	285	829.00	1001.33
90	612.67	870.67	290	975.50	1042.00
95	656.67	916.00	295	998.00	1200.67
100	614.00	888.33	300	1037.50	1172.67
105	584.00	835.33	305	984.00	1019.67
110	619.67	776.33	310	972.50	990.33
115	666.00	811.67	315	834.00	1173.33
120	695.00	874.67	320	675.00	1165.67
125	702.00	846.00	325	686.00	1142.00
130	739.67	920.67	330	963.00	1030.67
135	790.67	896.33	335	961.50	1089.67
140	730.33	810.33	340	932.00	1154.33
145	674.00	912.33	345	1054.00	1238.50
150	749.00	862.33	350	1038.00	1208.67
155	709.67	828.00	355	1238.00	1134.67
160	769.00	812.67	360	927.00	1088.00
165	663.00	795.67	365	850.00	1004.00
170	679.33	897.67	370	1066.00	1104.00
175	740.67	881.00	375	962.50	970.33
180	776.50	819.67	380	1025.50	1054.50
185	688.00	853.33	385	1205.50	1143.50
190	761.67	844.33	390	1168.00	1044.00
195	800.00	919.00	395	1032.50	978.33
200	845.50	933.33	400	1162.00	1104.00

Source: R. Penner and D.G. Watts. Mining information. *The American Statistician*, 45:4–9, 1991; Table 1 on page 6.

Dry holes
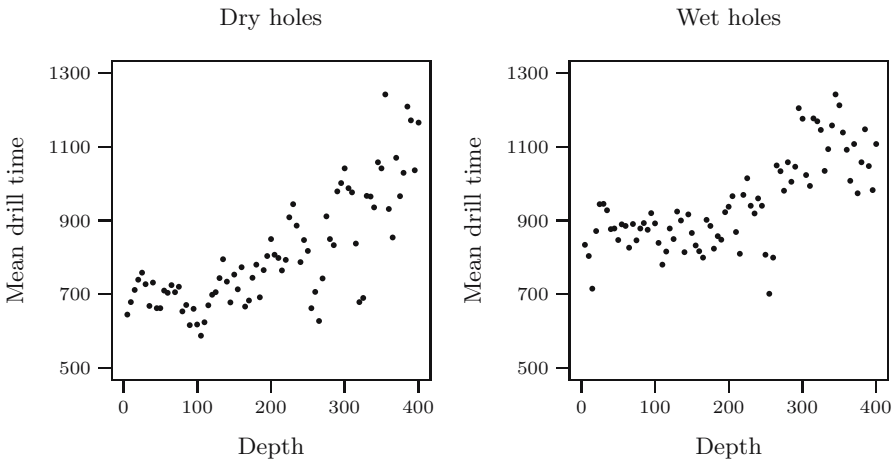
Wet holes

Fig. 15.11. Scatterplots of mean drill time versus depth.

the resulting scatterplots for the dry and wet holes. The scatterplots seem to indicate that in the beginning the drill time hardly depends on depth, at least up to, let's say, 250 feet. At greater depth, the drill time seems to vary over a larger range and increases somewhat with depth. A possible explanation for this is that the drill moved from softer to harder material. This was suggested by the fact that the drill hit an ore lens at about 250 feet and that the natural place such ore lenses occur is between two different materials (see [23] for details).

A more important question is whether one can drill holes faster using dry drilling or wet drilling. The scatterplots seem to suggest that dry drilling might be faster. We will come back to this later.

Predicting Janka hardness of Australian timber

The Janka hardness test is a standard test to measure the hardness of wood. It measures the force required to push a steel ball with a diameter of 11.28 millimeters (0.444 inch) into the wood to a depth of half the ball's diameter. To measure Janka hardness directly is difficult. However, it is related to the density of the wood, which is comparatively easy to measure. In Table 15.5 a bivariate dataset is given of density (x) and Janka hardness (y) of 36 Australian eucalypt hardwoods.

In order to get an impression of the relationship between hardness and density, we made a scatterplot of the bivariate dataset, which is displayed in Figure 15.12. It consists of all points (x_i, y_i) for $i = 1, 2, \ldots, 36$. The scatterplot might provide suggestions for the formula that describes the relationship between the variables x and y. In this case, a linear relationship between the two variables does not seem unreasonable. Later (Chapter 22) we will discuss

Table 15.5. Density and hardness of Australian timber.

Density	Hardness	Density	Hardness	Density	Hardness
24.7	484	39.4	1210	53.4	1880
24.8	427	39.9	989	56.0	1980
27.3	413	40.3	1160	56.5	1820
28.4	517	40.6	1010	57.3	2020
28.4	549	40.7	1100	57.6	1980
29.0	648	40.7	1130	59.2	2310
30.3	587	42.9	1270	59.8	1940
32.7	704	45.8	1180	66.0	3260
35.6	979	46.9	1400	67.4	2700
38.5	914	48.2	1760	68.8	2890
38.8	1070	51.5	1710	69.1	2740
39.3	1020	51.5	2010	69.1	3140

Source: E.J. Williams. *Regression analysis.* John Wiley & Sons Inc., New York, 1959; Table 3.1 on page 43.

how one can establish such a linear relationship by means of the observed pairs.

QUICK EXERCISE 15.7 Suppose we have a eucalypt hardwood tree with density 65. What would your prediction be for the corresponding Janka hardness?
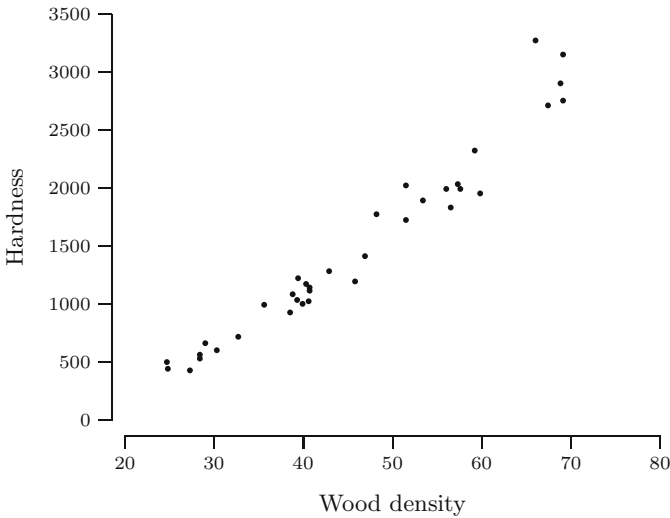
Fig. 15.12. Scatterplot of Janka hardness versus density of wood.

15.6 Solutions to the quick exercises

15.1 There are 272 elements in the dataset. The 91st and 182nd elements of the ordered data divide the dataset in three groups, each consisting of 90 elements. From a closer look at Table 15.2 we find that these two elements are 145 and 260.

15.2 In Table 15.2 one can easily count the number of observations in each of the bins $(90, 120], \ldots, (300, 330]$. The heights on each bin can be computed by dividing the number of observations in each bin by $272 \cdot 30 = 8160$. We get the following:

Bin	Count	Height	Bin	Count	Height
$(90, 120]$	55	0.0067	$(210, 240]$	34	0.0042
$(120, 150]$	37	0.0045	$(240, 270]$	75	0.0092
$(150, 180]$	5	0.0006	$(270, 300]$	54	0.0066
$(180, 210]$	9	0.0011	$(300, 330]$	3	0.0004

15.3 From Table 15.2 we see that we must cover an interval of length of at least $306 - 96 = 210$ with bins of width $b = 3.49 \cdot 68.48 \cdot 272^{-1/3} = 36.89$. Since $210/36.89 = 5.69$, we need at least six bins to cover the whole dataset.

15.4 By means of formula (15.1), we can write

$$\int_{-\infty}^{\infty} f_{n,h}(t) \, dt = \frac{1}{nh} \sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(\frac{t - x_i}{h}\right) \, dt.$$

For any $i = 1, \ldots, n$, we find by change of integration variables $t = hu + x_i$ that

$$\int_{-\infty}^{\infty} K\left(\frac{t - x_i}{h}\right) \, dt = h \int_{-\infty}^{\infty} K(u) \, du = h,$$

where we also use condition (K1). This directly yields

$$\int_{-\infty}^{\infty} f_{n,h}(t) \, dt = \frac{1}{nh} \cdot n \cdot h = 1.$$

15.5 The kernel density estimate will be strictly positive between the minimum minus h and the maximum plus h. The bandwidth equals $h = 1.06 \cdot 68.48 \cdot 272^{-1/5} = 23.66$. From Table 15.2, we see that this will be between $96 - 23.66 = 72.34$ and $306 + 23.66 = 329.66$.

15.6 By definition the number of elements less than or equal to 1.5 is $F_{300}(1.5) \cdot 300 = 210$. Hence 90 elements are strictly greater than 1.5.

15.7 Just by drawing a straight line that seems to fit the datapoints well, the authors predicted a Janka hardness of about 2700.

15.7 Exercises

15.1 In [33] Stephen Stigler discusses data from the *Edinburgh Medical and Surgical Journal* (1817). These concern the chest circumference of 5732 Scottish soldiers, measured in inches. The following information is given about the histogram with bin width 1, the first bin starting at 32.5.

Bin	Count	Bin	Count
(32.5, 33.5]	3	(40.5, 41.5]	935
(33.5, 34.5]	19	(41.5, 42.5]	646
(34.5, 35.5]	81	(42.5, 43.5]	313
(35.5, 36.5]	189	(43.5, 44.5]	168
(36.5, 37.5]	409	(44.5, 45.5]	50
(37.5, 38.5]	753	(45.5, 46.5]	18
(38.5, 39.5]	1062	(46.5, 47.5]	3
(39.5, 40.5]	1082	(47.5, 48.5]	1

Source: S.M. Stigler. *The history of statistics – The measurement of uncertainty before 1900.* Cambridge, Massachusetts, 1986.

a. Compute the height of the histogram on each bin.

b. Make a sketch of the histogram. Would you view the dataset as being symmetric or skewed?

15.2 Recall the example of the space shuttle *Challenger* in Section 1.4. The following list contains the launch temperatures in degrees Fahrenheit during previous takeoffs.

$$66 \quad 70 \quad 69 \quad 68 \quad 67 \quad 72 \quad 73 \quad 70 \quad 57 \quad 63 \quad 70 \quad 78$$
$$67 \quad 53 \quad 67 \quad 75 \quad 70 \quad 81 \quad 76 \quad 79 \quad 75 \quad 76 \quad 58$$

Source: Presidential commission on the space shuttle *Challenger* accident. Report on the space shuttle *Challenger* accident. Washington, DC, 1986; table on pages 129–131.

a. Compute the heights of a histogram with bin width 5, the first bin starting at 50.

b. On January 28, 1986, during the launch of the space shuttle *Challenger*, the temperature was 31 degrees Fahrenheit. Given the dataset of launch temperatures of previous takeoffs, would you consider 31 as a representative launch temperature?

15.3 □ In an article in *Biometrika*, an example is discussed about mine disasters during the period from March 15, 1851, to March, 22, 1962. A dataset has been obtained of 190 recorded time intervals (in days) between successive coal mine disasters involving ten or more men killed. The ordered data are listed in Table 15.6.

Table 15.6. Number of days between successive coal mine disasters.

0	1	1	2	2	3	4	4	4	6
7	10	11	12	12	12	13	15	15	16
16	16	17	17	18	19	19	19	20	20
22	23	24	25	27	28	29	29	29	31
31	32	33	34	34	36	36	37	40	41
41	42	43	45	47	48	49	50	53	54
54	55	56	59	59	61	61	65	66	66
70	72	75	78	78	78	80	80	81	88
91	92	93	93	95	95	96	96	97	99
101	108	110	112	113	114	120	120	123	123
124	124	125	127	129	131	134	137	139	143
144	145	151	154	156	157	176	182	186	187
188	189	190	193	194	197	202	203	208	215
216	217	217	217	218	224	225	228	232	233
250	255	275	275	275	276	286	292	307	307
312	312	315	324	326	326	329	330	336	345
348	354	361	364	368	378	388	420	431	456
462	467	498	517	536	538	566	632	644	745
806	826	871	952	1205	1312	1358	1630	1643	2366

Source: R.G. Jarrett. A note on the intervals between coal mining disasters.
Biometrika, 66:191-193, 1979; by permission of the Biometrika Trustees.

a. Compute the height on each bin of the histogram with bins $[0, 250]$, $(250, 500], \ldots, (2250, 2500]$.

b. Make a sketch of the histogram. Would you view the dataset as being symmetric or skewed?

15.4 ⊡ The ordered software data (see also Table 15.3) are given in the following list.

0	0	0	2	4	6	8	9	10	10
10	12	15	15	16	21	22	24	26	30
30	31	33	36	44	50	55	58	65	68
75	77	79	81	88	91	97	100	108	108
112	113	114	115	120	122	129	134	138	143
148	160	176	180	193	193	197	227	232	233
236	242	245	255	261	263	281	290	296	300
300	325	330	357	365	369	371	379	386	422
445	446	447	452	457	482	529	529	543	600
648	670	700	707	724	729	748	790	810	816
828	843	860	865	868	875	943	948	983	990
1011	1045	1064	1071	1082	1146	1160	1222	1247	1351
1435	1461	1755	1783	1800	1864	1897	2323	2930	3110
3321	4116	5485	5509	6150					

a. Compute the heights on each bin of the histogram with bins $[0, 500]$, $(500, 1000]$, and so on.

b. Compute the value of the empirical distribution function in the endpoints of the bins.

c. Check that the area under the histogram on bin $(1000, 1500]$ is equal to the increase $F_n(1500) - F_n(1000)$ of the empirical distribution function on this bin. Actually, this is true for each single bin (see Exercise 15.11).

15.5 ⊡ Suppose we construct a histogram with bins $[0,1]$, $(1,3]$, $(3,5]$, $(5,8]$, $(8,11]$, $(11,14]$, and $(14,18]$. Given are the values of the empirical distribution function at the boundaries of the bins:

t	0	1	3	5	8	11	14	18
$F_n(t)$	0	0.225	0.445	0.615	0.735	0.805	0.910	1.000

Compute the height of the histogram on each bin.

15.6 ⊞ Given is the following information about a histogram:

Bin	Height
(0,2]	0.245
(2,4]	0.130
(4,7]	0.050
(7,11]	0.020
(11,15]	0.005

Compute the value of the empirical distribution function in the point $t = 7$.

15.7 In Exercise 15.2 a histogram was constructed for the *Challenger* data. On which bin does the empirical distribution function have the largest increase?

15.8 Define a function K by

$$K(u) = \cos(\pi u) \quad \text{for } -1 \le u \le 1$$

and $K(u) = 0$ elsewhere. Check whether K satisfies the conditions (K1)–(K3) for a kernel function.

15.9 On the basis of the duration of an eruption of the Old Faithful geyser, park rangers try to predict the waiting time to the next eruption. In Figure 15.13 a scatterplot is displayed of the duration and the time to the next eruption in seconds.

a. Does the scatterplot give reason to believe that the duration of an eruption influences the time to the next eruption?
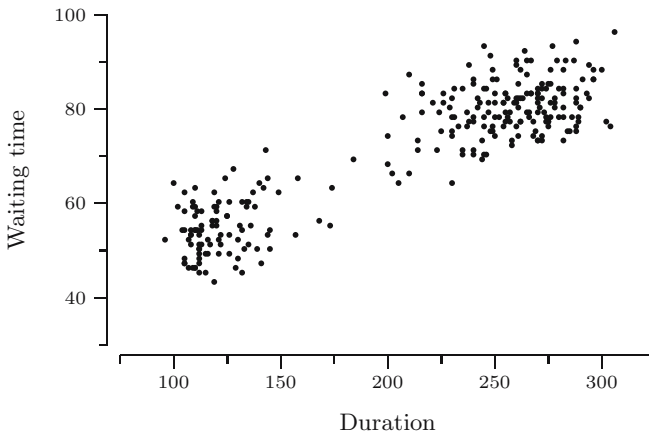
Fig. 15.13. Scatterplot of the Old Faithful data.

b. Suppose you have just observed an eruption that lasted 250 seconds. What would you predict for the time to the next eruption?

c. The dataset of durations shows two modes, i.e., there are two places where the data accumulate (see, for instance, the histogram in Figure 15.1). How many modes does the dataset of waiting times show?

15.10 Figure 15.14 displays the graph of an empirical distribution function of a dataset consisting of 200 elements. How many modes does the dataset show?
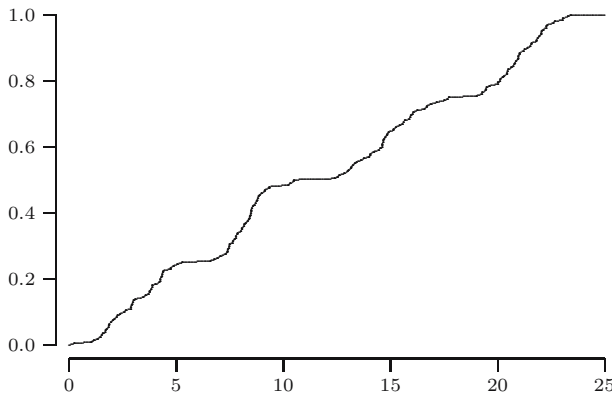
Fig. 15.14. Empirical distribution function.

15.11 ⊞ Given is a histogram and the empirical distribution function F_n of the same dataset. Show that the height of the histogram on a bin $(a, b]$ is

equal to

$$\frac{F_n(b) - F_n(a)}{b - a}.$$

15.12 ⊞ Let $f_{n,h}$ be a kernel estimate. As mentioned in Section 15.3, $f_{n,h}$ itself is a probability density.

a. Show that the corresponding expectation is equal to

$$\int_{-\infty}^{\infty} t f_{n,h}(t) \, dt = \bar{x}_n.$$

Hint: you might consult the solution to Quick exercise 15.4.

b. Show that the second moment corresponding to $f_{n,h}$ satisfies

$$\int_{-\infty}^{\infty} t^2 f_{n,h}(t) \, dt = \frac{1}{n} \sum_{i=1}^{n} x_i^2 + h^2 \int_{-\infty}^{\infty} u^2 K(u) \, du.$$

16

Exploratory data analysis: numerical summaries

The classical way to describe important features of a dataset is to give several numerical summaries. We discuss numerical summaries for the center of a dataset and for the amount of variability among the elements of a dataset, and then we introduce the notion of quantiles for a dataset. To distinguish these quantities from corresponding notions for probability distributions of random variables, we will often add the word *sample* or *empirical*; for instance, we will speak of the sample mean and empirical quantiles. We end this chapter with the *boxplot*, which combines some of the numerical summaries in a graphical display.

16.1 The center of a dataset

The best-known method to identify the *center* of a dataset is to compute the *sample mean*

$$\bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}. \tag{16.1}$$

For the sake of notational convenience we will sometimes drop the subscript n and write $\bar{x}$ instead of $\bar{x}_n$. The following dataset consists of hourly temperatures in degrees Fahrenheit (rounded to the nearest integer), recorded at Wick in northern Scotland from 5 p.m. December 31, 1960, to 3 a.m. January 1, 1961. The sample mean of the 11 measurements is equal to 44.7.

$$43 \quad 43 \quad 41 \quad 41 \quad 41 \quad 42 \quad 43 \quad 58 \quad 58 \quad 41 \quad 41$$

Source: V. Barnett and T. Lewis. *Outliers in statistical data.* Third edition, 1994. © John Wiley & Sons Limited. Reproduced with permission.

Another way to identify the center of a dataset is by means of the *sample median*, which we will denote by $\mathrm{Med}(x_1, x_2, \ldots, x_n)$ or briefly Med_n. The sample median is defined as the middle element of the dataset when it is put in ascending order. When n is odd, it is clear what this means. When n is even,

we take the average of the two middle elements. For the Wick temperature data the sample median is equal to 42.

QUICK EXERCISE 16.1 Compute the sample mean and sample median of the dataset

$$4.6 \quad 3.0 \quad 3.2 \quad 4.2 \quad 5.0.$$

Both methods have pros and cons. The sample mean is the natural analogue for a dataset of what the expectation is for a probability distribution. However, it is very sensitive to *outliers*, by which we mean observations in the dataset that deviate a lot from the bulk of the data.

To illustrate the sensitivity of the sample mean, consider the Wick temperature data displayed in Figure 16.1. The values 58 and 58 recorded at midnight and 1 a.m. are clearly far from the bulk of the data and give grounds for concern whether they are genuine (58 degrees Fahrenheit seems very warm at midnight for New Year's in northern Scotland). To investigate their effect on the sample mean we compute the average of the data, leaving out these measurements, which gives 41.8 (instead of 44.7). The sample median of the data is equal to 41 (instead of 42) when leaving out the measurements with value 58. The median is more robust in the sense that it is hardly affected by a few outliers.
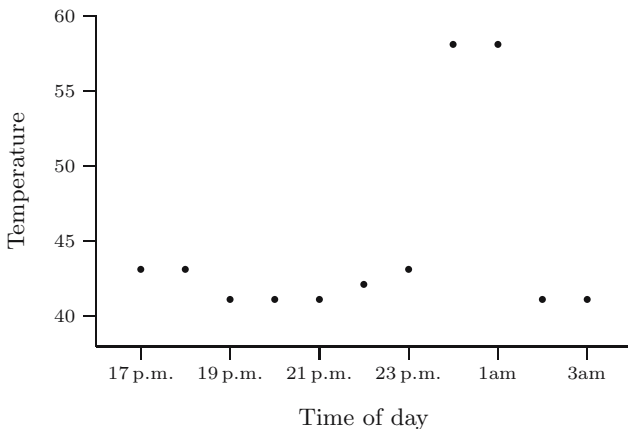
Fig. 16.1. The Wick temperature data.

It should be emphasized that this discussion is only meant to illustrate the sensitivity of the sample mean and by no means is intended to suggest we leave out measurements that deviate a lot from the bulk of the data! It is important to be *aware* of the presence of an outlier. In that case, one could try to find out whether there is perhaps something suspicious about this measurement. This might lead to assigning a smaller weight to such a measurement or even to

removing it from the dataset. However, sometimes it is possible to reconstruct the exact circumstances and correct the measurement. For instance, after further inquiry in the temperature example it turned out that at midnight the meteorological office changed its recording unit from degrees Fahrenheit to 1/10th degree Celsius (so 58 and 41 should read 5.8°C and 4.1°C). The corrected values in degrees Fahrenheit (to the nearest integer) are

$$43 \quad 43 \quad 41 \quad 41 \quad 41 \quad 42 \quad 43 \quad \mathit{42} \quad \mathit{42} \quad \mathit{39} \quad \mathit{39.}$$

For the corrected data the sample mean is 41.5 and the sample median is 42.

QUICK EXERCISE 16.2 Consider the same dataset as in Quick exercise 16.1. Suppose that someone misreads the dataset as

$$4.6 \quad 30 \quad 3.2 \quad 4.2 \quad 50.$$

Compute the sample mean and sample median and compare these values with the ones you found in Quick exercise 16.1.

16.2 The amount of variability of a dataset

To quantify the amount of variability among the elements of a dataset, one often uses the *sample variance* defined by

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2.$$

Up to a scaling factor this is equal to the average squared deviation from $\bar{x}_n$. At first sight, it seems more natural to define the sample variance by

$$\tilde{s}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2.$$

Why we choose the factor $1/(n-1)$ instead of $1/n$ will be explained later (see Chapter 19). Because s_n^2 is in different units from the elements of the dataset, one often prefers the *sample standard deviation*

$$s_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2},$$

which is measured in the same units as the elements of the dataset itself.

Just as the sample mean, the sample standard deviation is very sensitive to outliers. For the (uncorrected) Wick temperature data the sample standard deviation is 6.62, or 0.97 if we leave out the two measurements with value 58.

For the corrected data the standard deviation is 1.44. A more robust measure of variability is the *median of absolute deviations* or MAD, which is defined as follows. Consider the absolute deviation of every element x_i with respect to the sample median:

$$|x_i - \text{Med}(x_1, x_2, \ldots, x_n)|$$

or briefly

$$|x_i - \text{Med}_n|.$$

The MAD is obtained by taking the median of all these absolute deviations

$$\text{MAD}(x_1, x_2, \ldots, x_n) = \text{Med}(|x_1 - \text{Med}_n|, \ldots, |x_n - \text{Med}_n|). \qquad (16.2)$$

QUICK EXERCISE 16.3 Compute the sample standard deviation for the dataset of Quick exercise 16.1 for which it is given that the values of $x_i - \bar{x}_n$ are:

$$-1.0, \ 0.6, \ -0.8, \ 0.2, \ 1.0.$$

Also compute the MAD for this dataset.

Just as the sample median, the MAD is hardly affected by outliers. For the (uncorrected) Wick temperature data the MAD is 1 and equal to 0 if we leave out the two measurements with value 58 (the value 0 seems a bit strange, but is a consequence of the fact that the observations are given in degrees Fahrenheit rounded to the nearest integer). For the corrected data the MAD is 1.

QUICK EXERCISE 16.4 Compute the sample standard deviation for the mis-read dataset of Quick exercise 16.2 for which it is given that the values of $x_i - \bar{x}_n$ are:

$$11.6, \ -13.8, \ -15.2, \ -14.2, \ 31.6.$$

Also compute the MAD for this dataset and compare both values with the ones you found in Quick exercise 16.3.

16.3 Empirical quantiles, quartiles, and the IQR

The sample median divides the dataset in two more or less equal parts: about half of the elements are less than the median and about half of the elements are greater than the median. More generally, we can divide the dataset in two parts in such a way that a proportion p is less than a certain number and a proportion $1 - p$ is greater than this number. Such a number is called the $100p$ *empirical percentile* or the *pth empirical quantile* and is denoted by $q_n(p)$. For a suitable introduction of empirical quantiles we need the notion of order statistics.

The *order statistics* consist of the same elements as in the original dataset $x_1, x_2, \ldots, x_n$, but in ascending order. Denote by $x_{(k)}$ the kth element in the ordered list. Then

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

are called the order statistics of $x_1, x_2, \ldots, x_n$. The order statistics of the Wick temperature data are

$$41 \quad 41 \quad 41 \quad 41 \quad 41 \quad 42 \quad 43 \quad 43 \quad 43 \quad 58 \quad 58.$$

Note that by putting the elements in order, it is possible that successive order statistics are the same, for instance, $x_{(1)} = \cdots = x_{(5)} = 41$. Another example is Table 15.2, which lists the order statistics of the Old Faithful dataset.

To compute empirical quantiles one linearly interpolates between order statistics of the dataset. Let $0 < p < 1$, and suppose we want to compute the pth empirical quantile for a dataset $x_1, x_2, \ldots, x_n$. The following computation is based on requiring that the ith order statistic is the $i/(n+1)$ quantile. If we denote the integer part of a by $\lfloor a \rfloor$, then the computation of $q_n(p)$ runs as follows:

$$q_n(p) = x_{(k)} + \alpha(x_{(k+1)} - x_{(k)})$$

with $k = \lfloor p(n+1) \rfloor$ and $\alpha = p(n+1) - k$. On the left in Figure 16.2 the relation between the pth quantile and the empirical distribution function is illustrated for the Old Faithful data.
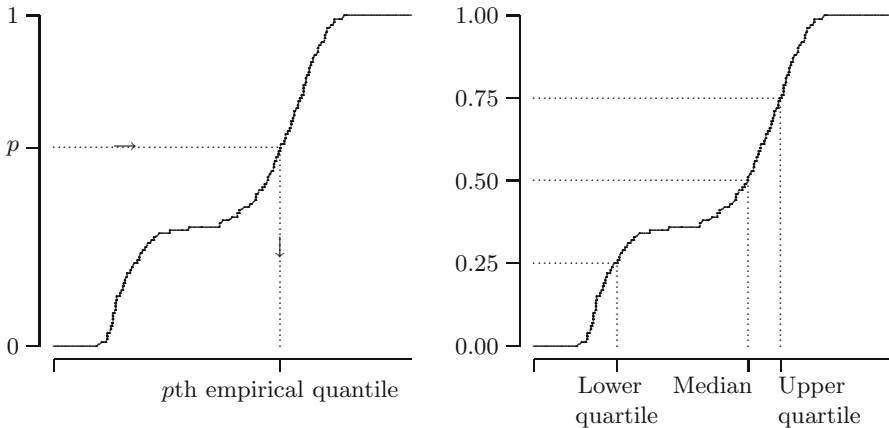
Fig. 16.2. Empirical quantile and quartiles for the Old Faithful data.

QUICK EXERCISE 16.5 Compute the 55th empirical percentile for the Wick temperature data.

Lower and upper quartiles

Instead of identifying only the center of the dataset, Tukey [35] suggested to give a five-number summary of the dataset: the minimum, the maximum, the sample median, and the 25th and 75th empirical percentiles. The 25th empirical percentile $q_n(0.25)$ is called the *lower quartile* and the 75th empirical percentile $q_n(0.75)$ is called the *upper quartile*. Together with the median, the lower and upper quartiles divide the dataset in four more or less equal parts consisting of about one quarter of the number of elements. The relation of the two quartiles and the median with the empirical distribution function is illustrated for the Old Faithful data on the right of Figure 16.2. The distance between the lower quartile and the median, relative to the distance between the upper quartile and the median, gives some indication on the skewness of the dataset. The distance between the upper and lower quartiles is called the *interquartile range*, or IQR:

$$\text{IQR} = q_n(0.75) - q_n(0.25).$$

The IQR specifies the range of the middle half of the dataset. It could also serve as a robust measure of the amount of variability among the elements of the dataset. For the Old Faithful data the five-number summary is

Minimum	Lower quartile	Median	Upper quartile	Maximum
96	129.25	240	267.75	306

and the IQR is 138.5.

QUICK EXERCISE 16.6 Compute the five-number summary for the (uncorrected) Wick temperature data.

16.4 The box-and-whisker plot

Tukey [35] also proposed visualizing the five-number summary discussed in the previous section by a so-called box-and-whisker plot, briefly *boxplot*. Figure 16.3 displays a boxplot. The data are now on the vertical axis, where we left out the numbers on the axis in order to explain the construction of the figure. The horizontal width of the box is irrelevant. In the vertical direction the box extends from the lower to the upper quartile, so that the height of the box is precisely the IQR. The horizontal line inside the box corresponds to the sample median. Up from the upper quartile we measure out a distance of 1.5 times the IQR and draw a so-called *whisker* up to the largest observation that lies within this distance, where we put a horizontal line. Similarly, down from the lower quartile we measure out a distance of 1.5 times the IQR and draw a whisker to the smallest observation that lies within this distance, where we also put a horizontal line. *All* other observations beyond the whiskers are marked by ∘. Such an observation is called an *outlier*.
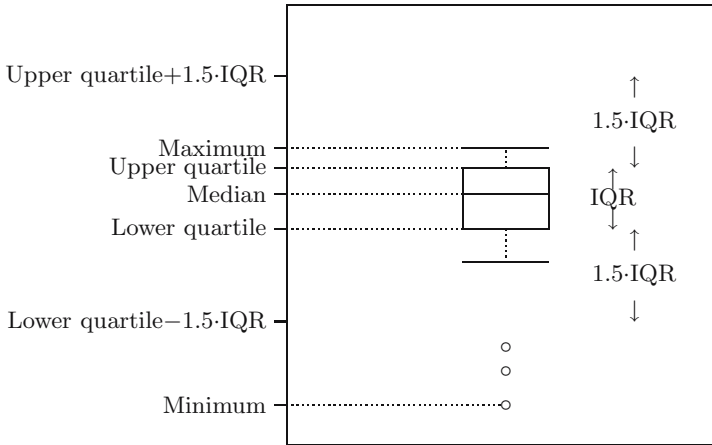
Fig. 16.3. A boxplot.

In Figure 16.4 the boxplots of the Old Faithful data and of the software reliability data (see also Chapter 15) are displayed. The skewness of the software reliability data produces a boxplot with whiskers of very different length and with several observations beyond the upper quartile plus 1.5 times the IQR. The boxplot of the Old Faithful data illustrates one of the shortcomings of the boxplot; it does not capture the fact that the data show two separate peaks. However, the position of the sample median inside the box does suggest that the dataset is skewed.

QUICK EXERCISE 16.7 Suppose we want to construct a boxplot of the (uncorrected) Wick temperature data. What is the height of the box, the length of both whiskers, and which measurements fall outside the box and whiskers? Would you consider the two values 58 extreme outliers?
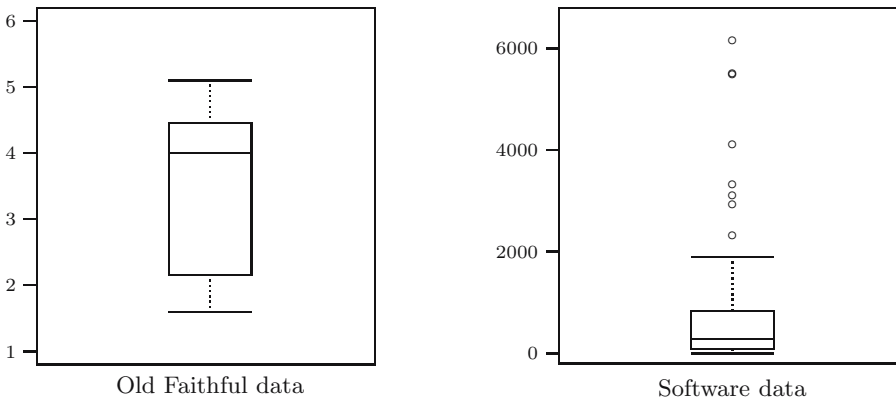
Fig. 16.4. Boxplot of the Old Faithful data and the software data.

Using boxplots to compare several datasets

Although the boxplot provides some information about the structure of the data, such as center, range, skewness or symmetry, it is a poor graphical display of the dataset. Graphical summaries such as the histogram and kernel density estimate are more informative displays of a single dataset. Boxplots become useful if we want to compare several sets of data in a simple graphical display. In Figure 16.5 boxplots are displayed of the average drill time for dry and wet drilling up to a depth of 250 feet for the drill data discussed in Section 15.5 (see also Table 15.4). It is clear that the boxplot corresponding to dry drilling differs from that corresponding to wet drilling. However, the question is whether this difference can still be attributed to chance or is caused by the drilling technique used. We will return to this type of question in Chapter 25.
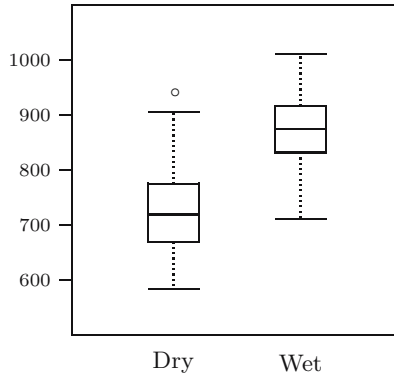
Fig. 16.5. Boxplot of average drill times.

16.5 Solutions to the quick exercises

16.1 The average is

$$\bar{x}_n = \frac{4.6 + 3.0 + 3.2 + 4.2 + 5.0}{5} = \frac{20}{5} = 4.$$

The median is the middle element of 3.0, 3.2, 4.2, 4.6, and 5.0, which gives $\text{Med}_n = 4.2$.

16.2 The average is

$$\bar{x}_n = \frac{4.6 + 30 + 3.2 + 4.2 + 50}{5} = \frac{90}{5} = 18,$$

which differs 14.4 from the average we found in Quick exercise 16.1. The median is the middle element of 3.2, 4.2, 4.6, 30, and 50. This gives $\text{Med}_n = 4.6$, which only differs 0.4 from the median we found in Quick exercise 16.1. As one can see, the median is hardly affected by the two outliers.

16.3 The sample variance is

$$s_n^2 = \frac{(-1)^2 + (0.6)^2 + (-0.8)^2 + (0.2)^2 + (1.0)^2}{5 - 1} = \frac{3.04}{4} = 0.76$$

so that the sample standard deviation is $s_n = \sqrt{0.76} = 0.872$. The median is 4.2, so that the absolute deviations from the median are given by

$$0.4 \quad 1.2 \quad 1.0 \quad 0.0 \quad 0.8.$$

The MAD is the median of these numbers, which is 0.8.

16.4 The sample variance is

$$s_n^2 = \frac{(11.6)^2 + (-13.8)^2 + (-15.2)^2 + (-14.2)^2 + (31.6)^2}{5 - 1} = \frac{1756.24}{4} = 439.06$$

so that the sample standard deviation is $s_n = \sqrt{439.06} = 20.95$, which is a difference of 20.19 from the value we found in Quick exercise 16.3. The median is 4.6, so that the absolute deviations from the median are given by

$$0.0 \quad 25.4 \quad 1.4 \quad 0.4 \quad 45.4.$$

The MAD is the median of these numbers, which is 1.4. Just as the median, the MAD is hardly affected by the two outliers.

16.5 We have $k = \lfloor 0.55 \cdot 12 \rfloor = \lfloor 6.6 \rfloor = 6$, so that $\alpha = 0.6$. This gives

$$q_n(0.55) = x_{(6)} + 0.6 \cdot (x_{(7)} - x_{(6)}) = 42 + 0.6 \cdot (43 - 42) = 42.6.$$

16.6 From the order statistics of the Wick temperature data

$$41 \quad 41 \quad 41 \quad 41 \quad 41 \quad 42 \quad 43 \quad 43 \quad 43 \quad 58 \quad 58$$

it can be seen immediately that minimum, maximum, and median are given by 41, 58, and 42. For the lower quartile we have $k = \lfloor 0.25 \cdot 12 \rfloor = 3$, so that $\alpha = 0$ and $q_n(0.25) = x_{(3)} = 41$. For the upper quartile we have $k = \lfloor 0.75 \cdot 12 \rfloor = 9$, so that again $\alpha = 0$ and $q_n(0.75) = x_{(9)} = 43$. Hence for the Wick temperature data the five-number summary is

Minimum	Lower quartile	Median	Upper quartile	Maximum
41	41	42	43	58

16.7 From the five-number summary for the Wick temperature data (see Quick exercise 16.6), it follows immediately that the height of the box is the IQR: $43 - 41 = 2$. If we measure out a distance of 1.5 times 2 down from the lower quartile 41, we see that the smallest observation within this range is 41, which means that the lower whisker has length zero. Similarly, the upper whisker has length zero. The two measurements with value 58 are outside the box and whiskers. The two values 58 are clearly far away from the bulk of the data and should be considered extreme outliers.
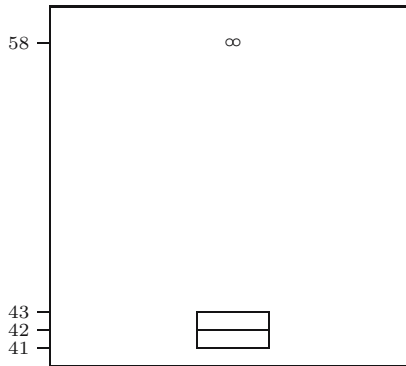
16.6 Exercises

16.1 ⊡ Use the order statistics of the software data as given in Exercise 15.4 to answer the following questions.

 a. Compute the sample median.
 b. Compute the lower and upper quartiles and the IQR.
 c. Compute the 37th empirical percentile.

16.2 Compute for the Old Faithful data the distance of the lower and upper quartiles to the median and explain the difference.

16.3 ⊞ Recall the example about the space shuttle *Challenger* in Section 1.4. The following table lists the order statistics of launch temperatures during take-offs in degrees Fahrenheit, including the launch temperature on January 28, 1986.

$$31 \quad 53 \quad 57 \quad 58 \quad 63 \quad 66 \quad 67 \quad 67 \quad 67 \quad 68 \quad 69 \quad 70$$
$$70 \quad 70 \quad 70 \quad 72 \quad 73 \quad 75 \quad 75 \quad 76 \quad 76 \quad 78 \quad 79 \quad 81$$

 a. Find the sample median and the lower and upper quartiles.
 b. Sketch the boxplot of this dataset.

c. On January 28, 1986, the launch temperature was 31 degrees Fahrenheit. Comment on the value 31 with respect to the other data points.

16.4 ⊡ The sample mean and sample median of the uncorrected Wick temperature data (in degrees Fahrenheit) are 44.7 and 42. We transform the data from degrees Fahrenheit (x_i) to degrees Celsius (y_i) by means of the formula

$$y_i = \frac{5}{9}(x_i - 32),$$

which gives the following dataset

$$\frac{55}{9} \quad \frac{55}{9} \quad 5 \quad 5 \quad 5 \quad \frac{50}{9} \quad \frac{55}{9} \quad \frac{130}{9} \quad \frac{130}{9} \quad 5 \quad 5.$$

a. Check that $\bar{y}_n = \frac{5}{9}(\bar{x}_n - 32)$.

b. Is it also true that $\text{Med}(y_1, \ldots, y_n) = \frac{5}{9}(\text{Med}(x_1, \ldots, x_n) - 32)$?

c. Suppose we have a dataset $x_1, x_2, \ldots, x_n$ and construct $y_1, y_2, \ldots, y_n$ where $y_i = ax_i + b$ with a and b being real numbers. Do similar relations hold for the sample mean and sample median? If so, state them.

16.5 Consider the uncorrected Wick temperature data in degrees Fahrenheit (x_i) and the corresponding temperatures in degrees Celsius (y_i) as given in Exercise 16.4. The sample standard deviation and the MAD for the Wick data are 6.62 and 1.

a. Let s_F and s_C denote the sample standard deviations of $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ respectively. Check that $s_C = \frac{5}{9}s_F$.

b. Let MAD_F and MAD_C denote the MAD of $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ respectively. Is it also true that $\text{MAD}_C = \frac{5}{9}\text{MAD}_F$?

c. Suppose we have a dataset $x_1, x_2, \ldots, x_n$ and construct $y_1, y_2, \ldots, y_n$ where $y_i = ax_i + b$ with a and b being real numbers. Do similar relations hold for the sample standard deviation and the MAD? If so, state them.

16.6 ⊞ Consider two datasets: $1, 5, 9$ and $2, 4, 6, 8$.

a. Denote the sample means of the two datasets by $\bar{x}$ and $\bar{y}$. Is it true that the average $(\bar{x} + \bar{y})/2$ of $\bar{x}$ and $\bar{y}$ is equal to the sample mean of the combined dataset with 7 elements?

b. Suppose we have two other datasets: one of size n with sample mean $\bar{x}_n$ and another dataset of size m with sample mean $\bar{y}_m$. Is it always true that the average $(\bar{x}_n + \bar{y}_m)/2$ of $\bar{x}_n$ and $\bar{y}_m$ is equal to the sample mean of the combined dataset with $n + m$ elements? If no, then provide a counterexample. If yes, then explain this.

c. If $m = n$, is $(\bar{x}_n + \bar{y}_m)/2$ equal to the sample mean of the combined dataset with $n + m$ elements?

16.7 Consider the two datasets from Exercise 16.6.

a. Denote the sample medians of the two datasets by Med_x and Med_y. Is it true that the sample median $(\mathrm{Med}_x + \mathrm{Med}_y)/2$ of the two sample medians is equal to the sample median of the combined dataset with 7 elements?

b. Suppose we have two other datasets: one of size n with sample median Med_x and another dataset of size m with sample median Med_y. Is it always true that the sample median $(\mathrm{Med}_x + \mathrm{Med}_y)/2$ of the two sample medians is equal to the sample median of the combined dataset with $n+m$ elements? If no, then provide a counterexample. If yes, then explain this.

c. What if $m = n$?

16.8 ⊞ Compute the MAD for the combined dataset of 7 elements from Exercise 16.6.

16.9 Consider a dataset $x_1, x_2, \ldots, x_n$ with $x_i \neq 0$. We construct a second dataset $y_1, y_2, \ldots, y_n$, where

$$ y_i = \frac{1}{x_i}. $$

a. Suppose dataset $x_1, x_2, \ldots, x_n$ consists of $-6, 1, 15$. Is it true that $\bar{y}_3 = 1/\bar{x}_3$?

b. Suppose that n is odd. Is it true that $\bar{y}_n = 1/\bar{x}_n$?

c. Suppose that n is odd and each $x_i > 0$. Is it true that $\mathrm{Med}(y_1, \ldots, y_n) = 1/\mathrm{Med}(x_1, \ldots, x_n)$? What about when n is even?

16.10 ⊡ A method to investigate the sensitivity of the sample mean and the sample median to extreme outliers is to replace one or more elements in a given dataset by a number y and investigate the effect when y goes to infinity. To illustrate this, consider the dataset from Quick Exercise 16.1:

$$ 4.6 \quad 3.0 \quad 3.2 \quad 4.2 \quad 5.0 $$

with sample mean 4 and sample median 4.2.

a. We replace the element 3.2 by some real number y. What happens with the sample mean and the sample median of this new dataset as $y \to \infty$?

b. We replace a number of elements by some real number y. How many elements do we need to replace so that the sample median of the new dataset goes to infinity as $y \to \infty$?

c. Suppose we have another dataset of size n. How many elements do we need to replace by some real number y, so that the sample mean of the new dataset goes to infinity as $y \to \infty$? And how many elements do we need to replace, so that the sample median of the new dataset goes to infinity?

16.11 Just as in Exercise 16.10 we investigate the sensitivity of the sample standard deviation and the MAD to extreme outliers, by considering the same dataset with sample standard deviation 0.872 and MAD equal to 0.8. Answer the same three questions for the sample standard deviation and the MAD instead of the sample mean and sample median.

16.12 ⊡ Compute the sample mean and sample median for the dataset

$$1, 2, \ldots, N$$

in case N is odd and in case N is even. You may use the fact that

$$1 + 2 + \cdots + N = \frac{N(N+1)}{2}.$$

16.13 Compute the sample standard deviation and MAD for the dataset

$$-N, \ldots, -1, 0, 1, \ldots, N.$$

You may use the fact that

$$1^2 + 2^2 + \cdots + N^2 = \frac{N(N+1)(2N+1)}{6}.$$

16.14 Check that the 50th empirical percentile is the sample median.

16.15 ⊞ The following rule is useful for the computation of the sample variance (and standard deviation). Show that

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - (\bar{x}_n)^2$$

where $\bar{x}_n = \left(\sum_{i=1}^{n} x_i \right)/n$.

16.16 Recall Exercise 15.12, where we computed the mean and second moment corresponding to a density estimate $f_{n,h}$. Show that the variance corresponding to $f_{n,h}$ satisfies:

$$\int_{-\infty}^{\infty} t^2 f_{n,h}(t) \, dt - \left(\int_{-\infty}^{\infty} t f_{n,h}(t) \, dt \right)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + h^2 \int_{-\infty}^{\infty} u^2 K(u) \, du.$$

16.17 Suppose we have a dataset $x_1, x_2, \ldots, x_n$. Check that if $p = i/(n+1)$ the pth empirical quantile is the ith order statistic.

17

Basic statistical models

In this chapter we introduce a common statistical model. It corresponds to the situation where the elements of the dataset are repeated measurements of the same quantity and where different measurements do not influence each other. Next, we discuss the probability distribution of the random variables that model the measurements and illustrate how *sample statistics* can help to select a suitable statistical model. Finally, we discuss the *simple linear regression model* that corresponds to the situation where the elements of the dataset are paired measurements.

17.1 Random samples and statistical models

In Chapter 1 we briefly discussed Michelson's experiment conducted between June 5 and July 2 in 1879, in which 100 measurements were obtained on the speed of light. The values are given in Table 17.1 and represent the speed of light in air in km/sec minus 299 000. The variation among the 100 values suggests that measuring the speed of light is subject to random influences. As we have seen before, we describe random phenomena by means of a probability model, i.e., we interpret the outcome of an experiment as a realization of some random variable. Hence the first measurement is modeled by a random variable X_1 and the value 850 is interpreted as the realization of X_1. Similarly, the second measurement is modeled by a random variable X_2 and the value 740 is interpreted as the realization of X_2. Since both measurements are obtained under the same experimental conditions, it is justified to assume that the probability distributions of X_1 and X_2 are the same. More generally, the 100 measurements are modeled by random variables

$$X_1, X_2, \ldots, X_{100}$$

with the same probability distribution, and the values in Table 17.1 are interpreted as realizations of $X_1, X_2, \ldots, X_{100}$. Moreover, because we believe that

Table 17.1. Michelson data on the speed of light.

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	770	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

Source: E.N. Dorsey. The velocity of light. *Transactions of the American Philosophical Society.* 34(1):1-110, 1944; Table 22 on pages 60-61.

Michelson took great care not to have the measurements influence each other, the random variables $X_1, X_2, \ldots, X_{100}$ are assumed to be *mutually independent* (see also Remark 3.1 about physical and stochastic independence). Such a collection of random variables is called a random sample or briefly, sample.

> RANDOM SAMPLE. A *random sample* is a collection of random variables $X_1, X_2, \ldots, X_n$, that have the same probability distribution and are mutually independent.

If F is the distribution function of each random variable X_i in a random sample, we speak of *a random sample from F*. Similarly, we speak of a random sample from a density f, a random sample from an $N(\mu, \sigma^2)$ distribution, etc.

QUICK EXERCISE 17.1 Suppose we have a random sample X_1, X_2 from a distribution with variance 1. Compute the variance of $X_1 + X_2$.

Properties that are inherent to the random phenomenon under study may provide additional knowledge about the distribution of the sample. Recall the software data discussed in Chapter 15. The data are observed lengths in CPU seconds between successive failures that occur during the execution of a certain real-time command. Typically, in a situation like this, in a small time interval, either 0 or 1 failure occurs. Moreover, failures occur with small probability and in disjoint time intervals failures occur independent of each other. In addition, let us assume that the rate at which the failures occur is constant over time. According to Chapter 12, this justifies the choice of a Poisson process to model the series of failures. From the properties of the Poisson process we know that the interfailure times are independent and have the same exponential distribution. Hence we model the software data as the realization of a random sample from an exponential distribution.

In some cases we may not be able to specify the type of distribution. Take, for instance, the Old Faithful data consisting of observed durations of eruptions of the Old Faithful geyser. Due to lack of specific geological knowledge about the subsurface and the mechanism that governs the eruptions, we prefer not to assume a particular type of distribution. However, we *do* model the durations as the realization of a random sample from a continuous distribution on $(0, \infty)$.

In each of the three examples the dataset was obtained from repeated measurements performed under the same experimental conditions. The basic statistical model for such a dataset is to consider the measurements as a random sample and to interpret the dataset as the realization of the random sample. Knowledge about the phenomenon under study and the nature of the experiment may lead to partial specification of the probability distribution of each X_i in the sample. This should be included in the model.

> STATISTICAL MODEL FOR REPEATED MEASUREMENTS. A dataset consisting of values $x_1, x_2, \ldots, x_n$ of repeated measurements of the same quantity is modeled as the realization of a random sample $X_1, X_2, \ldots, X_n$. The model may include a partial specification of the probability distribution of each X_i.

The probability distribution of each X_i is called the *model distribution*. Usually it refers to a collection of distributions: in the Old Faithful example to the collection of all continuous distributions on $(0, \infty)$, in the software example to the collection of all exponential distributions. In the latter case the parameter of the exponential distribution is called the *model parameter*. The unique distribution from which the sample actually originates is assumed to be one particular member of this collection and is called the *"true" distribution*. Similarly, in the software example, the parameter corresponding to the "true" exponential distribution is called the *"true" parameter*. The word *true* is put between quotation marks because it does not refer to something in the real world, but only to a distribution (or parameter) in the statistical model, which is merely an approximation of the real situation.

QUICK EXERCISE 17.2 We obtain a dataset of ten elements by tossing a coin ten times and recording the result of each toss. What is an appropriate statistical model and corresponding model distribution for this dataset?

Of course there are situations where the assumption of *independence* or *identical distributions* is unrealistic. In that case a different statistical model would be more appropriate. However, we will restrict ourselves mainly to the case where the dataset can be modeled as the realization of a random sample.

Once we have formulated a statistical model for our dataset, we can use the dataset to infer knowledge about the model distribution. Important questions about the corresponding model distribution are

- *which feature of the model distribution* represents the quantity of interest and *how do we use our dataset* to determine a value for this?
- *which model distribution* fits a particular dataset best?

These questions can be diverse, and answering them may be difficult. For instance, the Old Faithful data are modeled as a realization of a random sample from a continuous distribution. Suppose we are interested in a complete characterization of the "true" distribution, such as the distribution function F or the probability density f. Since there are no further specifications about the type of distribution, our problem would be to estimate the *complete curve* of F or f on the basis of our dataset.

On the other hand, the software data are modeled as the realization of a random sample from an exponential distribution. In that case F and f are completely characterized by a single parameter λ:

$$F(x) = 1 - e^{-\lambda x} \quad \text{and} \quad f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0.$$

Even if we are interested in the curves of F and f, our problem would reduce to estimating a *single parameter* on the basis of our dataset.

In other cases we may not be interested in the distribution as a whole, but only in a specific feature of the model distribution that represents the quantity of interest. For instance, in a physical experiment, such as the one performed by Michelson, one usually thinks of each measurement as

$$\text{measurement} = \text{quantity of interest} + \text{measurement error}.$$

The quantity of interest, in this case the speed of light, is thought of as being some (unknown) constant and the measurement error is some random fluctuation. In the absence of systematic error, the measurement error can be modeled by a random variable with zero expectation and finite variance. In that case the measurements are modeled by a random sample from a distribution with some unknown expectation and finite variance. The speed of light is represented by the expectation of the model distribution. Our problem would be to estimate the *expectation of the model distribution* on the basis of our dataset.

In the remaining chapters, we will develop several statistical methods to infer knowledge about the "true" distribution or about a specific feature of it, by means of a dataset. In the remainder of this chapter we will investigate how the graphical and numerical summaries of our dataset can serve as a first indication of what an appropriate choice would be for this distribution or for a specific feature, such as its expectation.

17.2 Distribution features and sample statistics

In Chapters 15 and 16 we have discussed several empirical summaries of datasets. They are examples of numbers, curves, and other objects that are a

function
$$h(x_1, x_2, \ldots, x_n)$$
of the dataset $x_1, x_2, \ldots, x_n$ only. Since datasets are modeled as realizations of random samples $X_1, X_2, \ldots, X_n$, an object $h(x_1, x_2, \ldots, x_n)$ is a realization of the corresponding random object
$$h(X_1, X_2, \ldots, X_n).$$

Such an object, which depends on the random sample $X_1, X_2, \ldots, X_n$ only, is called a *sample statistic*.

If a statistical model adequately describes the dataset at hand, then the sample statistics corresponding to the empirical summaries should somehow reflect corresponding features of the model distribution. We have already seen a mathematical justification for this in Chapter 13 for the sample statistic
$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$
based on a sample $X_1, X_2, \ldots, X_n$ from a probability distribution with expectation μ. According to the law of large numbers,
$$\lim_{n \to \infty} P\big(|\bar{X}_n - \mu| > \varepsilon\big) = 0$$

for every $\varepsilon > 0$. This means that for large sample size n, the sample mean of most realizations of the random sample is close to the expectation of the corresponding distribution. In fact, all sample statistics discussed in Chapters 15 and 16 are close to corresponding distribution features. To illustrate this we generate an artificial dataset from a normal distribution with parameters $\mu = 5$ and $\sigma = 2$, using a technique similar to the one described in Section 6.2. Next, we compare the sample statistics with corresponding features of this distribution.

The empirical distribution function

Let $X_1, X_2, \ldots, X_n$ be a random sample from distribution function F, and let
$$F_n(a) = \frac{\text{number of } X_i \text{ in } (-\infty, a]}{n}$$
be the empirical distribution function of the sample. Another application of the law of large numbers (see Exercise 13.7) yields that for every $\varepsilon > 0$,
$$\lim_{n \to \infty} P(|F_n(a) - F(a)| > \varepsilon) = 0.$$

This means that for most realizations of the random sample the empirical distribution function F_n is close to F:
$$F_n(a) \approx F(a).$$
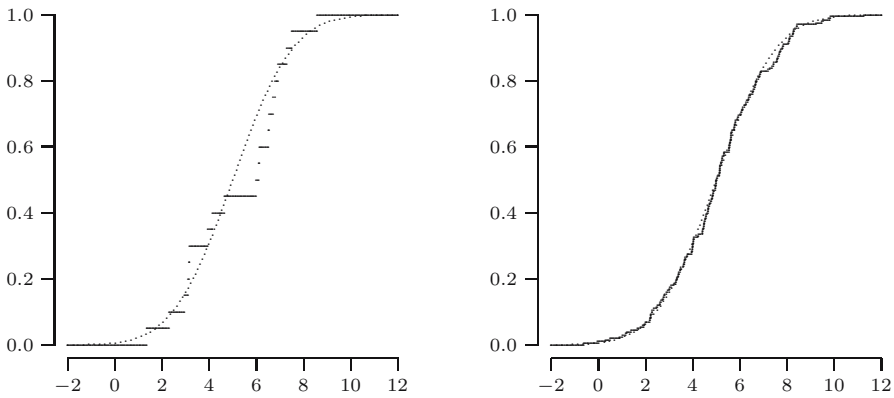
Fig. 17.1. Empirical distribution functions of normal samples.

Hence the empirical distribution function of the normal dataset should resemble the distribution function

$$F(a) = \int_{-\infty}^{a} \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-5}{2}\right)^2} \, dx$$

of the $N(5,4)$ distribution, and the fit should become better as the sample size n increases. An illustration of this can be found in Figure 17.1. We displayed the empirical distribution functions of datasets generated from an $N(5,4)$ distribution together with the "true" distribution function F (dotted lines), for sample sizes $n = 20$ (left) and $n = 200$ (right).

The histogram and the kernel density estimate

Suppose the random sample $X_1, X_2, \ldots, X_n$ is generated from a continuous distribution with probability density f. In Section 13.4 we have seen yet another consequence of the law of large numbers:

$$\frac{\text{number of } X_i \text{ in } (x-h, x+h]}{2hn} \approx f(x).$$
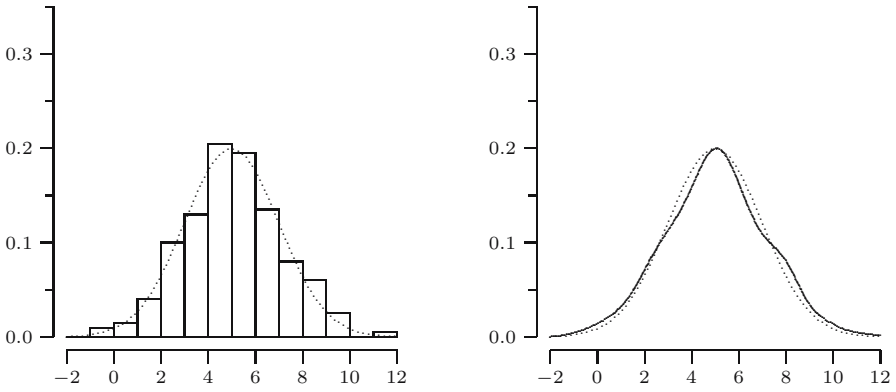
When $(x-h, x+h]$ is a bin of a histogram of the random sample, this means that the height of the histogram approximates the value of f at the midpoint of the bin:

$$\text{height of the histogram on } (x-h, x+h] \approx f(x).$$

Similarly, the kernel density estimate of a random sample approximates the corresponding probability density f:

$$f_{n,h}(x) \approx f(x).$$

Fig. 17.2. Histogram and kernel density estimate of a sample of size 200.

So the histogram and kernel density estimate of the normal dataset should resemble the graph of the probability density

$$f(x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-5}{2}\right)^2}$$

of the $N(5,4)$ distribution. This is illustrated in Figure 17.2, where we displayed a histogram and a kernel density estimate of our dataset consisting of 200 values generated from the $N(5,4)$ distribution. It should be noted that with a smaller dataset the similarity can be much worse. This is demonstrated in Figure 17.3, which is based on the dataset consisting of 20 values generated from the same distribution.
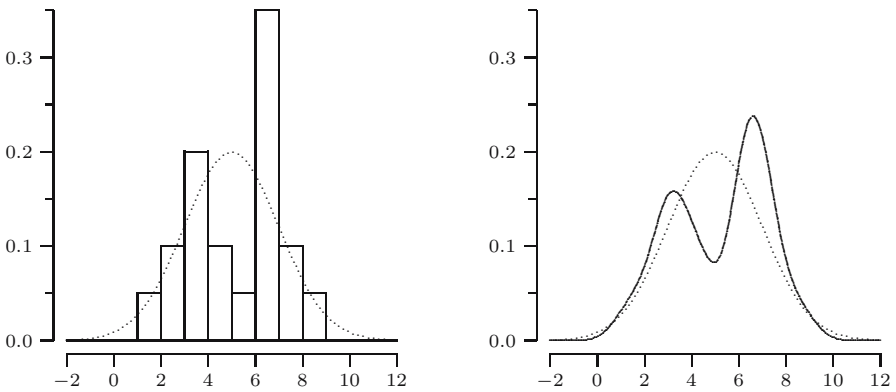
Fig. 17.3. Histogram and kernel density estimate of a sample of size 20.

Remark 17.1 (About the approximations). Let H_n be the height of the histogram on the interval $(x - h, x + h]$, which is assumed to be a bin of the histogram. Direct application of the law of large numbers merely yields that H_n converges to

$$\frac{1}{2h} \int_{x-h}^{x+h} f(u)\, du.$$

Only for small h this is close to $f(x)$. However, if we let h tend to 0 as n increases, a variation on the law of large numbers will guarantee that H_n converges to $f(x)$: for every $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathrm{P}(|H_n - f(x)| > \varepsilon) = 0.$$

A possible choice is the optimal bin width mentioned in Remark 15.1. Similarly, direct application of the law of large numbers yields that a kernel density estimator with fixed bandwidth h converges to

$$\int_{-\infty}^{\infty} f(x + hu)K(u)\, du.$$

Once more, only for small h this is close to $f(x)$, provided that K is symmetric and integrates to one. However, by letting the bandwidth h tend to 0 as n increases, yet another variation on the law of large numbers will guarantee that $f_{n,h}(x)$ converges to $f(x)$: for every $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathrm{P}(|f_{n,h}(x) - f(x)| > \varepsilon) = 0.$$

A possible choice is the optimal bandwidth mentioned in Remark 15.2.

The sample mean, the sample median, and empirical quantiles

As we saw in Section 5.5, the expectation of an $N(\mu, \sigma^2)$ distribution is μ; so the $N(5, 4)$ distribution has expectation 5. According to the law of large numbers: $\bar{X}_n \approx \mu$. This is illustrated by our dataset of 200 values generated from the $N(5, 4)$ distribution for which we find

$$\bar{x}_{200} = 5.012.$$

For the sample median we find

$$\mathrm{Med}(x_1, \ldots, x_{200}) = 5.018.$$

This illustrates the fact that the sample median of a random sample from F approximates the median $q_{0.5} = F^{\mathrm{inv}}(0.5)$. In fact, we have the following general property for the pth empirical quantile:

$$q_n(p) \approx F^{\mathrm{inv}}(p) = q_p.$$

In the special case of the $N(\mu, \sigma^2)$ distribution, the expectation and the median coincide, which explains why the sample mean and sample median of the normal dataset are so close to each other.

The sample variance and standard deviation, and the MAD

As we saw in Section 5.5, the standard deviation and variance of an $N(\mu, \sigma^2)$ distribution are σ and σ^2; so for the $N(5, 4)$ distribution these are 2 and 4. Another consequence of the law of large numbers is that

$$S_n^2 \approx \sigma^2 \quad \text{and} \quad S_n \approx \sigma.$$

This is illustrated by our normal dataset of size 200, for which we find

$$s_{200}^2 = 4.761 \quad \text{and} \quad s_{200} = 2.182$$

for the sample variance and sample standard deviation.

For the MAD of the dataset we find 1.334, which clearly differs from the standard deviation 2 of the $N(5, 4)$ distribution. The reason is that

$$\text{MAD}(X_1, X_2, \ldots, X_n) \approx F^{\text{inv}}(0.75) - F^{\text{inv}}(0.5),$$

for any distribution that is symmetric around its median $F^{\text{inv}}(0.5)$. For the $N(5, 4)$ distribution $F^{\text{inv}}(0.75) - F^{\text{inv}}(0.5) = 2\Phi^{\text{inv}}(0.75) = 1.3490$, where Φ denotes the distribution function of the standard normal distribution (see Exercise 17.10).

Relative frequencies

For continuous distributions the histogram and kernel density estimates of a random sample approximate the corresponding probability density f. For discrete distributions we would like to have a sample statistic that approximates the probability mass function. In Section 13.4 we saw that, as a consequence of the law of large numbers, relative frequencies based on a random sample approximate corresponding probabilities. As a special case, for a random sample $X_1, X_2, \ldots, X_n$ from a discrete distribution with probability mass function p, one has that

$$\frac{\text{number of } X_i \text{ equal to } a}{n} \approx p(a).$$

This means that the relative frequency of a's in the sample approximates the value of the probability mass function at a. Table 17.2 lists the sample statistics and the corresponding distribution features they approximate.

17.3 Estimating features of the "true" distribution

In the previous section we generated a dataset of 200 elements from a probability distribution, and we have seen that certain features of this distribution are approximated by corresponding sample statistics. In practice, the situation is reversed. In that case we have a dataset of n elements that is modeled as the realization of a random sample with a probability distribution that is unknown to us. Our goal is to use our dataset to estimate a certain feature of this distribution that represents the quantity of interest. In this section we will discuss a few examples.

Table 17.2. Some sample statistics and corresponding distribution features.

Sample statistic	Distribution feature
Graphical	
Empirical distribution function F_n	Distribution function F
Kernel density estimate $f_{n,h}$ and histogram	Probability density f
(Number of X_i equal to a)$/n$	Probability mass function $p(a)$
Numerical	
Sample mean $\bar{X}_n$	Expectation μ
Sample median $\text{Med}(X_1, X_2, \ldots, X_n)$	Median $q_{0.5} = F^{\text{inv}}(0.5)$
pth empirical quantile $q_n(p)$	$100p$th percentile $q_p = F^{\text{inv}}(p)$
Sample variance S_n^2	Variance σ^2
Sample standard deviation S_n	Standard deviation σ
$\text{MAD}(X_1, X_2, \ldots, X_n)$	$F^{\text{inv}}(0.75) - F^{\text{inv}}(0.5)$, for symmetric F

The Old Faithful data

We stick to the assumptions of Section 17.1: by lack of knowledge on this phenomenon we prefer not to specify a particular parametric type of distribution, and we model the Old Faithful data as the realization of a random sample of size 272 from a continuous probability distribution. From the previous section we know that the kernel density estimate and the empirical distribution function of the dataset approximate the probability density f and the distribution function F of this distribution. In Figure 17.4 a kernel density estimate (left) and the empirical distribution function (right) are displayed. Indeed, neither graph resembles the probability density function or distribution function of any of the familiar parametric distributions. Instead of viewing both graphs
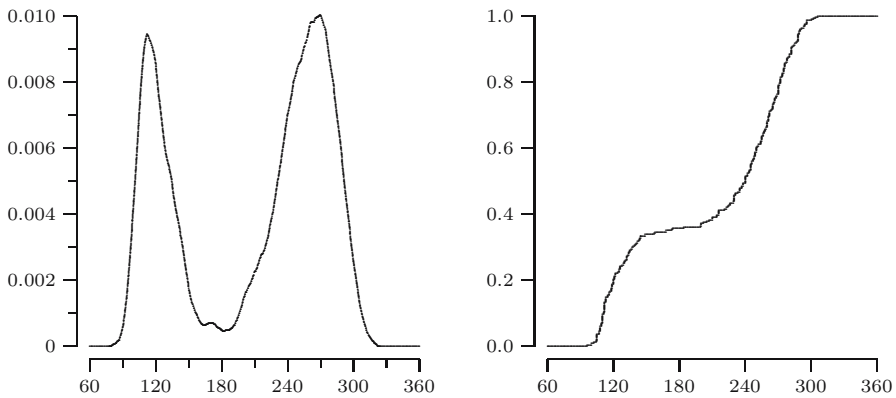

Fig. 17.4. Nonparametric estimates for f and F based on the Old Faithful data.

only as graphical summaries of the data, we can also use both curves as estimates for f and F. We estimate the model probability density f by means of the kernel density estimate and the model distribution function F by means of the empirical distribution function. Since neither estimate assumes a particular parametric model, they are called *nonparametric* estimates.
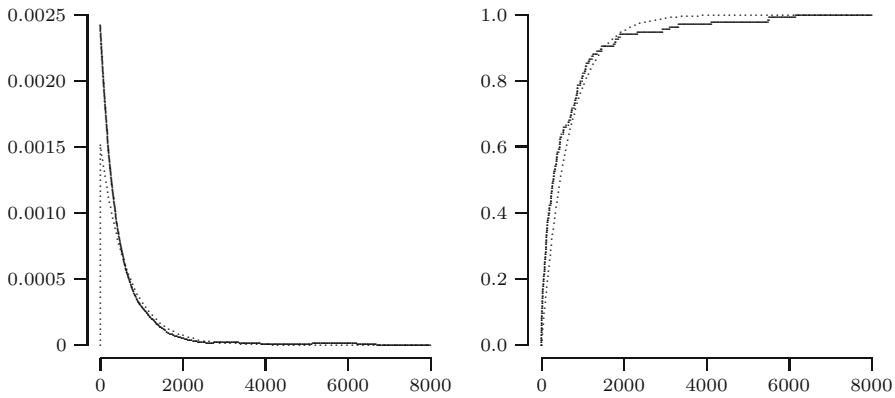
The software data

Next consider the software reliability data. As motivated in Section 17.1, we model interfailure times as the realization of a random sample from an exponential distribution. To see whether an exponential distribution is indeed a reasonable model, we plot a histogram and a kernel density estimate using a boundary kernel in Figure 17.5.
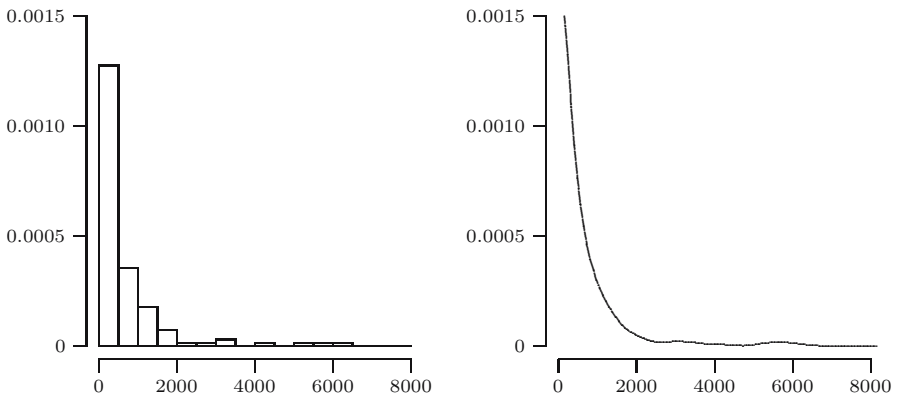
Fig. 17.5. Histogram and kernel density estimate for the software data.

Both seem to corroborate the assumption of an exponential distribution. Accepting this, we are left with estimating the parameter λ. Because for the exponential distribution $E[X] = 1/\lambda$, the law of large numbers suggests $1/\bar{x}$ as an estimate for λ. For our dataset $\bar{x} = 656.88$, which yields $1/\bar{x} = 0.0015$. In Figure 17.6 we compare the estimated exponential density (left) and distribution function (right) with the corresponding nonparametric estimates. Note that the nonparametric estimates do *not* assume an exponential model for the data. But, *if* an exponential distribution were the right model, the kernel density estimate and empirical distribution function should resemble the estimated exponential density and distribution function. At first sight the fit seems reasonable, although near zero the data accumulate more than one might perhaps expect for a sample of size 135 from an exponential distribution, and the other way around at the other end of the data range. The question is whether this phenomenon can be attributed to chance or is caused by the fact that the exponential model is the wrong model. We will return to this type of question in Chapter 25 (see also Chapter 18).

Fig. 17.6. Kernel density estimate and empirical cdf for software data (solid) compared to f and F of the estimated exponential distribution.

Michelson data

Consider the Michelson data on the speed of light. In this case we are not particularly interested in estimation of the "true" distribution, but solely in the expectation of this distribution, which represents the speed of light. The law of large numbers suggests to estimate the expectation by the sample mean $\bar{x}$, which equals 852.4.

17.4 The linear regression model

Recall the example about predicting Janka hardness of wood from the density of the wood in Section 15.5. The idea is, of course, that Janka hardness is related to the density: the higher the density of the wood, the higher the value of Janka hardness. This suggests a relationship of the type

$$\text{hardness} = g(\text{density of timber})$$

for some increasing function g. This is supported by the scatterplot of the data in Figure 17.7. A closer look at the bivariate dataset in Table 15.5 suggests that randomness is also involved. For instance, for the value 51.5 of the density, different corresponding values of Janka hardness were observed. One way to model such a situation is by means of a *regression model*:

$$\text{hardness} = g(\text{density of timber}) + \text{random fluctuation}.$$

The important question now is *what sort of function g* fits well to the points in the scatterplot?

In general, this may be a difficult question to answer. We may have so little knowledge about the phenomenon under study, and the data points may be
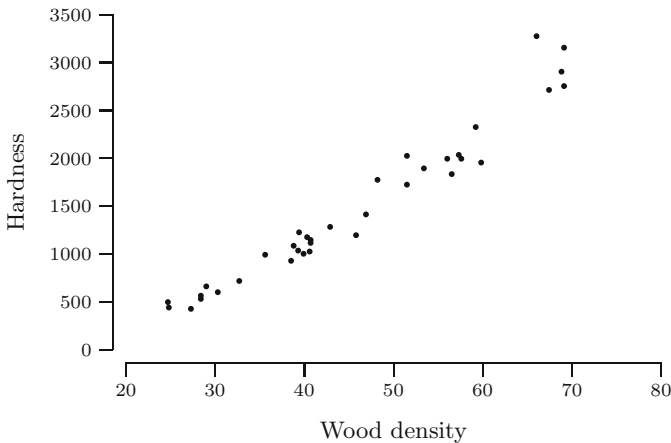
Fig. 17.7. Scatterplot of Janka hardness versus wood density.

scattered in such a way, that there is no reason to assume a specific type of function for g. However, for the Janka hardness data it makes sense to assume that g is increasing, but this still leaves us with many possibilities. Looking at the scatterplot, at first sight it does not seem unreasonable to assume that g is a straight line, i.e., Janka hardness depends linearly on the density of timber. The fact that the points are not exactly on a straight line is then modeled by a random fluctuation with respect to the straight line:

$$\text{hardness} = \alpha + \beta \cdot (\text{density of timber}) + \text{random fluctuation}.$$

This is a loose description of a simple linear regression model. A more complete description is given below.

> SIMPLE LINEAR REGRESSION MODEL. In a *simple linear regression model* for a bivariate dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we assume that $x_1, x_2, \ldots, x_n$ are nonrandom and that $y_1, y_2, \ldots, y_n$ are realizations of random variables $Y_1, Y_2, \ldots, Y_n$ satisfying
>
> $$Y_i = \alpha + \beta x_i + U_i \quad \text{for } i = 1, 2, \ldots, n,$$
>
> where $U_1, \ldots, U_n$ are *independent* random variables with $\mathrm{E}[U_i] = 0$ and $\mathrm{Var}(U_i) = \sigma^2$.

The line $y = \alpha + \beta x$ is called the *regression line*. The parameters α and β represent the *intercept* and *slope* of the regression line. Usually, the x-variable is called the *explanatory variable* and the y-variable is called the *response variable*. One also refers to x and y as *independent* and *dependent* variables. The random variables $U_1, U_2, \ldots, U_n$ are assumed to be independent when the different measurements do not influence each other. They are assumed to have

expectation zero, because the random fluctuation is considered to be around the regression line $y = \alpha + \beta x$. Finally, because each random fluctuation is supposed to have the same amount of variability, we assume that all U_i have the same variance. Note that by the propagation of independence rule in Section 9.4, independence of the U_i implies independence of Y_i. However, $Y_1, Y_2, \ldots, Y_n$ *do not* form a random sample. Indeed, the Y_i have different distributions because every Y_i has a different expectation

$$\mathrm{E}\,[Y_i] = \mathrm{E}\,[\alpha + \beta x_i + U_i] = \alpha + \beta x_i + \mathrm{E}\,[U_i] = \alpha + \beta x_i.$$

QUICK EXERCISE 17.3 Consider the simple linear regression model as defined earlier. Compute the variance of Y_i.

The parameters α and β are unknown and our task will be to estimate them on the basis of the data. We will come back to this in Chapter 22. In Figure 17.8 the scatterplot for the Janka hardness data is displayed with the estimated
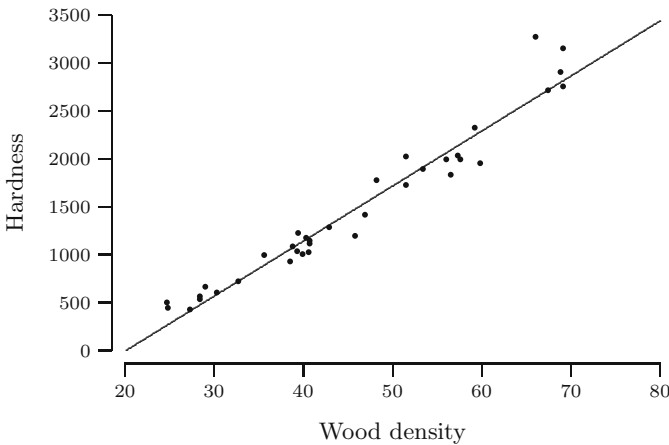
Fig. 17.8. Estimated regression line for the Janka hardness data.

regression line

$$y = -1160.5 + 57.51x.$$

Taking a closer look at Figure 17.8, you might wonder whether

$$y = \alpha + \beta x + \gamma x^2$$

would be a more appropriate model. By trying to answer this question we enter the area of *multiple* linear regression. We will not pursue this topic; we restrict ourselves to *simple* linear regression.

17.5 Solutions to the quick exercises

17.1 Because X_1, X_2 form a random sample, they are independent. Using the rule about the variance of the sum of independent random variables, this means that $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = 1 + 1 = 2$.

17.2 The result of each toss of a coin can be modeled by a Bernoulli random variable taking values 1 (heads) and 0 (tails). In the case when it is known that we are tossing a *fair* coin, heads and tails occur with equal probability. Since it is reasonable to assume that the tosses do not influence each other, the outcomes of the ten tosses are modeled as the realization of a random sample $X_1, \ldots, X_{10}$ from a Bernoulli distribution with parameter $p = 1/2$. In this case the model distribution is completely specified and coincides with the "true" distribution: a $Ber(\frac{1}{2})$ distribution.

In the case when we are dealing with a *possibly unfair* coin, the outcomes of the ten tosses are still modeled as the realization of a random sample $X_1, \ldots, X_{10}$ from a Bernoulli distribution, but we cannot specify the value of the parameter p. The model distribution is a Bernoulli distribution. The "true" distribution is a Bernoulli distribution with one particular value for p, unknown to us.

17.3 Note that the x_i are considered nonrandom. By the rules for the variance, we find $\text{Var}(Y_i) = \text{Var}(\alpha + \beta x_i + U_i) = \text{Var}(U_i) = \sigma^2$.

17.6 Exercises

17.1 ⊡ Figure 17.9 displays several histograms, kernel density estimates, and empirical distribution functions. It is known that all figures correspond to datasets of size 200 that are generated from normal distributions $N(0, 1)$, $N(0, 9)$, and $N(3, 1)$, and from exponential distributions $Exp(1)$ and $Exp(1/3)$. Report for each figure from which distribution the dataset has been generated.

17.2 ⊡ Figure 17.10 displays several boxplots. It is known that all figures correspond to datasets of size 200 that are generated from the same five distributions as in Exercise 17.1. Report for each boxplot from which distribution the dataset has been generated.

17.3 ⊞ At a London underground station, the number of women was counted in each of 100 queues of length 10. In this way a dataset $x_1, x_2, \ldots, x_{100}$ was obtained, where x_i denotes the observed number of women in the ith queue. The dataset is summarized in the following table and lists the number of queues with 0 women, 1 woman, 2 women, etc.
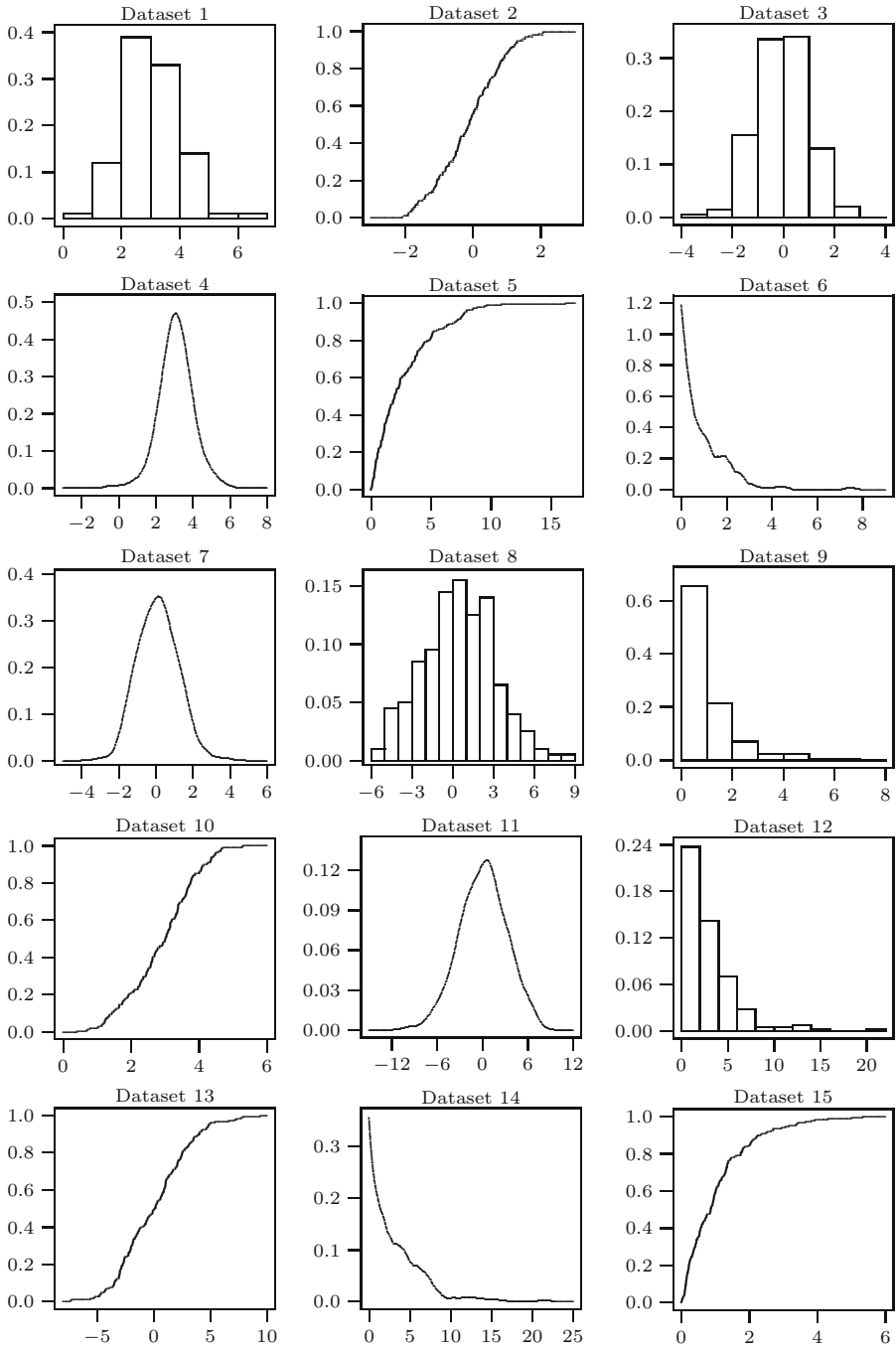
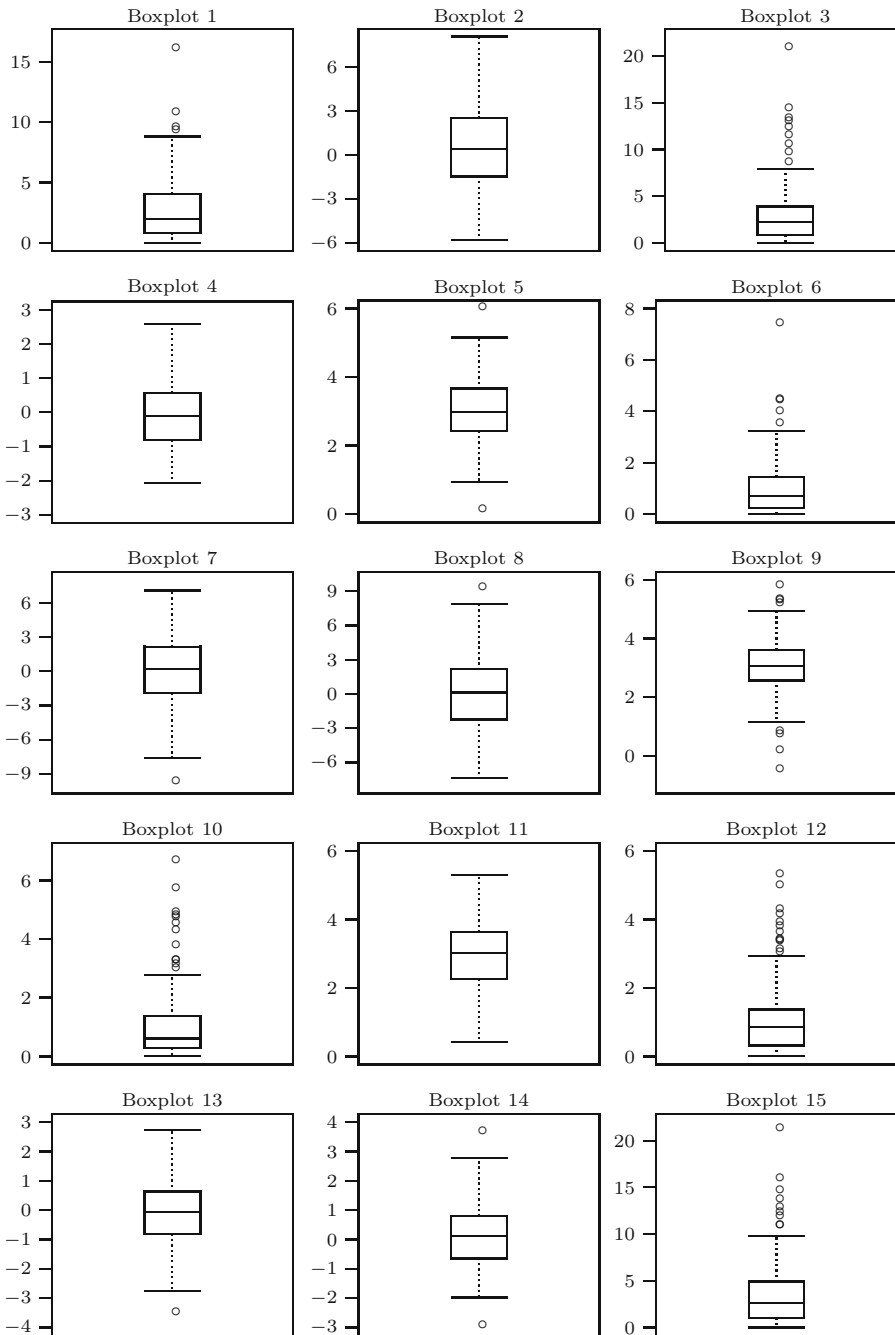**Fig. 17.9.** Graphical representations of different datasets from Exercise 17.1.

Fig. 17.10. Boxplot of different datasets from Exercise 17.2.

Count	0	1	2	3	4	5	6	7	8	9	10
Frequency	1	3	4	23	25	19	18	5	1	1	0

Source: R.A. Jinkinson and M. Slater. Critical discussion of a graphical method for identifying discrete distributions. *The Statistician*, 30:239–248, 1981; Table 1 on page 240.

In the statistical model for this dataset, we assume that the observed counts are a realization of a random sample $X_1, X_2, \ldots, X_{100}$.

 a. Assume that people line up in such a way that a man or woman in a certain position is independent of the other positions, and that in each position one has a woman with equal probability. What is an appropriate choice for the model distribution?

 b. Use the table to find an estimate for the parameter(s) of the model distribution chosen in part **a**.

17.4 During the Second World War, London was hit by numerous flying bombs. The following data are from an area in South London of 36 square kilometers. The area was divided into 576 squares with sides of length 1/4 kilometer. For each of the 576 squares the number of hits was recorded. In this way we obtain a dataset $x_1, x_2, \ldots, x_{576}$, where x_i denotes the number of hits in the ith square. The data are summarized in the following table which lists the number of squares with no hits, 1 hit, 2 hits, etc.

Number of hits	0	1	2	3	4	5	6	7
Number of squares	229	211	93	35	7	0	0	1

Source: R.D. Clarke. An application of the Poisson distribution. *Journal of the Institute of Actuaries*, 72:48, 1946; Table 1 on page 481. © Faculty and Institute of Actuaries.

An interesting question is whether London was hit in a completely random manner. In that case a Poisson distribution should fit the data.

 a. If we model the dataset as the realization of a random sample from a Poisson distribution with parameter μ, then what would you choose as an estimate for μ?

 b. Check the fit with a Poisson distribution by comparing some of the observed relative frequencies of 0's, 1's, 2's, etc., with the corresponding probabilities for the Poisson distribution with μ estimated as in part **a**.

17.5 ⊡ We return to the example concerning the number of menstrual cycles up to pregnancy, where the number of cycles was modeled by a geometric random variable (see Section 4.4). The original data concerned 100 smoking and 486 nonsmoking women. For 7 smokers and 12 nonsmokers, the exact number of cycles up to pregnancy was unknown. In the following tables we only

incorporated the 93 smokers and 474 nonsmokers, for which the exact number of cycles was observed. Another analysis, based on the complete dataset, is done in Section 21.1.

a. Consider the dataset $x_1, x_2, \ldots, x_{93}$ corresponding to the smoking women, where x_i denotes the number of cycles for the ith smoking woman. The data are summarized in the following table.

Cycles	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	29	16	17	4	3	9	4	5	1	1	1	3

Source: C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

The table lists the number of women that had to wait 1 cycle, 2 cycles, etc. If we model the dataset as the realization of a random sample from a geometric distribution with parameter p, then what would you choose as an estimate for p?

b. Also estimate the parameter p for the 474 nonsmoking women, which is also modeled as the realization of a random sample from a geometric distribution. The dataset $y_1, y_2, \ldots, y_{474}$, where y_j denotes the number of cycles for the jth nonsmoking woman, is summarized here:

Cycles	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	198	107	55	38	18	22	7	9	5	3	6	6

Source: C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

You may use that $y_1 + y_2 + \cdots + y_{474} = 1285$.

c. Compare the estimates of the probability of becoming pregnant in three or fewer cycles for smoking and nonsmoking women.

17.6 Recall Exercise 15.1 about the chest circumference of 5732 Scottish soldiers, where we constructed the histogram displayed in Figure 17.11. The histogram suggests modeling the data as the realization of a random sample from a normal distribution.

a. Suppose that for the dataset $\sum x_i = 228377.2$ and $\sum x_i^2 = 9124064$. What would you choose as estimates for the parameters μ and σ of the $N(\mu, \sigma^2)$ distribution?
Hint: you may want to use the relation from Exercise 16.15.

b. Give an estimate for the probability that a Scottish soldier has a chest circumference between 38.5 and 42.5 inches.
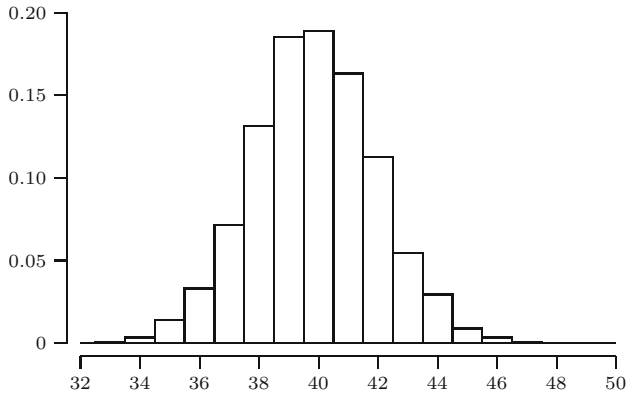
Fig. 17.11. Histogram of chest circumferences.

17.7 ⊞ Recall Exercise 15.3 about time intervals between successive coal mine disasters. Let us assume that the rate at which the disasters occur is constant over time and that on a single day a disaster takes place with small probability independently of what happens on other days. According to Chapter 12 this suggests modeling the series of disasters with a Poisson process. Figure 17.12 displays a histogram and empirical distribution function of the observed time intervals.

a. In the statistical model for this dataset we model the 190 time intervals as the realization of a random sample. What would you choose for the model distribution?

b. The sum of the observed time intervals is 40 549 days. Give an estimate for the parameter(s) of the distribution chosen in part **a**.
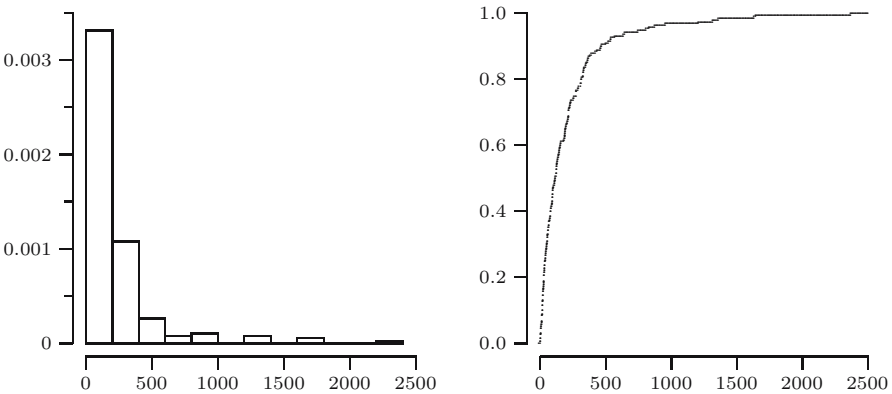
Fig. 17.12. Histogram of time intervals between successive disasters.

17.8 The following data represent the number of revolutions to failure (in millions) of 22 deep-groove ball-bearings.

17.88	28.92	33.00	41.52	42.12
45.60	48.48	51.84	51.96	54.12
55.56	67.80	68.64	68.88	84.12
93.12	98.64	105.12	105.84	127.92
128.04	173.40			

Source: J. Lieblein and M. Zelen. Statistical investigation of the fatigue-life of deep-groove ball-bearings. *Journal of Research, National Bureau of Standards*, 57:273–316, 1956; specimen worksheet on page 286.

Lieblein and Zelen propose modeling the dataset as a realization of a random sample from a Weibull distribution, which has distribution function

$$F(x) = 1 - e^{-(\lambda x)^{\alpha}} \quad \text{for } x \geq 0,$$

and $F(x) = 0$, for $x < 0$, where $\alpha, \lambda > 0$.

a. Suppose that X is a random variable with a Weibull distribution. Check that the random variable $Y = X^{\alpha}$ has an exponential distribution with parameter λ^{α} and conclude that $E[X^{\alpha}] = 1/\lambda^{\alpha}$.

b. Use part **a** to explain how one can use the data in the table to find an estimate for the parameter λ, if it is given that the parameter α is estimated by 2.102.

17.9 ⊞ The volume (i.e., the effective wood production in cubic meters), height (in meters), and diameter (in meters) (measured at 1.37 meter above the ground) are recorded for 31 black cherry trees in the Allegheny National Forest in Pennsylvania. The data are listed in Table 17.3. They were collected to find an estimate for the volume of a tree (and therefore for the timber yield), given its height and diameter. For each tree the volume y and the value of $x = d^2h$ are recorded, where d and h are the diameter and height of the tree. The resulting points $(x_1, y_1), \ldots, (x_{31}, y_{31})$ are displayed in the scatterplot in Figure 17.13.

We model the data by the following linear regression model (without intercept)

$$Y_i = \beta x_i + U_i$$

for $i = 1, 2, \ldots, 31$.

a. What physical reasons justify the linear relationship between y and d^2h? *Hint:* how does the volume of a cylinder relate to its diameter and height?

b. We want to find an estimate for the slope β of the line $y = \beta x$. Two natural candidates are the average slope $\bar{z}_n$, where $z_i = y_i/x_i$, and the

Table 17.3. Measurements on black cherry trees.

Diameter	Height	Volume
0.21	21.3	0.29
0.22	19.8	0.29
0.22	19.2	0.29
0.27	21.9	0.46
0.27	24.7	0.53
0.27	25.3	0.56
0.28	20.1	0.44
0.28	22.9	0.52
0.28	24.4	0.64
0.28	22.9	0.56
0.29	24.1	0.69
0.29	23.2	0.59
0.29	23.2	0.61
0.30	21.0	0.60
0.30	22.9	0.54
0.33	22.6	0.63
0.33	25.9	0.96
0.34	26.2	0.78
0.35	21.6	0.73
0.35	19.5	0.71
0.36	23.8	0.98
0.36	24.4	0.90
0.37	22.6	1.03
0.41	21.9	1.08
0.41	23.5	1.21
0.44	24.7	1.57
0.44	25.0	1.58
0.45	24.4	1.65
0.46	24.4	1.46
0.46	24.4	1.44
0.52	26.5	2.18

slope of the averages $\bar{y}/\bar{x}$. In Chapter 22 we will encounter the so-called least squares estimate:

$$\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$
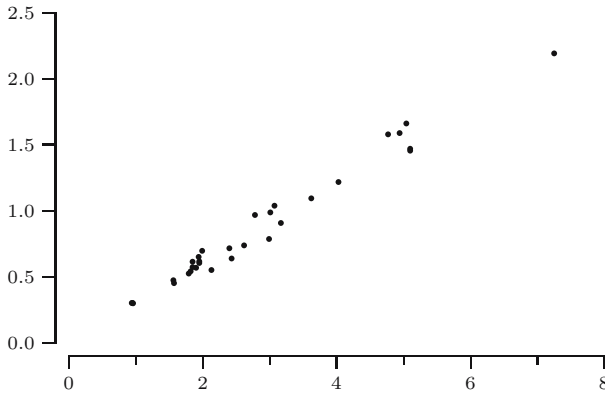
Fig. 17.13. Scatterplot of the black cherry tree data.

Compute all three estimates for the data in Table 17.3. You need at least 5 digits accuracy, and you may use that $\sum x_i = 87.456$, $\sum y_i = 26.486$, $\sum y_i/x_i = 9.369$, $\sum x_i y_i = 95.498$, and $\sum x_i^2 = 314.644$.

17.10 Let X be a random variable with (continuous) distribution function F. Let $m = q_{0.5} = F^{\mathrm{inv}}(0.5)$ be the median of F and define the random variable

$$Y = |X - m|.$$

a. Show that Y has distribution function G, defined by

$$G(y) = F(m + y) - F(m - y).$$

b. The MAD of F is the median of G. Show that if the density f corresponding to F is symmetric around its median m, then

$$G(y) = 2F(m + y) - 1$$

and derive that

$$G^{\mathrm{inv}}\left(\tfrac{1}{2}\right) = F^{\mathrm{inv}}\left(\tfrac{3}{4}\right) - F^{\mathrm{inv}}\left(\tfrac{1}{2}\right).$$

c. Use **b** to conclude that the MAD of an $N(\mu, \sigma^2)$ distribution is equal to $\sigma\Phi^{\mathrm{inv}}(3/4)$, where Φ is the distribution function of a standard normal distribution. Recall that the distribution function F of an $N(\mu, \sigma^2)$ can be written as

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

You might check that, as stated in Section 17.2, the MAD of the $N(5, 4)$ distribution is equal to $2\Phi^{\mathrm{inv}}(3/4) = 1.3490$.

17.11 In this exercise we compute the MAD of the $Exp(\lambda)$ distribution.

a. Let X have an $Exp(\lambda)$ distribution, with median $m = (\ln 2)/\lambda$. Show that $Y = |X - m|$ has distribution function

$$G(y) = \frac{1}{2} \left(e^{\lambda y} - e^{-\lambda y} \right).$$

b. Argue that the MAD of the $Exp(\lambda)$ distribution is a solution of the equation $e^{2\lambda y} - e^{\lambda y} - 1 = 0$.

c. Compute the MAD of the $Exp(\lambda)$ distribution.
Hint: put $x = e^{\lambda y}$ and first solve for x.

18
The bootstrap

In the forthcoming chapters we will develop statistical methods to infer knowledge about the model distribution and encounter several sample statistics to do this. In the previous chapter we have seen examples of sample statistics that can be used to estimate different model features, for instance, the empirical distribution function to estimate the model distribution function F, and the sample mean to estimate the expectation μ corresponding to F. One of the things we would like to know is how close a sample statistic is to the model feature it is supposed to estimate. For instance, what is the probability that the sample mean and μ differ more than a given tolerance ε? For this we need to know the distribution of $\bar{X}_n - \mu$. More generally, it is important to know how a sample statistic is distributed in relation to the corresponding model feature. For the distribution of the sample mean we saw a normal *limit* approximation in Chapter 14. In this chapter we discuss a simulation procedure that approximates the distribution of the sample mean for *finite* sample size. Moreover, the method is more generally applicable to sample statistics other than the sample mean.

18.1 The bootstrap principle

Consider the Old Faithful data introduced in Chapter 15, which we modeled as the realization of a random sample of size $n = 272$ from some distribution function F. The sample mean $\bar{x}_n$ of the observed durations equals 209.3. What does this say about the expectation μ of F? As we saw in Chapter 17, the value 209.3 is a natural estimate for μ, but to conclude that μ is *equal* to 209.3 is unwise. The reason is that, if we would observe a new dataset of durations, we will obtain a different sample mean as an estimate for μ. This should not come as a surprise. Since the dataset $x_1, x_2, \ldots, x_n$ is just one possible realization of the random sample $X_1, X_2, \ldots, X_n$, the observed sample mean is just one possible realization of the random variable

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

A new dataset is another realization of the random sample, and the corresponding sample mean is another realization of the random variable $\bar{X}_n$. Hence, to infer something about μ, one should take into account how realizations of $\bar{X}_n$ vary. This variation is described by the probability distribution of $\bar{X}_n$.

In principle[1] it is possible to determine the distribution function of $\bar{X}_n$ from the distribution function F of the random sample $X_1, X_2, \ldots, X_n$. However, F is *unknown*. Nevertheless, in Chapter 17 we saw that the observed dataset reflects most features of the "true" probability distribution. Hence the natural thing to do is to compute an *estimate* $\hat{F}$ for the distribution function F and then to consider a random sample from $\hat{F}$ and the corresponding sample mean as substitutes for the random sample $X_1, X_2, \ldots, X_n$ from F and the random variable $\bar{X}_n$. A random sample from $\hat{F}$ is called a *bootstrap random sample*, or briefly bootstrap sample, and is denoted by

$$X_1^*, X_2^*, \ldots, X_n^*$$

to distinguish it from the random sample $X_1, X_2, \ldots, X_n$ from the "true" F. The corresponding average is called the *bootstrapped sample mean*, and this random variable is denoted by

$$\bar{X}_n^* = \frac{X_1^* + X_2^* + \cdots + X_n^*}{n}$$

to distinguish it from the random variable $\bar{X}_n$. The idea is now to use the distribution of $\bar{X}_n^*$ to approximate the distribution of $\bar{X}_n$.

The preceding procedure is called the *bootstrap principle* for the sample mean. Clearly, it can be applied to *any* sample statistic $h(X_1, X_2, \ldots, X_n)$ by approximating its probability distribution by that of the corresponding bootstrapped sample statistic $h(X_1^*, X_2^*, \ldots, X_n^*)$.

> BOOTSTRAP PRINCIPLE. Use the dataset $x_1, x_2, \ldots, x_n$ to compute an estimate $\hat{F}$ for the "true" distribution function F. Replace the random sample $X_1, X_2, \ldots, X_n$ from F by a random sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{F}$, and approximate the probability distribution of $h(X_1, X_2, \ldots, X_n)$ by that of $h(X_1^*, X_2^*, \ldots, X_n^*)$.

Returning to the sample mean, the first question that comes to mind is, of course, how well does the distribution of $\bar{X}_n^*$ approximate the distribution

---

[1] In Section 11.1 we saw how the distribution of the sum of independent random variables can be computed. Together with the change-of-units rule (see page 106), the distribution of $\bar{X}_n$ can be determined. See also Section 13.1, where this is done for independent $Gam(2, 1)$ variables.

of $\bar{X}_n$? Or more generally, how well does the distribution of a bootstrapped sample statistic $h(X_1^*, X_2^*, \ldots, X_n^*)$ approximate the distribution of the sample statistic of interest $h(X_1, X_2, \ldots, X_n)$? Applied in such a straightforward manner, the bootstrap approximation for the distribution of $\bar{X}_n$ by that of $\bar{X}_n^*$ may not be so good (see Remark 18.1). The bootstrap approximation will improve if we approximate the distribution of the *centered* sample mean:

$$\bar{X}_n - \mu,$$

where μ is the expectation corresponding to F. The bootstrapped version would be the random variable

$$\bar{X}_n^* - \mu^*,$$

where μ^* is the expectation corresponding to $\hat{F}$. Often the bootstrap approximation of the distribution of a sample statistic will improve if we somehow normalize the sample statistic by relating it to a corresponding feature of the "true" distribution. An example is the centered sample median

$$\mathrm{Med}(X_1, X_2, \ldots, X_n) - F^{\mathrm{inv}}(0.5),$$

where we subtract the median $F^{\mathrm{inv}}(0.5)$ of F. Another example is the normalized sample variance

$$\frac{S_n^2}{\sigma^2},$$

where we divide by the variance σ^2 of F.

QUICK EXERCISE 18.1 Describe how the bootstrap principle should be applied to approximate the distribution of $\mathrm{Med}(X_1, X_2, \ldots, X_n) - F^{\mathrm{inv}}(0.5)$.

Remark 18.1 (The bootstrap for the sample mean). To see why the bootstrap approximation for $\bar{X}_n$ may be bad, consider a dataset $x_1, x_2, \ldots, x_n$ that is a realization of a random sample $X_1, X_2, \ldots, X_n$ from an $N(\mu, 1)$ distribution. In that case the corresponding sample mean $\bar{X}_n$ has an $N(\mu, 1/n)$ distribution. We estimate μ by $\bar{x}_n$ and replace the random sample from an $N(\mu, 1)$ distribution by a bootstrap random sample $X_1^*, X_2^*, \ldots, X_n^*$ from an $N(\bar{x}_n, 1)$ distribution. The corresponding bootstrapped sample mean $\bar{X}_n^*$ has an $N(\bar{x}_n, 1/n)$ distribution. Therefore the distribution functions G_n and G_n^* of the random variables $\bar{X}_n$ and $\bar{X}_n^*$ can be determined:

$$G_n(a) = \Phi(\sqrt{n}(a - \mu)) \quad \text{and} \quad G_n^*(a) = \Phi(\sqrt{n}(a - \bar{x}_n)).$$

In this case it turns out that the maximum distance between the two distribution functions is equal to

$$2\Phi\left(\tfrac{1}{2}\sqrt{n}|\bar{x}_n - \mu|\right) - 1.$$

Since $\bar{X}_n$ has an $N(\mu, 1/n)$ distribution, this value is approximately equal to $2\Phi\left(|z|/2\right) - 1$, where z is a realization of an $N(0, 1)$ random variable Z. This only equals zero for $z = 0$, so that the distance between the distribution functions of $\bar{X}_n$ and $\bar{X}_n^*$ will almost always be strictly positive, even for large n.

The question that remains is what to take as an estimate $\hat{F}$ for F. This will depend on how well F can be specified. For the Old Faithful data we cannot say anything about the type of distribution. However, for the software data it seems reasonable to model the dataset as a realization of a random sample from an $Exp(\lambda)$ distribution and then we only have to estimate the parameter λ. Different assumptions about F give rise to different bootstrap procedures. We will discuss two of them in the next sections.

18.2 The empirical bootstrap

Suppose we consider our dataset $x_1, x_2, \ldots, x_n$ as a realization of a random sample from a distribution function F. When we cannot make any assumptions about the type of F, we can always estimate F by the empirical distribution function of the dataset:

$$\hat{F}(a) = F_n(a) = \frac{\text{number of } x_i \text{ less than or equal to } a}{n}.$$

Since we estimate F by the empirical distribution function, the corresponding bootstrap principle is called the *empirical bootstrap*. Applying this principle to the centered sample mean, the random sample $X_1, X_2, \ldots, X_n$ from F is replaced by a bootstrap random sample $X_1^*, X_2^*, \ldots, X_n^*$ from F_n, and the distribution of $\bar{X}_n - \mu$ is approximated by that of $\bar{X}_n^* - \mu^*$, where μ^* denotes the expectation corresponding to F_n. The question is, of course, how good this approximation is. A mathematical theorem tells us that the empirical bootstrap works for the centered sample mean, i.e., the distribution of $\bar{X}_n - \mu$ is well approximated by that of $\bar{X}_n^* - \mu^*$ (see Remark 18.2). On the other hand, there are (normalized) sample statistics for which the empirical bootstrap fails, such as

$$1 - \frac{\text{maximum of } X_1, X_2, \ldots, X_n}{\theta},$$

based on a random sample $X_1, X_2, \ldots, X_n$ from a $U(0, \theta)$ distribution (see Exercise 18.12).

Remark 18.2 (The empirical bootstrap for $\bar{X}_n - \mu$). For the centered sample mean the bootstrap approximation works, even if we estimate F by the empirical distribution function F_n. If G_n denotes the distribution function of $\bar{X}_n - \mu$ and G_n^* the distribution function of its bootstrapped version $\bar{X}_n^* - \mu^*$, then the maximum distance between G_n^* and G_n goes to zero with probability one:

$$\mathrm{P}\left(\lim_{n\to\infty}\sup_{t\in\mathbb{R}}|G_n^*(t) - G_n(t)| = 0\right) = 1$$

(see, for instance, Singh [32]). In fact, the empirical bootstrap approximation can be improved by approximating the distribution of the standardized average $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ by its bootstrapped version $\sqrt{n}(\bar{X}_n^* - \mu^*)/\sigma^*$, where σ and σ^* denote the standard deviations of F and F_n. This approximation is even better than the normal approximation by the central limit theorem! See, for instance, Hall [14].

Let us continue with approximating the distribution of $\bar{X}_n - \mu$ by that of $\bar{X}_n^* - \mu^*$. First note that the empirical distribution function F_n of the original dataset is the distribution function of a discrete random variable that attains the values $x_1, x_2, \ldots, x_n$, each with probability $1/n$. This means that each of the bootstrap random variables X_i^* has expectation

$$\mu^* = \mathrm{E}[X_i^*] = x_1 \cdot \frac{1}{n} + x_2 \cdot \frac{1}{n} + \cdots + x_n \cdot \frac{1}{n} = \bar{x}_n.$$

Therefore, applying the empirical bootstrap to $\bar{X}_n - \mu$ means approximating its distribution by that of $\bar{X}_n^* - \bar{x}_n$. In principle it would be possible to determine the probability distribution of $\bar{X}_n^* - \bar{x}_n$. Indeed, the random variable $\bar{X}_n^*$ is based on the random variables X_i^*, whose distribution we know precisely: it takes values $x_1, x_2, \ldots, x_n$ with equal probability $1/n$. Hence we could determine the possible values of $\bar{X}_n^* - \bar{x}_n$ and the corresponding probabilities. For small n this can be done (see Exercise 18.5), but for large n this becomes cumbersome. Therefore we invoke a second approximation.

Recall the jury example in Section 6.3, where we investigated the variation of two different rules that a jury might use to assign grades. In terms of the present chapter, the jury example deals with a random sample from a $U(-0.5, 0.5)$ distribution and two different sample statistics T and M, corresponding to the two rules. To investigate the distribution of T and M, a simulation was carried out with one thousand runs, where in every run we generated a realization of a random sample from the $U(-0.5, 0.5)$ distribution and computed the corresponding realization of T and M. The one thousand realizations give a good impression of how T and M vary around the deserved score (see Figure 6.4).

Returning to the distribution of $\bar{X}_n^* - \bar{x}_n$, the analogue would be to repeatedly generate a realization of the bootstrap random sample from F_n and every time compute the corresponding realization of $\bar{X}_n^* - \bar{x}_n$. The resulting realizations would give a good impression about the distribution of $\bar{X}_n^* - \bar{x}_n$. A realization of the bootstrap random sample is called a *bootstrap dataset* and is denoted by

$$x_1^*, x_2^*, \ldots, x_n^*$$

to distinguish it from the original dataset $x_1, x_2, \ldots, x_n$. For the centered sample mean the simulation procedure is as follows.

EMPIRICAL BOOTSTRAP SIMULATION (FOR $\bar{X}_n - \mu$). Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F, and compute the expectation

$$\mu^* = \bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

corresponding to F_n.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - \bar{x}_n,$$

where

$$\bar{x}_n^* = \frac{x_1^* + x_2^* + \cdots + x_n^*}{n}.$$

Repeat steps 1 and 2 many times.

Note that generating a value x_i^* from F_n is equivalent to choosing one of the elements $x_1, x_2, \ldots, x_n$ of the original dataset with equal probability $1/n$.

The empirical bootstrap simulation is described for the centered sample mean, but clearly a similar simulation procedure can be formulated for any (normalized) sample statistic.

> **Remark 18.3 (Some history).** Although Efron [7] in 1979 drew attention to diverse applications of the empirical bootstrap simulation, it already existed before that time, but not as a unified widely applicable technique. See Hall [14] for references to earlier ideas along similar lines and to further development of the bootstrap. One of Efron's contributions was to point out how to combine the bootstrap with modern computational power. In this way, the interest in this procedure is a typical consequence of the influence of computers on the development of statistics in the past decades. Efron also coined the term "bootstrap," which is inspired by the American version of one of the tall stories of the Baron von Münchhausen, who claimed to have lifted himself out of a swamp by pulling the strap on his boot (in the European version he lifted himself by pulling his hair).

QUICK EXERCISE 18.2 Describe the empirical bootstrap simulation for the centered sample median $\text{Med}(X_1, X_2, \ldots, X_n) - F^{\text{inv}}(0.5)$.

For the Old Faithful data we carried out the empirical bootstrap simulation for the centered sample mean with one thousand repetitions. In Figure 18.1 a histogram (left) and kernel density estimate (right) are displayed of one thousand centered bootstrap sample means

$$\bar{x}_{n,1}^* - \bar{x}_n \quad \bar{x}_{n,2}^* - \bar{x}_n \quad \cdots \quad \bar{x}_{n,1000}^* - \bar{x}_n.$$
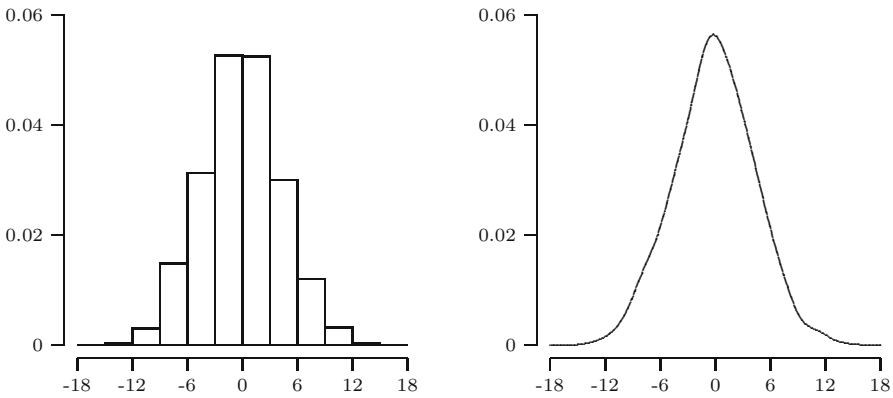
Fig. 18.1. Histogram and kernel density estimate of centered bootstrap sample means.

Since these are realizations of the random variable $\bar{X}_n^* - \bar{x}_n$, we know from Section 17.2 that they reflect the distribution of $\bar{X}_n^* - \bar{x}_n$. Hence, as the distribution of $\bar{X}_n^* - \bar{x}_n$ approximates that of $\bar{X}_n - \mu$, the centered bootstrap sample means also reflect the distribution of $\bar{X}_n - \mu$. This leads to the following application.

An application of the empirical bootstrap

Let us return to our example about the Old Faithful data, which are modeled as a realization of a random sample from some F. Suppose we estimate the expectation μ corresponding to F by $\bar{x}_n = 209.3$. Can we say how far away 209.3 is from the "true" expectation μ? To be honest, the answer is *no*...(oops). In a situation like this, the measurements and their corresponding average are subject to randomness, so that we cannot say anything with absolute certainty about how far away the average will be from μ. One of the things we can say is how *likely* it is that the average is within a given distance from μ.

To get an impression of how close the average of a dataset of $n = 272$ observed durations of the Old Faithful geyser is to μ, we want to compute the probability that the sample mean deviates more than 5 from μ:

$$P\left(|\bar{X}_n - \mu| > 5\right).$$

Direct computation of this probability is impossible, because we do not know the distribution of the random variable $\bar{X}_n - \mu$. However, since the distribution of $\bar{X}_n^* - \bar{x}_n$ approximates the distribution of $\bar{X}_n - \mu$, we can approximate the probability as follows

$$P\left(|\bar{X}_n - \mu| > 5\right) \approx P\left(|\bar{X}_n^* - \bar{x}_n| > 5\right) = P\left(|\bar{X}_n^* - 209.3| > 5\right),$$

where we have also used that for the Old Faithful data, $\bar{x}_n = 209.3$. As we mentioned before, in principle it is possible to compute the last probability exactly. Since this is too cumbersome, we approximate $\mathrm{P}\big(|\bar{X}_n^* - 209.3| > 5\big)$ by means of the one thousand centered bootstrap sample means obtained from the empirical bootstrap simulation:

$$\bar{x}_{n,1}^* - 209.3 \quad \bar{x}_{n,2}^* - 209.3 \quad \cdots \quad \bar{x}_{n,1000}^* - 209.3.$$

In view of Table 17.2, a natural estimate for $\mathrm{P}\big(|\bar{X}_n^* - 209.3| > 5\big)$ is the relative frequency of centered bootstrap sample means that are greater than 5 in absolute value:

$$\frac{\text{number of } i \text{ with } |\bar{x}_{n,i}^* - 209.3| \text{ greater than } 5}{1000}.$$

For the centered bootstrap sample means of Figure 18.1, this relative frequency is 0.227. Hence, we obtain the following bootstrap approximation

$$\mathrm{P}\big(|\bar{X}_n - \mu| > 5\big) \approx \mathrm{P}\big(|\bar{X}_n^* - 209.3| > 5\big) \approx 0.227.$$

It should be emphasized that the second approximation can be made arbitrarily accurate by increasing the number of repetitions in the bootstrap procedure.

18.3 The parametric bootstrap

Suppose we consider our dataset as a realization of a random sample from a distribution of a specific parametric type. In that case the distribution function is completely determined by a parameter or vector of parameters θ: $F = F_\theta$. Then we do *not* have to estimate the whole distribution function F, but it suffices to estimate the parameter(vector) θ by $\hat{\theta}$ and estimate F by

$$\hat{F} = F_{\hat{\theta}}.$$

The corresponding bootstrap principle is called the *parametric bootstrap.*

Let us investigate what this would mean for the centered sample mean. First we should realize that the expectation of F_θ is also determined by θ: $\mu = \mu_\theta$. The parametric bootstrap for the centered sample mean now amounts to the following. The random sample $X_1, X_2, \ldots, X_n$ from the "true" distribution function F_θ is replaced by a bootstrap random sample $X_1^*, X_2^*, \ldots, X_n^*$ from $F_{\hat{\theta}}$, and the probability distribution of $\bar{X}_n - \mu_\theta$ is approximated by that of $\bar{X}_n^* - \mu^*$, where

$$\mu^* = \mu_{\hat{\theta}}$$

denotes the expectation corresponding to $F_{\hat{\theta}}$.

Often the parametric bootstrap approximation is better than the empirical bootstrap approximation, as illustrated in the next quick exercise.

QUICK EXERCISE 18.3 Suppose the dataset $x_1, x_2, \ldots, x_n$ is a realization of a random sample $X_1, X_2, \ldots, X_n$ from an $N(\mu, 1)$ distribution. Estimate μ by $\bar{x}_n$ and consider a bootstrap random sample $X_1^*, X_2^*, \ldots, X_n^*$ from an $N(\bar{x}_n, 1)$ distribution. Check that the probability distributions of $\bar{X}_n - \mu$ and $\bar{X}_n^* - \bar{x}_n$ are the *same*: an $N(0, 1/n)$ distribution.

Once more, in principle it is possible to determine the distribution of $\bar{X}_n^* - \mu_{\hat{\theta}}$ exactly. However, in contrast with the situation considered in the previous quick exercise, in some cases this is still cumbersome. Again a simulation procedure may help us out. For the centered sample mean the procedure is as follows.

> PARAMETRIC BOOTSTRAP SIMULATION (FOR $\bar{X}_n - \mu$). Given a dataset $x_1, x_2, \ldots, x_n$, compute an estimate $\hat{\theta}$ for θ. Determine $F_{\hat{\theta}}$ as an estimate for F_θ, and compute the expectation $\mu^* = \mu_{\hat{\theta}}$ corresponding to $F_{\hat{\theta}}$.
>
> 1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from $F_{\hat{\theta}}$.
> 2. Compute the centered sample mean for the bootstrap dataset:
>
> $$\bar{x}_n^* - \mu_{\hat{\theta}},$$
>
> where
>
> $$\bar{x}_n^* = \frac{x_1^* + x_2^* + \cdots + x_n^*}{n}.$$
>
> Repeat steps 1 and 2 many times.

As an application we will use the parametric bootstrap simulation to investigate whether the exponential distribution is a reasonable model for the software data.

Are the software data exponential?

Consider fitting an exponential distribution to the software data, as discussed in Section 17.3. At first sight, Figure 17.6 shows a reasonable fit with the exponential distribution. One way to quantify the difference between the dataset and the exponential model is to compute the maximum distance between the empirical distribution function F_n of the dataset and the exponential distribution function $F_{\hat{\lambda}}$ estimated from the dataset:

$$t_{\mathrm{ks}} = \sup_{a \in \mathbb{R}} |F_n(a) - F_{\hat{\lambda}}(a)|.$$

Here $F_{\hat{\lambda}}(a) = 0$ for $a < 0$ and

$$F_{\hat{\lambda}}(a) = 1 - \mathrm{e}^{-\hat{\lambda}a} \quad \text{for } a \geq 0,$$

where $\hat{\lambda} = 1/\bar{x}_n$ is estimated from the dataset. The quantity t_{ks} is called the *Kolmogorov-Smirnov distance* between F_n and $F_{\hat{\lambda}}$.

The idea behind the use of this distance is the following. If F denotes the "true" distribution function, then according to Section 17.2 the empirical distribution function F_n will resemble F whether F equals the distribution function F_λ of some $Exp(\lambda)$ distribution or not. On the other hand, if the "true" distribution function *is* F_λ, then the estimated exponential distribution function $F_{\hat\lambda}$ will resemble F_λ, because $\hat\lambda = 1/\bar x_n$ is close to the "true" λ. Therefore, if $F = F_\lambda$, then both F_n and $F_{\hat\lambda}$ will be close to the same distribution function, so that t_{ks} is small; if F is different from F_λ, then F_n and $F_{\hat\lambda}$ are close to two different distribution functions, so that t_{ks} is large. The value t_{ks} is always between 0 and 1, and the further away this value is from 0, the more it is an indication that the exponential model is inappropriate. For the software dataset we find $\hat\lambda = 1/\bar x_n = 0.0015$ and $t_{\mathrm{ks}} = 0.176$. Does this speak against the believed exponential model?

One way to investigate this is to find out whether, in the case when the data are truly a realization of an exponential random sample from F_λ, the value 0.176 is unusually large. To answer this question we consider the sample statistic that corresponds to t_{ks}. The estimate $\hat\lambda = 1/\bar x_n$ is replaced by the random variable $\hat\Lambda = 1/\bar X_n$, and the empirical distribution function of the dataset is replaced by the empirical distribution function of the random sample $X_1, X_2, \ldots, X_n$ (again denoted by F_n):

$$F_n(a) = \frac{\text{number of } X_i \text{ less than or equal to } a}{n}.$$

In this way, t_{ks} is a realization of the sample statistic

$$T_{\mathrm{ks}} = \sup_{a \in \mathbb{R}} |F_n(a) - F_{\hat\Lambda}(a)|.$$

To find out whether 0.176 is an exceptionally large value for the random variable T_{ks}, we must determine the probability distribution of T_{ks}. However, this is impossible because the parameter λ of the $Exp(\lambda)$ distribution is unknown. We will approximate the distribution of T_{ks} by a parametric bootstrap. We use the dataset to estimate λ by $\hat\lambda = 1/\bar x_n = 0.0015$ and replace the random sample $X_1, X_2, \ldots, X_n$ from F_λ by a bootstrap random sample $X_1^*, X_2^*, \ldots, X_n^*$ from $F_{\hat\lambda}$. Next we approximate the distribution of T_{ks} by that of its bootstrapped version

$$T_{\mathrm{ks}}^* = \sup_{a \in \mathbb{R}} |F_n^*(a) - F_{\hat\Lambda^*}(a)|,$$

where F_n^* is the empirical distribution function of the bootstrap random sample:

$$F_n^*(a) = \frac{\text{number of } X_i^* \text{ less than or equal to } a}{n},$$

and $\hat\Lambda^* = 1/\bar X_n^*$, with $\bar X_n^*$ being the average of the bootstrap random sample.

The bootstrapped sample statistic T_{ks}^* is too complicated to determine its probability distribution, and hence we perform a parametric bootstrap simulation:

1. We generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_{135}^*$ from an exponential distribution with parameter $\hat{\lambda} = 0.0015$.

2. We compute the bootstrapped KS distance

$$t_{\text{ks}}^* = \sup_{a \in \mathbb{R}} |F_n^*(a) - F_{\hat{\lambda}^*}(a)|,$$

where F_n^* denotes the empirical distribution function of the bootstrap dataset and $F_{\hat{\lambda}^*}$ denotes the estimated exponential distribution function, where $\hat{\lambda}^* = 1/\bar{x}_n^*$ is computed from the bootstrap dataset.

We repeat steps 1 and 2 one thousand times, which results in one thousand values of the bootstrapped KS distance. In Figure 18.2 we have displayed a histogram and kernel density estimate of the one thousand bootstrapped KS distances. It is clear that if the software data would come from an exponential distribution, the value 0.176 of the KS distance would be very unlikely! This strongly suggests that the exponential distribution is not the right model for the software data. The reason for this is that the Poisson process is the wrong model for the series of failures. A closer inspection shows that the rate at which failures occur over time is not constant, as was assumed in Chapter 17, but decreases.
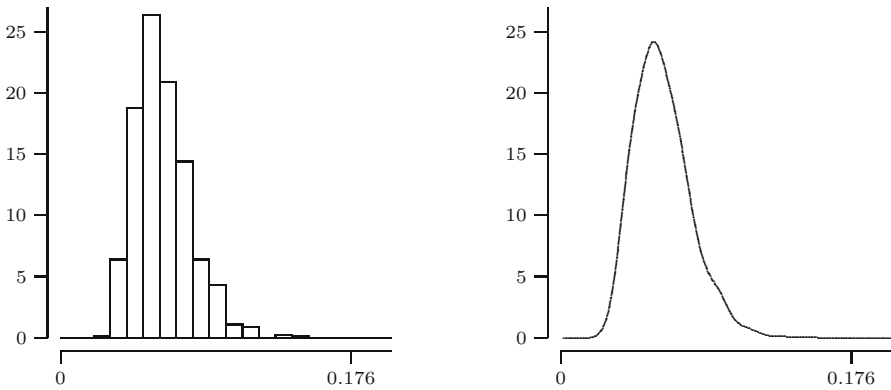

Fig. 18.2. One thousand bootstrapped KS distances.

18.4 Solutions to the quick exercises

18.1 You could have written something like the following: *"Use the dataset $x_1, x_2, \ldots, x_n$ to compute an estimate $\hat{F}$ for F. Replace the random sample $X_1, X_2, \ldots, X_n$ from F by a random sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{F}$, and approximate the probability distribution of*

$$Med(X_1, X_2, \ldots, X_n) - F^{\text{inv}}(0.5)$$

by that of $Med(X_1^*, X_2^*, \ldots, X_n^*) - \hat{F}^{\text{inv}}(0.5)$, where $\hat{F}^{\text{inv}}(0.5)$ is the median of $\hat{F}$."

18.2 You could have written something like the following: "*Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F, and the median $F^{\text{inv}}(0.5)$ of F_n.*

1. *Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.*
2. *Compute the sample median for the bootstrap dataset:*

$$Med_n^* - F^{\text{inv}}(0.5),$$

where $Med_n^* =$ sample median of $x_1^*, x_2^*, \ldots, x_n^*$.

Repeat steps 1 and 2 many times."

Note that if n is odd, then $F^{\text{inv}}(0.5)$ equals the sample median of the original dataset, but this is not necessarily so for n even.

18.3 According to Remark 11.2 about the sum of independent normal random variables, the sum of n independent $N(\mu, 1)$ distributed random variables has an $N(n\mu, n)$ distribution. Hence by the change-of-units rule for the normal distribution (see page 106), it follows that $\bar{X}_n$ has an $N(\mu, 1/n)$ distribution, and that $\bar{X}_n - \mu$ has an $N(0, 1/n)$ distribution. Similarly, the average $\bar{X}_n^*$ of n independent $N(\bar{x}_n, 1)$ distributed bootstrap random variables has a normal distribution $N(\bar{x}_n, 1/n)$ distribution, and therefore $\bar{X}_n^* - \bar{x}_n$ again has an $N(0, 1/n)$ distribution.

18.5 Exercises

18.1 ☐ We generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_6^*$ from the empirical distribution function of the dataset

$$2 \quad 1 \quad 1 \quad 4 \quad 6 \quad 3,$$

i.e., we draw (with replacement) six values from these numbers with equal probability $1/6$. How many different bootstrap datasets are possible? Are they all equally likely to occur?

18.2 We generate a bootstrap dataset $x_1^*, x_2^*, x_3^*, x_4^*$ from the empirical distribution function of the dataset

$$1 \quad 3 \quad 4 \quad 6.$$

a. Compute the probability that the bootstrap sample mean is equal to 1.

b. Compute the probability that the maximum of the bootstrap dataset is equal to 6.

c. Compute the probability that exactly two elements in the bootstrap sample are less than 2.

18.3 ⊞ We generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_{10}^*$ from the empirical distribution function of the dataset

$$0.39 \quad 0.41 \quad 0.38 \quad 0.44 \quad 0.40$$
$$0.36 \quad 0.34 \quad 0.46 \quad 0.35 \quad 0.37.$$

a. Compute the probability that the bootstrap dataset has exactly three elements equal to 0.35.

b. Compute the probability that the bootstrap dataset has at most two elements less than or equal to 0.38.

c. Compute the probability that the bootstrap dataset has exactly two elements less than or equal to 0.38 and all other elements greater than 0.42.

18.4 ⊡ Consider the dataset from Exercise 18.3, with maximum 0.46.

a. We generate a bootstrap random sample $X_1^*, X_2^*, \ldots, X_{10}^*$ from the empirical distribution function of the dataset. Compute $\mathrm{P}(M_{10}^* < 0.46)$, where $M_{10}^* = \max\{X_1^*, X_2^*, \ldots, X_{10}^*\}$.

b. The same question as in **a**, but now for a dataset with distinct elements $x_1, x_2, \ldots, x_n$ and maximum m_n. Compute $\mathrm{P}(M_n^* < m_n)$, where M_n^* is the maximum of a bootstrap random sample $X_1^*, X_2^*, \ldots, X_n^*$ generated from the empirical distribution function of the dataset.

18.5 ⊡ Suppose we have a dataset

$$0 \quad 3 \quad 6,$$

which is the realization of a random sample from a distribution function F. If we estimate F by the empirical distribution function, then according to the bootstrap principle applied to the centered sample mean $\bar{X}_3 - \mu$, we must replace this random variable by its bootstrapped version $\bar{X}_3^* - \bar{x}_3$. Determine the possible values for the bootstrap random variable $\bar{X}_3^* - \bar{x}_3$ and the corresponding probabilities.

18.6 Suppose that the dataset $x_1, x_2, \ldots, x_n$ is a realization of a random sample from an $Exp(\lambda)$ distribution with distribution function F_λ, and that $\bar{x}_n = 5$.

a. Check that the median of the $Exp(\lambda)$ distribution is $m_\lambda = (\ln 2)/\lambda$ (see also Exercise 5.11).

b. Suppose we estimate λ by $1/\bar{x}_n$. Describe the parametric bootstrap simulation for $\mathrm{Med}(X_1, X_2, \ldots, X_n) - m_\lambda$.

18.7 ⊞ To give an example in which the bootstrapped centered sample mean in the parametric and empirical bootstrap simulations may be *different*, consider the following situation. Suppose that the dataset $x_1, x_2, \ldots, x_n$ is a realization of a random sample from a $U(0, \theta)$ distribution with expectation $\mu = \theta/2$. We estimate θ by

$$\hat{\theta} = \frac{n+1}{n} m_n,$$

where $m_n = \max\{x_1, x_2, \ldots, x_n\}$. Describe the parametric bootstrap simulation for the centered sample mean $\bar{X}_n - \mu$.

18.8 ⊞ Here is an example in which the bootstrapped centered sample mean in the parametric and empirical bootstrap simulations are the *same*. Consider the software data with average $\bar{x}_n = 656.8815$ and median $m_n = 290$, modeled as a realization of a random sample $X_1, X_2, \ldots, X_n$ from a distribution function F with expectation μ. By means of bootstrap simulation we like to get an impression of the distribution of $\bar{X}_n - \mu$.

a. Suppose that we assume nothing about the distribution of the interfailure times. Describe the appropriate bootstrap simulation procedure with one thousand repetitions.

b. Suppose we assume that F is the distribution function of an $Exp(\lambda)$ distribution, where λ is estimated by $1/\bar{x}_n = 0.0015$. Describe the appropriate bootstrap simulation procedure with one thousand repetitions.

c. Suppose we assume that F is the distribution function of an $Exp(\lambda)$ distribution, and that (as suggested by Exercise 18.6 a) the parameter λ is estimated by $(\ln 2)/m_n = 0.0024$. Describe the appropriate bootstrap simulation procedure with one thousand repetitions.

18.9 ⊡ Consider the dataset from Exercises 15.1 and 17.6 consisting of measured chest circumferences of Scottish soldiers with average $\bar{x}_n = 39.85$ and sample standard deviation $s_n = 2.09$. The histogram in Figure 17.11 suggests modeling the data as the realization of a random sample $X_1, X_2, \ldots, X_n$ from an $N(\mu, \sigma^2)$ distribution. We estimate μ by the sample mean and we are interested in the probability that the sample mean deviates more than 1 from μ: $P(|\bar{X}_n - \mu| > 1)$. Describe how one can use the bootstrap principle to approximate this probability, i.e., describe the distribution of the bootstrap random sample $X_1^*, X_2^*, \ldots, X_n^*$ and compute $P(|\bar{X}_n^* - \mu^*| > 1)$. Note that one does not need a simulation to approximate this latter probability.

18.10 Consider the software data, with average $\bar{x}_n = 656.8815$, modeled as a realization of a random sample $X_1, X_2, \ldots, X_n$ from a distribution function F. We estimate the expectation μ of F by the sample mean and we are interested in the probability that the sample mean deviates more than ten from μ: $P(|\bar{X}_n - \mu| > 10)$.

a. Suppose we assume nothing about the distribution of the interfailure times. Describe how one can obtain a bootstrap approximation for the probability, i.e., describe the appropriate bootstrap simulation procedure with one thousand repetitions and how the results of this simulation can be used to approximate the probability.

b. Suppose we assume that F is the distribution function of an $Exp(\lambda)$ distribution. Describe how one can obtain a bootstrap approximation for the probability.

18.11 Consider the dataset of measured chest circumferences of 5732 Scottish soldiers (see Exercises 15.1, 17.6, and 18.9). The Kolmogorov-Smirnov distance between the empirical distribution function and the distribution function $F_{\bar{x}_n, s_n}$ of the normal distribution with estimated parameters $\hat{\mu} = \bar{x}_n = 39.85$ and $\hat{\sigma} = s_n = 2.09$ is equal to

$$t_{\mathrm{ks}} = \sup_{a \in \mathbb{R}} |F_n(a) - F_{\bar{x}_n, s_n}(a)| = 0.0987,$$

where $\bar{x}_n$ and s_n denote sample mean and sample standard deviation of the dataset. Suppose we want to perform a bootstrap simulation with one thousand repetitions for the KS distance to investigate to which degree the value 0.0987 agrees with the assumed normality of the dataset. Describe the appropriate bootstrap simulation that must be carried out.

18.12 To give an example where the empirical bootstrap fails, consider the following situation. Suppose our dataset $x_1, x_2, \ldots, x_n$ is a realization of a random sample $X_1, X_2, \ldots, X_n$ from a $U(0, \theta)$ distribution. Consider the normalized sample statistic

$$T_n = 1 - \frac{M_n}{\theta},$$

where M_n is the maximum of $X_1, X_2, \ldots, X_n$. Let $X_1^*, X_2^*, \ldots, X_n^*$ be a bootstrap random sample from the empirical distribution function of our dataset, and let M_n^* be the corresponding bootstrap maximum. We are going to compare the distribution functions of T_n and its bootstrap counterpart

$$T_n^* = 1 - \frac{M_n^*}{m_n},$$

where m_n is the maximum of $x_1, x_2, \ldots, x_n$.

a. Check that $P(T_n \le 0) = 0$ and show that

$$P(T_n^* \le 0) = 1 - \left(1 - \frac{1}{n}\right)^n.$$

Hint: first argue that $P(T_n^* \le 0) = P(M_n^* = m_n)$, and then use the result of Exercise 18.4.

b. Let $G_n(t) = P(T_n \leq t)$ be the distribution function of T_n, and similarly let $G_n^*(t) = P(T_n^* \leq t)$ be the distribution function of the bootstrap statistic T_n^*. Conclude from part **a** that the maximum distance between G_n^* and G_n can be bounded from below as follows:

$$\sup_{t \in \mathbb{R}} |G_n^*(t) - G_n(t)| \geq 1 - \left(1 - \frac{1}{n}\right)^n.$$

c. Use part **b** to argue that for all n, the maximum distance between G_n^* and G_n is greater than 0.632:

$$\sup_{t \in \mathbb{R}} |G_n^*(t) - G_n(t)| \geq 1 - e^{-1} = 0.632.$$

Hint: you may use that $e^{-x} \geq 1 - x$ for all x.

We conclude that even for very large sample sizes the maximum distance between the distribution functions of T_n and its bootstrap counterpart T_n^* is at least 0.632.

18.13 (Exercise 18.12 continued). In contrast to the empirical bootstrap, the parametric bootstrap for T_n *does* work. Suppose we estimate the parameter θ of the $U(0, \theta)$ distribution by

$$\hat{\theta} = \frac{n+1}{n} m_n, \quad \text{where } m_n = \text{maximum of } x_1, x_2, \ldots, x_n.$$

Let now $X_1^*, X_2^*, \ldots, X_n^*$ be a bootstrap random sample from a $U(0, \hat{\theta})$ distribution, and let M_n^* be the corresponding bootstrap maximum. Again, we are going to compare the distribution function G_n of $T_n = 1 - M_n/\theta$ with the distribution function G_n^* of its bootstrap counterpart $T_n^* = 1 - M_n^*/\hat{\theta}$.

a. Check that the distribution function F_θ of a $U(0, \theta)$ distribution is given by

$$F_\theta(a) = \frac{a}{\theta} \quad \text{for } 0 \leq a \leq \theta.$$

b. Check that the distribution function of T_n is

$$G_n(t) = P(T_n \leq t) = 1 - (1 - t)^n \quad \text{for } 0 \leq t \leq 1.$$

Hint: rewrite $P(T_n \leq t)$ as $1 - P(M_n \leq \theta(1 - t))$ and use the rule on page 109 about the distribution function of the maximum.

c. Show that T_n^* has the same distribution function:

$$G_n^*(t) = P(T_n^* \leq t) = 1 - (1 - t)^n \quad \text{for } 0 \leq t \leq 1.$$

This means that, in contrast to the empirical bootstrap (see Exercise 18.12), the parametric bootstrap works perfectly in this situation.

19

Unbiased estimators

In Chapter 17 we saw that a dataset can be modeled as a realization of a random sample from a probability distribution and that quantities of interest correspond to features of the model distribution. One of our tasks is to use the dataset to estimate a quantity of interest. We shall mainly deal with the situation where it is modeled as one of the parameters of the model distribution or as a certain function of the parameters. We will first discuss what we mean exactly by an *estimator* and then introduce the notion of *unbiasedness* as a desirable property for estimators. We end the chapter by providing unbiased estimators for the expectation and variance of a model distribution.

19.1 Estimators

Consider the arrivals of packages at a network server. One is interested in the intensity at which packages arrive on a generic day and in the percentage of minutes during which no packages arrive. If the arrivals occur completely at random in time, the arrival process can be modeled by a Poisson process. This would mean that the number of arrivals during one minute is modeled by a random variable having a Poisson distribution with (unknown) parameter μ. The intensity of the arrivals is then modeled by the parameter μ itself, and the percentage of minutes during which no packages arrive is modeled by the probability of zero arrivals: $e^{-\mu}$. Suppose one observes the arrival process for a while and gathers a dataset $x_1, x_2, \ldots, x_n$, where x_i represents the number of arrivals in the ith minute. Our task will be to estimate, based on the dataset, the parameter μ and a function of the parameter: $e^{-\mu}$.

This example is typical for the general situation in which our dataset is modeled as a realization of a random sample $X_1, X_2, \ldots, X_n$ from a probability distribution that is completely determined by one or more parameters. The parameters that determine the model distribution are called the *model parameters*. We focus on the situation where the quantity of interest corresponds

to a feature of the model distribution that can be described by the model parameters themselves or by some function of the model parameters. This distribution feature is referred to as the *parameter of interest*. In discussing this general setup we shall denote the parameter of interest by the Greek letter θ. So, for instance, in our network server example, μ is the model parameter. When we are interested in the arrival intensity, the role of θ is played by the parameter μ itself, and when we are interested in the percentage of idle minutes the role of θ is played by $e^{-\mu}$.

Whatever method we use to estimate the parameter of interest θ, the result depends only on our dataset.

> ESTIMATE. An *estimate* is a value t that only depends on the dataset $x_1, x_2, \ldots, x_n$, i.e., t is some function of the dataset only:
>
> $$t = h(x_1, x_2, \ldots, x_n).$$

This description of *estimate* is a bit formal. The idea is, of course, that the value t, computed from our dataset $x_1, x_2, \ldots, x_n$, gives some indication of the "true" value of the parameter θ. We have already met several estimates in Chapter 17; see, for instance, Table 17.2. This table illustrates that the value of an estimate can be anything: a single number, a vector of numbers, even a complete curve.

Let us return to our network server example in which our dataset $x_1, x_2, \ldots, x_n$ is modeled as a realization of a random sample from a $Pois(\mu)$ distribution. The intensity at which packages arrive is then represented by the parameter μ. Since the parameter μ is the expectation of the model distribution, the law of large numbers suggests the sample mean $\bar{x}_n$ as a natural estimate for μ. On the other hand, the parameter μ also represents the variance of the model distribution, so that by a similar reasoning another natural estimate is the sample variance s_n^2.

The percentage of idle minutes is modeled by the probability of zero arrivals. Similar to the reasoning in Section 13.4, a natural estimate is the relative frequency of zeros in the dataset:

$$\frac{\text{number of } x_i \text{ equal to zero}}{n}.$$

On the other hand, the probability of zero arrivals can be expressed as a function of the model parameter: $e^{-\mu}$. Hence, if we estimate μ by $\bar{x}_n$, we could also estimate $e^{-\mu}$ by $e^{-\bar{x}_n}$.

QUICK EXERCISE 19.1 Suppose we estimate the probability of zero arrivals $e^{-\mu}$ by the relative frequency of x_i equal to zero. Deduce an estimate for μ from this.

The preceding examples illustrate that one can often think of several estimates for the parameter of interest. This raises questions like

- When is one estimate better than another?
- Does there exist a best possible estimate?

For instance, can we say which of the values $\bar{x}_n$ or s_n^2 computed from the dataset is closer to the "true" parameter μ? The answer is *no*. The measurements and the corresponding estimates are subject to randomness, so that we cannot say anything with certainty about which of the two is closer to μ. One of the things we can say for each of them is *how likely* it is that they are within a given distance from μ. To this end, we consider the random variables that correspond to the estimates. Because our dataset $x_1, x_2, \ldots, x_n$ is modeled as a realization of a random sample $X_1, X_2, \ldots, X_n$, the estimate t is a realization of a random variable T.

> ESTIMATOR.   Let $t = h(x_1, x_2, \ldots, x_n)$ be an estimate based on the dataset $x_1, x_2, \ldots, x_n$. Then t is a realization of the random variable
>
> $$T = h(X_1, X_2, \ldots, X_n).$$
>
> The random variable T is called an *estimator*.

The word *estimator* refers to the method or device for estimation. This is distinguished from *estimate*, which refers to the actual value computed from a dataset. Note that estimators are special cases of sample statistics. In the remainder of this chapter we will discuss the notion of *unbiasedness* that describes to some extent the behavior of estimators.

19.2 Investigating the behavior of an estimator

Let us continue with our network server example. Suppose we have observed the network for 30 minutes and we have recorded the number of arrivals in each minute. The dataset is modeled as a realization of a random sample $X_1, X_2, \ldots, X_n$ of size $n = 30$ from a $Pois(\mu)$ distribution. Let us concentrate on estimating the probability p_0 of zero arrivals, which is an unknown number between 0 and 1. As motivated in the previous section, we have the following possible estimators:

$$S = \frac{\text{number of } X_i \text{ equal to zero}}{n} \quad \text{and} \quad T = e^{-\bar{X}_n}.$$

Our first estimator S can only attain the values $0, \frac{1}{30}, \frac{2}{30}, \ldots, 1$, so that in general it *cannot* give the exact value of p_0. Similarly for our second estimator T, which can only attain the values $1, e^{-1/30}, e^{-2/30}, \ldots$. So clearly, we

cannot expect our estimators always to give the exact value of p_0 on basis of 30 observations. Well, then what *can* we expect from a reasonable estimator?

To get an idea of the behavior of both estimators, we pretend we know μ and we simulate the estimation process in the case of $n = 30$ observations. Let us choose $\mu = \ln 10$, so that $p_0 = e^{-\mu} = 0.1$. We draw 30 values from a Poisson distribution with parameter $\mu = \ln 10$ and compute the value of estimators S and T. We repeat this 500 times, so that we have 500 values for each estimator. In Figure 19.1 a frequency histogram[1] of these values for estimator S is displayed on the left and for estimator T on the right. Clearly, the values of both estimators vary around the value 0.1, which they are supposed to estimate.
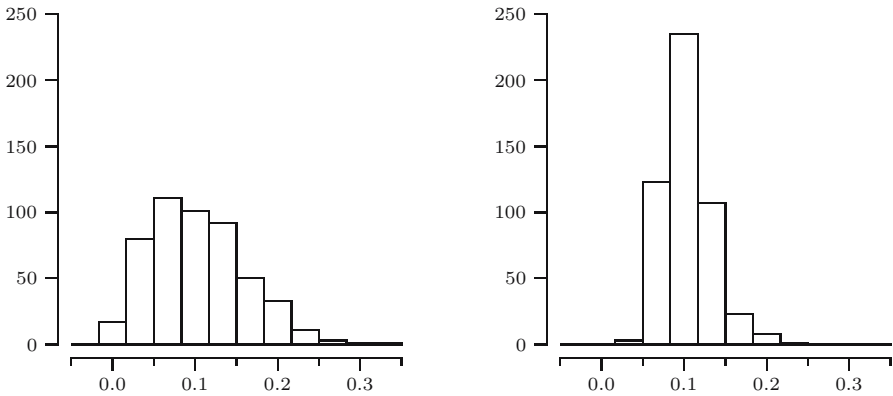
Fig. 19.1. Frequency histograms of 500 values for estimators S (left) and T (right) of $p_0 = 0.1$.

19.3 The sampling distribution and unbiasedness

We have just seen that the values generated for estimator S fluctuate around $p_0 = 0.1$. Although the value of this estimator is not always equal to 0.1, it is desirable that on average, S is on target, i.e., $\mathrm{E}[S] = 0.1$. Moreover, it is desirable that this property holds no matter what the actual value of p_0 is, i.e.,

$$\mathrm{E}[S] = p_0$$

irrespective of the value $0 < p_0 < 1$. In order to find out whether this is true, we need the probability distribution of the estimator S. Of course this

---

[1] In a frequency histogram the height of each vertical bar equals the frequency of values in the corresponding bin.

is simply the distribution of a random variable, but because estimators are constructed from a random sample $X_1, X_2, \ldots, X_n$, we speak of the sampling distribution.

THE SAMPLING DISTRIBUTION.   Let $T = h(X_1, X_2, \ldots, X_n)$ be an estimator based on a random sample $X_1, X_2, \ldots, X_n$. The probability distribution of T is called the *sampling distribution* of T.

The sampling distribution of S can be found as follows. Write

$$S = \frac{Y}{n},$$

where Y is the number of X_i equal to zero. If for each i we label $X_i = 0$ as a success, then Y is equal to the number of successes in n independent trials with p_0 as the probability of success. Similar to Section 4.3, it follows that Y has a $Bin(n, p_0)$ distribution. Hence the sampling distribution of S is that of a $Bin(n, p_0)$ distributed random variable divided by n. This means that S is a discrete random variable that attains the values k/n, where $k = 0, 1, \ldots, n$, with probabilities given by

$$p_S\left(\frac{k}{n}\right) = \mathrm{P}\left(S = \frac{k}{n}\right) = \mathrm{P}(Y = k) = \binom{n}{k} p_0^k (1 - p_0)^{n-k}.$$

The probability mass function of S for the case $n = 30$ and $p_0 = 0.1$ is displayed in Figure 19.2. Since $S = Y/n$ and Y has a $Bin(n, p_0)$ distribution, it follows that

$$\mathrm{E}[S] = \frac{\mathrm{E}[Y]}{n} = \frac{np_0}{n} = p_0.$$

So, indeed, the estimator S for p_0 has the property $\mathrm{E}[S] = p_0$. This property reflects the fact that estimator S has no systematic tendency to produce
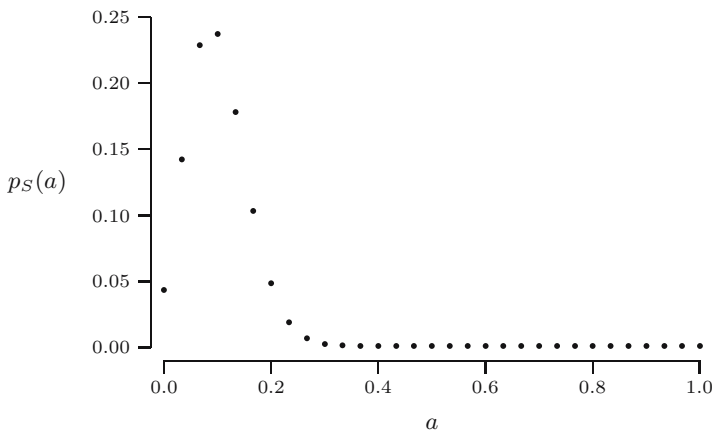
Fig. 19.2. Probability mass function of S.

estimates that are larger than p_0, and no systematic tendency to produce estimates that are smaller than p_0. This is a desirable property for estimators, and estimators that have this property are called unbiased.

> DEFINITION. An estimator T is called an *unbiased* estimator for the parameter θ, if
> $$\mathrm{E}[T] = \theta$$
> irrespective of the value of θ. The difference $\mathrm{E}[T] - \theta$ is called the *bias* of T; if this difference is nonzero, then T is called *biased*.

Let us return to our second estimator for the probability of zero arrivals in the network server example: $T = \mathrm{e}^{-\bar{X}_n}$. The sampling distribution can be obtained as follows. Write
$$T = \mathrm{e}^{-Z/n},$$
where $Z = X_1 + X_2 + \cdots + X_n$. From Exercise 12.9 we know that the random variable Z, being the sum of n independent $Pois(\mu)$ random variables, has a $Pois(n\mu)$ distribution. This means that T is a discrete random variable attaining values $\mathrm{e}^{-k/n}$, where $k = 0, 1, \ldots$ and the probability mass function of T is given by

$$p_T\left(\mathrm{e}^{-k/n}\right) = \mathrm{P}\left(T = \mathrm{e}^{-k/n}\right) = \mathrm{P}(Z = k) = \frac{\mathrm{e}^{-n\mu}(n\mu)^k}{k!}.$$

The probability mass function of T for the case $n = 30$ and $p_0 = 0.1$ is displayed in Figure 19.3. From the histogram in Figure 19.1 as well as from the probability mass function in Figure 19.3, you may get the impression that T is also an unbiased estimator. However, this not the case, which follows immediately from an application of Jensen's inequality:
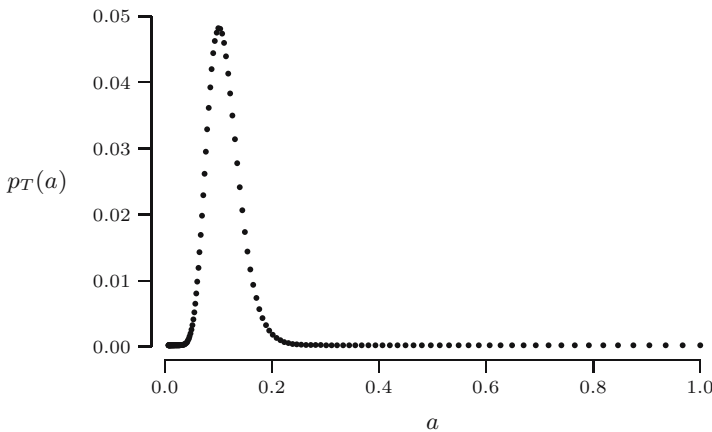
Fig. 19.3. Probability mass function of T.

$$\mathrm{E}[T] = \mathrm{E}\left[\mathrm{e}^{-\bar{X}_n}\right] > \mathrm{e}^{-\mathrm{E}[\bar{X}_n]},$$

where we have a strict inequality because the function $g(x) = \mathrm{e}^{-x}$ is strictly convex ($g''(x) = \mathrm{e}^{-x} > 0$). Recall that the parameter μ equals the expectation of the $Pois(\mu)$ model distribution, so that according to Section 13.1 we have $\mathrm{E}[\bar{X}_n] = \mu$. We find that

$$\mathrm{E}[T] > \mathrm{e}^{-\mu} = p_0,$$

which means that the estimator T for p_0 has positive bias. In fact we can compute $\mathrm{E}[T]$ exactly (see Exercise 19.9):

$$\mathrm{E}[T] = \mathrm{E}\left[\mathrm{e}^{-\bar{X}_n}\right] = \mathrm{e}^{-n\mu(1-\mathrm{e}^{-1/n})}.$$

Note that $n(1 - \mathrm{e}^{-1/n}) \to 1$, so that

$$\mathrm{E}[T] = \mathrm{e}^{-n\mu(1-\mathrm{e}^{-1/n})} \to \mathrm{e}^{-\mu} = p_0$$

as n goes to infinity. Hence, although T has positive bias, the bias decreases to zero as the sample size becomes larger. In Figure 19.4 the expectation of T is displayed as a function of the sample size n for the case $\mu = \ln(10)$. For $n = 30$ the difference between $\mathrm{E}[T]$ and $p_0 = 0.1$ equals 0.0038.
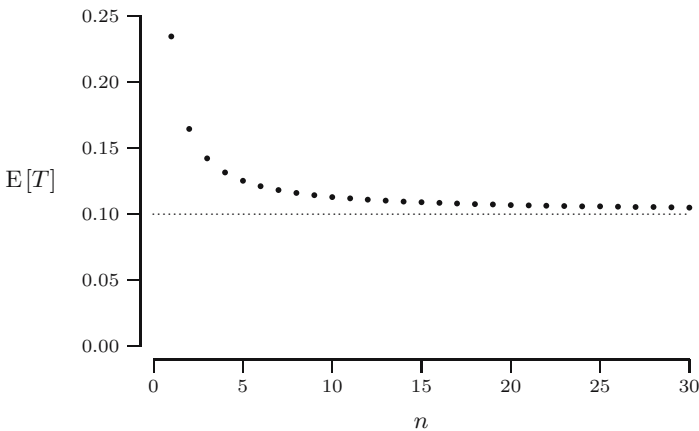
Fig. 19.4. $\mathrm{E}[T]$ as a function of n.

QUICK EXERCISE 19.2 If we estimate $p_0 = \mathrm{e}^{-\mu}$ by the relative frequency of zeros $S = Y/n$, then we could estimate μ by $U = -\ln(S)$. Argue that U is a biased estimator for μ. Is the bias positive or negative?

We conclude this section by returning to the estimation of the parameter μ. Apart from the (biased) estimator in Quick exercise 19.2 we also considered

the sample mean $\bar{X}_n$ and sample variance S_n^2 as possible estimators for μ. These are both unbiased estimators for the parameter μ. This is a direct consequence of a more general property of $\bar{X}_n$ and S_n^2, which is discussed in the next section.

19.4 Unbiased estimators for expectation and variance

Sometimes the quantity of interest can be described by the expectation or variance of the model distribution, and is it irrelevant whether this distribution is of a parametric type. In this section we propose unbiased estimators for these distribution features.

UNBIASED ESTIMATORS FOR EXPECTATION AND VARIANCE. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with finite expectation μ and finite variance σ^2. Then

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is an *unbiased estimator for* μ and

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

is an *unbiased estimator for* σ^2.

The first statement says that $E[\bar{X}_n] = \mu$, which was shown in Section 13.1. The second statement says $E[S_n^2] = \sigma^2$. To see this, use linearity of expectations to write

$$E[S_n^2] = \frac{1}{n-1} \sum_{i=1}^{n} E[(X_i - \bar{X}_n)^2].$$

Since $E[\bar{X}_n] = \mu$, we have $E[X_i - \bar{X}_n] = E[X_i] - E[\bar{X}_n] = 0$. Now note that for any random variable Y with $E[Y] = 0$, we have

$$\text{Var}(Y) = E[Y^2] - (E[Y])^2 = E[Y^2].$$

Applying this to $Y = X_i - \bar{X}_n$, it follows that

$$E[(X_i - \bar{X}_n)^2] = \text{Var}(X_i - \bar{X}_n).$$

Note that we can write

$$X_i - \bar{X}_n = \frac{n-1}{n} X_i - \frac{1}{n} \sum_{j \neq i} X_j.$$

Then from the rules concerning variances of sums of independent random variables we find that

$$\mathrm{Var}\left(X_i - \bar{X}_n\right) = \mathrm{Var}\left(\frac{n-1}{n}X_i - \frac{1}{n}\sum_{j\neq i}X_j\right)$$

$$= \frac{(n-1)^2}{n^2}\mathrm{Var}(X_i) + \frac{1}{n^2}\sum_{j\neq i}\mathrm{Var}(X_j)$$

$$= \left[\frac{(n-1)^2}{n^2} + \frac{n-1}{n^2}\right]\sigma^2 = \frac{n-1}{n}\sigma^2.$$

We conclude that

$$\mathrm{E}\left[S_n^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\left[(X_i - \bar{X}_n)^2\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{Var}\left(X_i - \bar{X}_n\right) = \frac{1}{n-1}\cdot n \cdot \frac{n-1}{n}\sigma^2 = \sigma^2.$$

This explains why we divide by $n-1$ in the formula for S_n^2; only in this case S_n^2 is an unbiased estimator for the "true" variance σ^2. If we would divide by n instead of $n-1$, we would obtain an estimator with negative bias; it would systematically produce too-small estimates for σ^2.

QUICK EXERCISE 19.3 Consider the following estimator for σ^2:

$$V_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Compute the bias $\mathrm{E}\left[V_n^2\right] - \sigma^2$ for this estimator, where you can keep computations simple by realizing that $V_n^2 = (n-1)S_n^2/n$.

Unbiasedness does not always carry over

We have seen that S_n^2 is an unbiased estimator for the "true" variance σ^2. A natural question is whether S_n is again an unbiased estimator for σ. This is not the case. Since the function $g(x) = x^2$ is strictly convex, Jensen's inequality yields that

$$\sigma^2 = \mathrm{E}\left[S_n^2\right] > (\mathrm{E}[S_n])^2,$$

which implies that $\mathrm{E}[S_n] < \sigma$. Another example is the network arrivals, in which $\bar{X}_n$ is an unbiased estimator for μ, whereas $\mathrm{e}^{-\bar{X}_n}$ is positively biased with respect to $\mathrm{e}^{-\mu}$. These examples illustrate a general fact: unbiasedness *does not* always carry over, i.e., if T is an unbiased estimator for a parameter θ, then $g(T)$ *does not* have to be an unbiased estimator for $g(\theta)$.

However, there is one special case in which unbiasedness does carry over, namely if $g(T) = aT + b$. Indeed, if T is unbiased for θ: $\mathrm{E}[T] = \theta$, then by the change-of-units rule for expectations,

$$\mathrm{E}[aT + b] = a\mathrm{E}[T] + b = a\theta + b,$$

which means that $aT + b$ is unbiased for $a\theta + b$.

19.5 Solutions to the quick exercises

19.1 Write y for the number of x_i equal to zero. Denote the probability of zero by p_0, so that $p_0 = \mathrm{e}^{-\mu}$. This means that $\mu = -\ln(p_0)$. Hence if we estimate p_0 by the relative frequency y/n, we can estimate μ by $-\ln(y/n)$.

19.2 The function $g(x) = -\ln(x)$ is strictly convex, since $g''(x) = 1/x^2 > 0$. Hence by Jensen's inequality

$$\mathrm{E}[U] = \mathrm{E}[-\ln(S)] > -\ln(\mathrm{E}[S]).$$

Since we have seen that $\mathrm{E}[S] = p_0 = \mathrm{e}^{-\mu}$, it follows that $\mathrm{E}[U] > -\ln(\mathrm{E}[S]) = -\ln(\mathrm{e}^{-\mu}) = \mu$. This means that U has positive bias.

19.3 Using that $\mathrm{E}[S_n^2] = \sigma^2$, we find that

$$\mathrm{E}[V_n^2] = \mathrm{E}\left[\frac{n-1}{n} S_n^2\right] = \frac{n-1}{n} \mathrm{E}[S_n^2] = \frac{n-1}{n}\sigma^2.$$

We conclude that the bias of V_n^2 equals $\mathrm{E}[V_n^2] - \sigma^2 = -\sigma^2/n < 0$.

19.6 Exercises

19.1 ⊞ Suppose our dataset is a realization of a random sample $X_1, X_2, \ldots, X_n$ from a uniform distribution on the interval $[-\theta, \theta]$, where θ is unknown.

a. Show that
$$T = \frac{3}{n}(X_1^2 + X_2^2 + \cdots + X_n^2)$$
is an unbiased estimator for θ^2.

b. Is $\sqrt{T}$ also an unbiased estimator for θ? If not, argue whether it has positive or negative bias.

19.2 Suppose the random variables $X_1, X_2, \ldots, X_n$ have the same expectation μ.

a. Is $S = \frac{1}{2}X_1 + \frac{1}{3}X_2 + \frac{1}{6}X_3$ an unbiased estimator for μ?

b. Under what conditions on constants $a_1, a_2, \ldots, a_n$ is

$$T = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

an unbiased estimator for μ?

19.3 ⊡ Suppose the random variables $X_1, X_2, \ldots, X_n$ have the same expectation μ. For which constants a and b is

$$T = a(X_1 + X_2 + \cdots + X_n) + b$$

an unbiased estimator for μ?

19.4 Recall Exercise 17.5 about the number of cycles to pregnancy. Suppose the dataset corresponding to the table in Exercise 17.5 a is modeled as a realization of a random sample $X_1, X_2, \ldots, X_n$ from a $Geo(p)$ distribution, where $0 < p < 1$ is unknown. Motivated by the law of large numbers, a natural estimator for p is

$$T = 1/\bar{X}_n.$$

a. Check that T is a biased estimator for p and find out whether it has positive or negative bias.

b. In Exercise 17.5 we discussed the estimation of the probability that a woman becomes pregnant within three or fewer cycles. One possible estimator for this probability is the relative frequency of women that became pregnant within three cycles

$$S = \frac{\text{number of } X_i \le 3}{n}.$$

Show that S is an unbiased estimator for this probability.

19.5 ⊡ Suppose a dataset is modeled as a realization of a random sample $X_1, X_2, \ldots, X_n$ from an $Exp(\lambda)$ distribution, where $\lambda > 0$ is unknown. Let μ denote the corresponding expectation and let M_n denote the minimum of $X_1, X_2, \ldots, X_n$. Recall from Exercise 8.18 that M_n has an $Exp(n\lambda)$ distribution. Find out for which constant c the estimator

$$T = cM_n$$

is an unbiased estimator for μ.

19.6 ⊡ Consider the following dataset of lifetimes of ball bearings in hours.

6278	3113	5236	11584	12628	7725	8604	14266	6125	9350
3212	9003	3523	12888	9460	13431	17809	2812	11825	2398

Source: J.E. Angus. Goodness-of-fit tests for exponentiality based on a loss-of-memory type functional equation. *Journal of Statistical Planning and Inference*, 6:241-251, 1982; example 5 on page 249.

One is interested in estimating the minimum lifetime of this type of ball bearing. The dataset is modeled as a realization of a random sample $X_1, \ldots, X_n$. Each random variable X_i is represented as

$$X_i = \delta + Y_i,$$

where Y_i has an $Exp(\lambda)$ distribution and $\delta > 0$ is an unknown parameter that is supposed to model the minimum lifetime. The objective is to construct an unbiased estimator for δ. It is known that

$$\mathrm{E}[M_n] = \delta + \frac{1}{n\lambda} \quad \text{and} \quad \mathrm{E}[\bar{X}_n] = \delta + \frac{1}{\lambda},$$

where $M_n = $ minimum of $X_1, X_2, \ldots, X_n$ and $\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n$.

a. Check that

$$T = \frac{n}{n-1}\left(\bar{X}_n - M_n\right)$$

is an unbiased estimator for $1/\lambda$.

b. Construct an unbiased estimator for δ.

c. Use the dataset to compute an estimate for the minimum lifetime δ. You may use that the average lifetime of the data is 8563.5.

19.7 Leaves are divided into four different types: starchy-green, sugary-white, starchy-white, and sugary-green. According to genetic theory, the types occur with probabilities $\frac{1}{4}(\theta + 2)$, $\frac{1}{4}\theta$, $\frac{1}{4}(1 - \theta)$, and $\frac{1}{4}(1 - \theta)$, respectively, where $0 < \theta < 1$. Suppose one has n leaves. Then the number of starchy-green leaves is modeled by a random variable N_1 with a $Bin(n, p_1)$ distribution, where $p_1 = \frac{1}{4}(\theta + 2)$, and the number of sugary-white leaves is modeled by a random variable N_2 with a $Bin(n, p_2)$ distribution, where $p_2 = \frac{1}{4}\theta$. The following table lists the counts for the progeny of self-fertilized heterozygotes among 3839 leaves.

Type	Count
Starchy-green	1997
Sugary-white	32
Starchy-white	906
Sugary-green	904

Source: R.A. Fisher. *Statistical methods for research workers*. Hafner, New York, 1958; Table 62 on page 299.

Consider the following two estimators for θ:

$$T_1 = \frac{4}{n}N_1 - 2 \quad \text{and} \quad T_2 = \frac{4}{n}N_2.$$

a. Check that both T_1 and T_2 are unbiased estimators for θ.

b. Compute the value of both estimators for θ.

19.8 ⊞ Recall the black cherry trees example from Exercise 17.9, modeled by a linear regression model without intercept

$$Y_i = \beta x_i + U_i \quad \text{for } i = 1, 2, \ldots, n,$$

where $U_1, U_2, \ldots, U_n$ are independent random variables with $\mathrm{E}[U_i] = 0$ and $\mathrm{Var}(U_i) = \sigma^2$. We discussed three estimators for the parameter β:

$$B_1 = \frac{1}{n}\left(\frac{Y_1}{x_1} + \cdots + \frac{Y_n}{x_n}\right),$$

$$B_2 = \frac{Y_1 + \cdots + Y_n}{x_1 + \cdots + x_n},$$

$$B_3 = \frac{x_1 Y_1 + \cdots + x_n Y_n}{x_1^2 + \cdots + x_n^2}.$$

Show that all three estimators are unbiased for β.

19.9 Consider the network example where the dataset is modeled as a realization of a random sample $X_1, X_2, \ldots, X_n$ from a *Pois*(μ) distribution. We estimate the probability of zero arrivals $\mathrm{e}^{-\mu}$ by means of $T = \mathrm{e}^{-\bar{X}_n}$. Check that

$$\mathrm{E}[T] = \mathrm{e}^{-n\mu(1-\mathrm{e}^{-1/n})}.$$

Hint: write $T = \mathrm{e}^{-Z/n}$, where $Z = X_1 + X_2 + \cdots + X_n$ has a *Pois*$(n\mu)$ distribution.

Efficiency and mean squared error

In the previous chapter we introduced the notion of unbiasedness as a desirable property of an estimator. If several unbiased estimators for the same parameter of interest exist, we need a criterion for comparison of these estimators. A natural criterion is some measure of spread of the estimators around the parameter of interest. For unbiased estimators we will use variance. For arbitrary estimators we introduce the notion of *mean squared error* (MSE), which combines variance and bias.

20.1 Estimating the number of German tanks

In this section we come back to the problem of estimating German war production as discussed in Section 1.5. We consider serial numbers on tanks, recoded to numbers running from 1 to some unknown largest number N. Given is a subset of n numbers of this set. The objective is to estimate the total number of tanks N on the basis of the observed serial numbers.

Denote the observed distinct serial numbers by $x_1, x_2, \ldots, x_n$. This dataset can be modeled as a realization of random variables $X_1, X_2, \ldots, X_n$ representing n draws *without replacement* from the numbers $1, 2, \ldots, N$ with equal probability. Note that in this example our dataset is *not* a realization of a random sample, because the random variables $X_1, X_2, \ldots, X_n$ are *dependent*. We propose two unbiased estimators. The first one is based on the sample mean

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

and the second one is based on the sample maximum

$$M_n = \max\{X_1, X_2, \ldots, X_n\}.$$

An estimator based on the sample mean

To construct an unbiased estimator for N based on the sample mean, we start by computing the expectation of $\bar{X}_n$. The linearity-of-expectations rule also applies to dependent random variables, so that

$$\mathrm{E}\left[\bar{X}_n\right] = \frac{\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n]}{n}.$$

In Section 9.3 we saw that the marginal distribution of each X_i is the same:

$$\mathrm{P}(X_i = k) = \frac{1}{N} \quad \text{for } k = 1, 2, \ldots, N.$$

Therefore the expectation of each X_i is given by

$$\mathrm{E}[X_i] = 1 \cdot \frac{1}{N} + 2 \cdot \frac{1}{N} + \cdots + N \cdot \frac{1}{N} = \frac{1 + 2 + \cdots + N}{N}$$
$$= \frac{\frac{1}{2}N(N+1)}{N} = \frac{N+1}{2}.$$

It follows that

$$\mathrm{E}\left[\bar{X}_n\right] = \frac{\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n]}{n} = \frac{N+1}{2}.$$

This directly implies that

$$T_1 = 2\bar{X}_n - 1$$

is an unbiased estimator for N, since the change-of-units rule yields that

$$\mathrm{E}[T_1] = \mathrm{E}\left[2\bar{X}_n - 1\right] = 2\mathrm{E}\left[\bar{X}_n\right] - 1 = 2 \cdot \frac{N+1}{2} - 1 = N.$$

QUICK EXERCISE 20.1 Suppose we have observed tanks with (recoded) serial numbers

$$61 \quad 19 \quad 56 \quad 24 \quad 16.$$

Compute the value of the estimator T_1 for the total number of tanks.

An estimator based on the sample maximum

To construct an unbiased estimator for N based on the maximum, we first compute the expectation of M_n. We start by computing the probability that $M_n = k$, where k takes the values $n, \ldots, N$. Similar to the combinatorics used in Section 4.3 to derive the binomial distribution, the number of ways to draw n numbers without replacement from $1, 2, \ldots, N$ is $\binom{N}{n}$. Hence each combination has probability $1/\binom{N}{n}$. In order to have $M_n = k$, we must have one number equal to k and choose the other $n-1$ numbers out of the numbers $1, 2, \ldots, k-1$. There are $\binom{k-1}{n-1}$ ways to do this. Hence for the possible values $k = n, n+1, \ldots, N$,

$$P(M_n = k) = \frac{\binom{k-1}{n-1}}{\binom{N}{n}} = \frac{(k-1)!}{(k-n)!(n-1)!} \cdot \frac{(N-n)!\,n!}{N!}$$

$$= n \cdot \frac{(k-1)!}{(k-n)!} \frac{(N-n)!}{N!}.$$

Thus the expectation of M_n is given by

$$E[M_n] = \sum_{k=n}^{N} kP(M_n = k) = \sum_{k=n}^{N} k \cdot n \cdot \frac{(k-1)!}{(k-n)!} \frac{(N-n)!}{N!}$$

$$= \sum_{k=n}^{N} n \cdot \frac{k!}{(k-n)!} \frac{(N-n)!}{N!}$$

$$= n \cdot \frac{(N-n)!}{N!} \sum_{k=n}^{N} \frac{k!}{(k-n)!}.$$

How to continue the computation of $E[M_n]$? We use a trick: we start by rearranging

$$1 = \sum_{j=n}^{N} P(M_n = j) = \sum_{j=n}^{N} n \cdot \frac{(j-1)!}{(j-n)!} \frac{(N-n)!}{N!},$$

finding that

$$\sum_{j=n}^{N} \frac{(j-1)!}{(j-n)!} = \frac{N!}{n\,(N-n)!}. \tag{20.1}$$

This holds for any N and any $n \le N$. In particular we could replace N by $N+1$ and n by $n+1$:

$$\sum_{j=n+1}^{N+1} \frac{(j-1)!}{(j-n-1)!} = \frac{(N+1)!}{(n+1)(N-n)!}.$$

Changing the summation variable to $k = j - 1$, we obtain

$$\sum_{k=n}^{N} \frac{k!}{(k-n)!} = \frac{(N+1)!}{(n+1)(N-n)!}. \tag{20.2}$$

This is exactly what we need to finish the computation of $E[M_n]$. Substituting (20.2) in what we obtained earlier, we find

$$E[M_n] = n \cdot \frac{(N-n)!}{N!} \sum_{k=n}^{N} \frac{k!}{(k-n)!}$$

$$= n \cdot \frac{(N-n)!}{N!} \cdot \frac{(N+1)!}{(n+1)(N-n)!} = n \cdot \frac{N+1}{n+1}.$$

QUICK EXERCISE 20.2 Choosing $n = N$ in this formula yields $\mathrm{E}[M_N] = N$. Can you argue that this is the right answer without doing any computations?

With the formula for $\mathrm{E}[M_n]$ we can derive immediately that

$$T_2 = \frac{n+1}{n} M_n - 1$$

is an unbiased estimator for N, since by the change-of-units rule,

$$\mathrm{E}[T_2] = \mathrm{E}\left[\frac{n+1}{n} M_n - 1\right] = \frac{n+1}{n}\mathrm{E}[M_n] - 1 = \frac{n+1}{n} \cdot \frac{n(N+1)}{n+1} - 1 = N.$$

QUICK EXERCISE 20.3 Compute the value of estimator T_2 for the total number of tanks on basis of the observed numbers from Quick exercise 20.1.

20.2 Variance of an estimator

In the previous section we saw that we can construct two completely different estimators for the total number of tanks N that are *both* unbiased. The obvious question is: which of the two is better? To answer this question, we investigate how both estimators vary around the parameter of interest N. Although we could in principle compute the distributions of T_1 and T_2, we carry out a small simulation study instead. Take $N = 1000$ and $n = 10$ fixed. We draw 10 numbers, without replacement, from $1, 2, \ldots, 1000$ and compute the value of the estimators T_1 and T_2. We repeat this two thousand times, so that we have 2000 values for both estimators. In Figure 20.1 we have displayed the histogram of the 2000 values for T_1 on the left and the histogram of the 2000 values for T_2 on the right. From the histograms, which reflect the probability
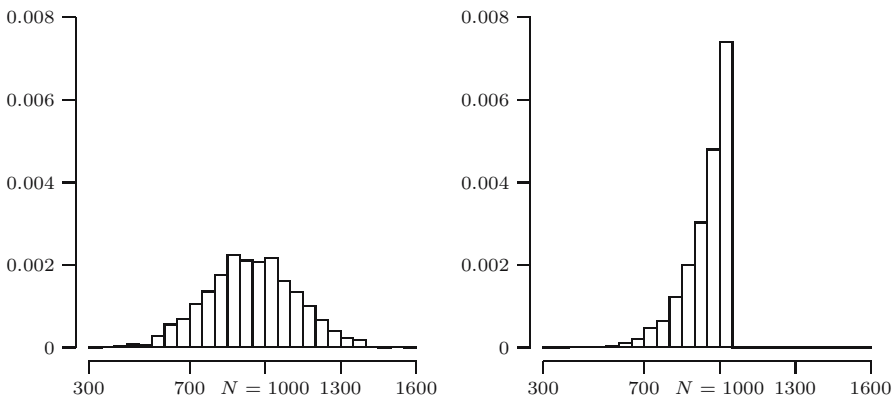

Fig. 20.1. Histograms of two thousand values for T_1 (left) and T_2 (right).

mass functions of both estimators, we see that the distributions of T_1 and T_2 are of completely different types. As can be expected from the fact that both estimators are unbiased, the values vary around the parameter of interest $N = 1000$. The most important difference between the histograms is that the variation in the values of T_2 is *less* than the variation in the values of T_1. This suggests that estimator T_2 estimates the total number of tanks more efficiently than estimator T_1, in the sense that it produces estimates that are more concentrated around the parameter of interest N than estimates produced by T_1. Recall that the variance measures the spread of a random variable. Hence the previous discussion motivates the use of the variance of an estimator to evaluate its performance.

EFFICIENCY. Let T_1 and T_2 be two unbiased estimators for the same parameter θ. Then estimator T_2 is called *more efficient* than estimator T_1 if $\mathrm{Var}(T_2) < \mathrm{Var}(T_1)$, irrespective of the value of θ.

Let us compare T_1 and T_2 using this criterion. For T_1 we have

$$\mathrm{Var}(T_1) = \mathrm{Var}\left(2\bar{X}_n - 1\right) = 4\mathrm{Var}\left(\bar{X}_n\right).$$

Although the X_i are not independent, it is true that all pairs (X_i, X_j) with $i \neq j$ have the *same* distribution (this follows in the same way in which we showed on page 122 that all X_i have the same distribution). With the variance-of-the-sum rule for n random variables (see Exercise 10.17), we find that

$$\mathrm{Var}(X_1 + \cdots + X_n) = n\mathrm{Var}(X_1) + n(n-1)\mathrm{Cov}(X_1, X_2).$$

In Exercises 9.18 and 10.18, we computed that

$$\mathrm{Var}(X_1) = \frac{1}{12}(N-1)(N+1), \quad \mathrm{Cov}(X_1, X_2) = -\frac{1}{12}(N+1).$$

We find therefore that

$$\begin{aligned}
\mathrm{Var}(T_1) &= 4\mathrm{Var}\left(\bar{X}_n\right) = \frac{4}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) \\
&= \frac{4}{n^2}\left[n \cdot \frac{1}{12}(N-1)(N+1) - n(n-1) \cdot \frac{1}{12}(N+1)\right] \\
&= \frac{1}{3n}(N+1)[N - 1 - (n-1)] \\
&= \frac{(N+1)(N-n)}{3n}.
\end{aligned}$$

Obtaining the variance of T_2 is a little more work. One can compute the variance of M_n in a way that is very similar to the way we obtained $\mathrm{E}[M_n]$. The result is (see Remark 20.1 for details)

$$\mathrm{Var}(M_n) = \frac{n(N+1)(N-n)}{(n+2)(n+1)^2}.$$

Remark 20.1 (How to compute this variance). The trick is to compute not $\mathrm{E}\left[M_n^2\right]$ but $\mathrm{E}\left[M_n(M_n+1)\right]$. First we derive an identity from Equation (20.1) as before, this time replacing N by $N+2$ and n by $n+2$:

$$\sum_{j=n+2}^{N+2} \frac{(j-1)!}{(j-n-2)!} = \frac{(N+2)!}{(n+2)(N-n)!}.$$

Changing the summation variable to $k = j-2$ yields

$$\sum_{k=n}^{N} \frac{(k+1)!}{(k-n)!} = \frac{(N+2)!}{(n+2)(N-n)!}.$$

With this formula one can obtain:

$$\mathrm{E}\left[M_n(M_n+1)\right] = \sum_{k=n}^{N} k(k+1) \cdot n \frac{(k-1)!}{(k-n)!} \frac{(N-n)!}{N!} = \frac{n(N+1)(N+2)}{n+2}.$$

Since we know $\mathrm{E}\left[M_n\right]$, we can determine $\mathrm{E}\left[M_n^2\right]$ from this, and subsequently the variance of M_n.

With the expression for the variance of M_n, we derive

$$\mathrm{Var}(T_2) = \mathrm{Var}\left(\frac{n+1}{n}M_n - 1\right) = \frac{(n+1)^2}{n^2}\mathrm{Var}(M_n) = \frac{(N+1)(N-n)}{n(n+2)}.$$

We see that $\mathrm{Var}(T_2) < \mathrm{Var}(T_1)$ for all N and $n \geq 2$. Hence T_2 is always more efficient than T_1, except when $n = 1$. In this case the variances are equal, simply because the estimators are the same—they both equal X_1.

The quotient $\mathrm{Var}(T_1)/\mathrm{Var}(T_2)$, is called the *relative efficiency* of T_2 with respect to T_1. In our case the relative efficiency of T_2 with respect to T_1 equals

$$\frac{\mathrm{Var}(T_1)}{\mathrm{Var}(T_2)} = \frac{(N+1)(N-n)}{3n} \cdot \frac{n(n+2)}{(N+1)(N-n)} = \frac{n+2}{3}.$$

Surprisingly, this quotient does not depend on N, and we see clearly the advantage of T_2 over T_1 as the sample size n gets larger.

QUICK EXERCISE 20.4 Let $n = 5$, and let the sample be

$$7 \quad 3 \quad 10 \quad 45 \quad 15.$$

Compute the value of the estimator T_1 for N. Do you notice anything strange?

The self-contradictory behavior of T_1 in Quick exercise 20.4 is not rare: this phenomenon will occur for up to 50% of the samples if n and N are large. This gives another reason to prefer T_2 over T_1.

Remark 20.2 (The Cramér-Rao inequality). Suppose we have a random sample from a continuous distribution with probability density function f_θ, where θ is the parameter of interest. Under certain smoothness conditions on the density f_θ, the variance of an unbiased estimator T for θ always has to be larger than or equal to a certain positive number, the so-called Cramér-Rao lower bound:

$$\text{Var}(T) \geq \frac{1}{n\text{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_\theta(X)\right)^2\right]} \qquad \text{for all } \theta.$$

Here n is the size of the sample and X a random variable whose density function is f_θ. In some cases we can find unbiased estimators attaining this bound. These are called *minimum variance unbiased estimators*. An example is the sample mean for the expectation of an exponential distribution. (We will consider this case in Exercise 20.3.)

20.3 Mean squared error

In the last section we compared two unbiased estimators by considering their spread around the value to be estimated, where the spread was measured by the variance. Although unbiasedness is a desirable property, the performance of an estimator should mainly be judged by the way it spreads around the parameter θ to be estimated. This leads to the following definition.

> DEFINITION. Let T be an estimator for a parameter θ. The *mean squared error* of T is the number $\text{MSE}(T) = \text{E}\left[(T - \theta)^2\right]$.

According to this criterion, an estimator T_1 performs better than an estimator T_2 if $\text{MSE}(T_1) < \text{MSE}(T_2)$. Note that

$$
\begin{aligned}
\text{MSE}(T) &= \text{E}\left[(T - \theta)^2\right] \\
&= \text{E}\left[(T - \text{E}[T] + \text{E}[T] - \theta)^2\right] \\
&= \text{E}\left[(T - \text{E}[T])^2\right] + 2\text{E}[T - \text{E}[T]](\text{E}[T] - \theta) + (\text{E}[T] - \theta)^2 \\
&= \text{Var}(T) + (\text{E}[T] - \theta)^2.
\end{aligned}
$$

So the MSE of T turns out to be the variance of T plus the square of the bias of T. In particular, when T is unbiased, the MSE of T is just the variance of T. This means that we already used mean squared errors to compare the estimators T_1 and T_2 in the previous section. We extend the notion of efficiency by saying that estimator T_2 is more *efficient* than estimator T_1 (for the same parameter of interest), if the MSE of T_2 is smaller than the MSE of T_1.

Unbiasedness and efficiency

A biased estimator with a small variance may be more useful than an unbiased estimator with a large variance. We illustrate this with the network server
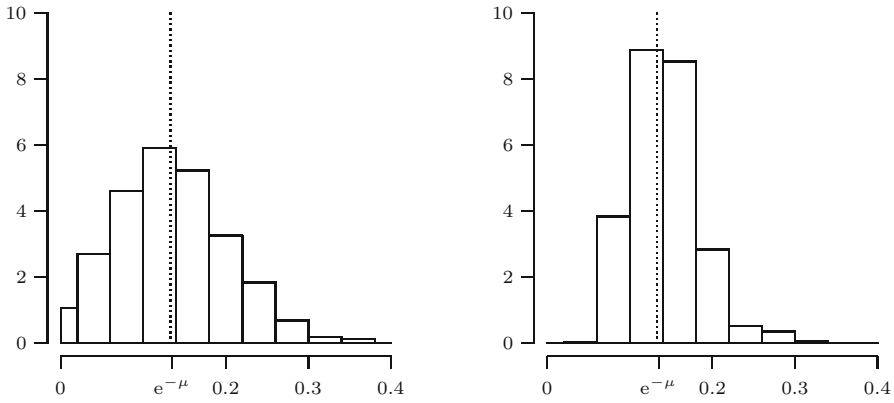
Fig. 20.2. Histograms of a thousand values for S (left) and T (right).

example from Section 19.2. Recall that our goal was to estimate the probability $p_0 = \mathrm{e}^{-\mu}$ of zero arrivals (of packages) in a minute. We did have two promising candidates as estimators:

$$S = \frac{\text{number of } X_i \text{ equal to zero}}{n} \quad \text{and} \quad T = \mathrm{e}^{-\bar{X}_n}.$$

In Figure 20.2 we depict histograms of one thousand simulations of the values of S and T computed for random samples of size $n = 25$ from a $Pois(\mu)$ distribution, where $\mu = 2$. Considering the way the values of the (biased!) estimator T are more concentrated around the true value $\mathrm{e}^{-\mu} = \mathrm{e}^{-2} = 0.1353$, we would be inclined to prefer T over S. This choice is strongly supported by the fact that T is more efficient than S: $\mathrm{MSE}(T)$ is always smaller than $\mathrm{MSE}(S)$, as illustrated in Figure 20.3.
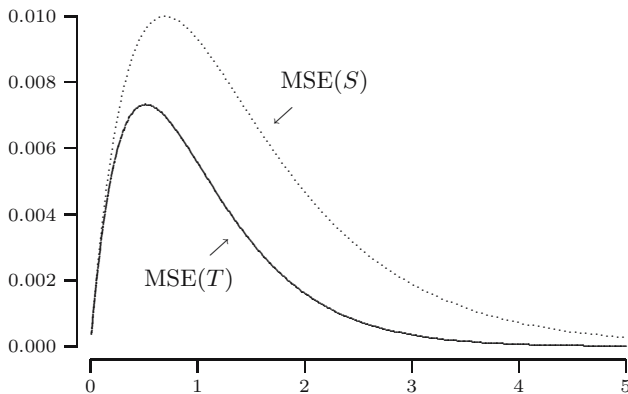
Fig. 20.3. MSEs of S and T as a function of μ.

20.4 Solutions to the quick exercises

20.1 We have $\bar{x}_5 = (61 + 19 + 56 + 24 + 16)/5 = 176/5 = 35.2$. Therefore $t_1 = 2 \cdot 35.2 - 1 = 69.4$.

20.2 When $n = N$, we have drawn *all* the numbers. But then the largest number M_N *is* N, and so $\mathrm{E}[M_N] = N$.

20.3 We have $t_2 = (6/5) \cdot 61 - 1 = 72.2$.

20.4 Since 45 is in the sample, N has to be at least 45. Adding the numbers yields $7 + 3 + 10 + 15 + 45 = 80$. So $t_1 = 2\bar{x}_n - 1 = 2 \cdot 16 - 1 = 31$. What is strange about this is that the estimate for N is far smaller than the number 45 in the sample!

20.5 Exercises

20.1 Given is a random sample $X_1, X_2, \ldots, X_n$ from a distribution with finite variance σ^2. We estimate the expectation of the distribution with the sample mean $\bar{X}_n$. Argue that the larger our sample, the more efficient our estimator. What is the relative efficiency $\mathrm{Var}(\bar{X}_n)/\mathrm{Var}(\bar{X}_{2n})$ of $\bar{X}_{2n}$ with respect to $\bar{X}_n$?

20.2 ⊞ Given are two estimators S and T for a parameter θ. Furthermore it is known that $\mathrm{Var}(S) = 40$ and $\mathrm{Var}(T) = 4$.

a. Suppose that we know that $\mathrm{E}[S] = \theta$ and $\mathrm{E}[T] = \theta + 3$. Which estimator would you prefer, and why?

b. Suppose that we know that $\mathrm{E}[S] = \theta$ and $\mathrm{E}[T] = \theta + a$ for some positive number a. For each a, which estimator would you prefer, and why?

20.3 ⊞ Suppose we have a random sample $X_1, \ldots, X_n$ from an *Exp*(λ) distribution. Suppose we want to estimate the mean $1/\lambda$. According to Section 19.4 the estimator

$$T_1 = \bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

is an unbiased estimator of $1/\lambda$. Let M_n be the minimum of $X_1, X_2, \ldots, X_n$. Recall from Exercise 8.18 that M_n has an *Exp*$(n\lambda)$ distribution. In Exercise 19.5 you have determined that

$$T_2 = nM_n$$

is another unbiased estimator for $1/\lambda$. Which of the estimators T_1 and T_2 would you choose for estimating the mean $1/\lambda$? Substantiate your answer.

20.4 □ Consider the situation of this chapter, where we have to estimate the parameter N from a sample $x_1, \ldots, x_n$ drawn without replacement from the numbers $\{1, \ldots, N\}$. To keep it simple, we consider $n = 2$. Let $M = M_2$ be the maximum of X_1 and X_2. We have found that $T_2 = 3M/2 - 1$ is a good unbiased estimator for N. We want to construct a new unbiased estimator T_3 based on the *minimum* L of X_1 and X_2. In the following you may use that the random variable L has the same distribution as the random variable $N + 1 - M$ (this follows from symmetry considerations).

a. Show that $T_3 = 3L - 1$ is an unbiased estimator for N.

b. Compute $\mathrm{Var}(T_3)$ using that $\mathrm{Var}(M) = (N + 1)(N - 2)/18$. (The latter has been computed in Remark 20.1.)

c. What is the relative efficiency of T_2 with respect to T_3?

20.5 Someone is proposing two unbiased estimators U and V, with the *same* variance $\mathrm{Var}(U) = \mathrm{Var}(V)$. It therefore appears that we would not prefer one estimator over the other. However, we could go for a third estimator, namely $W = (U + V)/2$. Note that W is unbiased. To judge the quality of W we want to compute its variance. Lacking information on the joint probability distribution of U and V, this is impossible. However, we should prefer W in any case! To see this, show by means of the variance-of-the-sum rule that the relative efficiency of U with respect to W is equal to

$$\frac{\mathrm{Var}((U + V)/2)}{\mathrm{Var}(U)} = \frac{1}{2} + \frac{1}{2}\rho(U, V).$$

Here $\rho(U, V)$ is the correlation coefficient. Why does this result imply that we should use W instead of U (or V)?

20.6 A geodesic engineer measures the three unknown angles α_1, α_2, and α_3 of a triangle. He models the uncertainty in the measurements by considering them as realizations of three independent random variables T_1, T_2, and T_3 with expectations

$$\mathrm{E}[T_1] = \alpha_1, \quad \mathrm{E}[T_2] = \alpha_2, \quad \mathrm{E}[T_3] = \alpha_3,$$

and all three with the same variance σ^2. In order to make use of the fact that the three angles must add to π, he also considers new estimators U_1, U_2, and U_3 defined by

$$U_1 = T_1 + \tfrac{1}{3}(\pi - T_1 - T_2 - T_3),$$
$$U_2 = T_2 + \tfrac{1}{3}(\pi - T_1 - T_2 - T_3),$$
$$U_3 = T_3 + \tfrac{1}{3}(\pi - T_1 - T_2 - T_3).$$

(Note that the "deviation" $\pi - T_1 - T_2 - T_3$ is evenly divided over the three measurements and that $U_1 + U_2 + U_3 = \pi$.)

a. Compute $\mathrm{E}[U_1]$ and $\mathrm{Var}(U_1)$.

b. What does he gain in efficiency when he uses U_1 instead of T_1 to estimate the angle α_1?

c. What kind of estimator would you choose for α_1 if it is known that the triangle is isosceles (i.e., $\alpha_1 = \alpha_2$)?

20.7 ⊡ (Exercise 19.7 continued.) Leaves are divided into four different types: starchy-green, sugary-white, starchy-white, and sugary-green. According to genetic theory, the types occur with probabilities $\frac{1}{4}(\theta + 2)$, $\frac{1}{4}\theta$, $\frac{1}{4}(1 - \theta)$, and $\frac{1}{4}(1 - \theta)$, respectively, where $0 < \theta < 1$. Suppose one has n leaves. Then the number of starchy-green leaves is modeled by a random variable N_1 with a $Bin(n, p_1)$ distribution, where $p_1 = \frac{1}{4}(\theta + 2)$, and the number of sugary-white leaves is modeled by a random variable N_2 with a $Bin(n, p_2)$ distribution, where $p_2 = \frac{1}{4}\theta$. Consider the following two estimators for θ:

$$T_1 = \frac{4}{n}N_1 - 2 \quad \text{and} \quad T_2 = \frac{4}{n}N_2.$$

In Exercise 19.7 you showed that both T_1 and T_2 are unbiased estimators for θ. Which estimator would you prefer? Motivate your answer.

20.8 ⊞ Let $\bar{X}_n$ and $\bar{Y}_m$ be the sample means of two independent random samples of size n (resp. m) from the same distribution with mean μ. We combine these two estimators to a new estimator T by putting

$$T = r\bar{X}_n + (1 - r)\bar{Y}_m,$$

where r is some number between 0 and 1.

a. Show that T is an unbiased estimator for the mean μ.

b. Show that T is most efficient when $r = n/(n + m)$.

20.9 Given is a random sample $X_1, X_2, \ldots, X_n$ from a $Ber(p)$ distribution. One considers the estimators

$$T_1 = \frac{1}{n}(X_1 + \cdots + X_n) \quad \text{and} \quad T_2 = \min\{X_1, \ldots, X_n\}.$$

a. Are T_1 and T_2 unbiased estimators for p?

b. Show that

$$\mathrm{MSE}(T_1) = \frac{1}{n}p(1 - p), \quad \mathrm{MSE}(T_2) = p^n - 2p^{n+1} + p^2.$$

c. Which estimator is more efficient when $n = 2$?

20.10 Suppose we have a random sample $X_1, \ldots, X_n$ from an $Exp(\lambda)$ distribution. We want to estimate the expectation $1/\lambda$. According to Section 19.4,

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

is an unbiased estimator of $1/\lambda$. Let us consider more generally estimators T of the form

$$T = c \cdot (X_1 + X_2 + \cdots + X_n),$$

where c is a real number. We are interested in the MSE of these estimators and would like to know whether there are choices for c that yield a smaller MSE than the choice $c = 1/n$.

a. Compute $\mathrm{MSE}(T)$ for each c.

b. For which c does the estimator perform best in the MSE sense? Compare this to the unbiased estimator $\bar{X}_n$ that one obtains for $c = 1/n$.

20.11 ⊡ In Exercise 17.9 we modeled diameters of black cherry trees with the linear regression model (without intercept)

$$Y_i = \beta x_i + U_i$$

for $i = 1, 2, \ldots, n$. As usual, the U_i here are independent random variables with $\mathrm{E}[U_i] = 0$, and $\mathrm{Var}(U_i) = \sigma^2$.

We considered three estimators for the slope β of the line $y = \beta x$: the so-called least squares estimator T_1 (which will be considered in Chapter 22), the average slope estimator T_2, and the slope of the averages estimator T_3. These estimators are defined by:

$$T_1 = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}, \qquad T_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{x_i}, \qquad T_3 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i}.$$

In Exercise 19.8 it was shown that all three estimators are unbiased. Compute the MSE of all three estimators.

Remark: it can be shown that T_1 is always more efficient than T_3, which in turn is more efficient than T_2. To prove the first inequality one uses a famous inequality called the Cauchy Schwartz inequality; for the second inequality one uses Jensen's inequality (can you see how?).

20.12 Let $X_1, X_2, \ldots, X_n$ represent n draws *without replacement* from the numbers $1, 2, \ldots, N$ with equal probability. The goal of this exercise is to compute the distribution of M_n in a way other than by the combinatorial analysis we did in this chapter.

a. Compute $\mathrm{P}(M_n \leq k)$, by using, as in Section 8.4, that:

$$\mathrm{P}(M_n \leq k) = \mathrm{P}(X_1 \leq k, X_2 \leq k, \ldots, X_n \leq k).$$

b. Derive that

$$P(M_n = n) = \frac{n!(N-n)!}{N!}.$$

c. Show that for $k = n+1, \ldots, N$

$$P(M_n = k) = n \cdot \frac{(k-1)!}{(k-n)!} \frac{(N-n)!}{N!}.$$

Maximum likelihood

In previous chapters we could easily construct estimators for various parameters of interest because these parameters had a natural sample analogue: expectation versus sample mean, probabilities versus relative frequencies, etc. However, in some situations such an analogue does not exist. In this chapter, a general principle to construct estimators is introduced, the so-called *maximum likelihood principle. Maximum likelihood estimators* have certain attractive properties that are discussed in the last section.

21.1 Why a general principle?

In Section 4.4 we modeled the number of cycles up to pregnancy by a random variable X with a geometric distribution with (unknown) parameter p. Weinberg and Gladen studied the effect of smoking on the number of cycles and obtained the data in Table 21.1 for 100 smokers and 486 nonsmokers.

Table 21.1. Observed numbers of cycles up to pregnancy.

Number of cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Nonsmokers	198	107	55	38	18	22	7	9	5	3	6	6	12

Source: C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

Is the parameter p, which equals the probability of becoming pregnant after one cycle, different for smokers and nonsmokers? Let us try to find out by estimating p in the two cases.

What would be reasonable ways to estimate p? Since $p = \mathrm{P}(X = 1)$, the law of large numbers (see Section 13.3) motivates use of

$$S = \frac{\text{number of } X_i \text{ equal to } 1}{n}$$

as an estimator for p. This yields estimates $p = 29/100 = 0.29$ for smokers and $p = 198/486 = 0.41$ for nonsmokers. We know from Section 19.4 that S is an unbiased estimator for p. However, one cannot escape the feeling that S is a "bad" estimator: S does not use all the information in the table, i.e., the way the women are distributed over the numbers $2, 3, \ldots$ of observed numbers of cycles is not used. One would like to have an estimator that incorporates all the available information. Due to the way the data are given, this seems to be difficult. For instance, estimators based on the average cannot be evaluated, because 7 smokers and 12 nonsmokers had an unknown number of cycles up to pregnancy (larger than 12). If one simply ignores the last column in Table 21.1 as we did in Exercise 17.5, the average can be computed and yields $1/\bar{x}_{93} = 0.2809$ as an estimate of p for smokers and $1/\bar{x}_{474} = 0.3688$ for nonsmokers. However, because we discard seven values larger than 12 in case of the smokers and twelve values larger than 12 in case of the nonsmokers, we overestimate p in both cases.

In the next section we introduce a general principle to find an estimate for a parameter of interest, the *maximum likelihood principle*. This principle yields good estimators and will solve problems such as those stated earlier.

21.2 The maximum likelihood principle

Suppose a dealer of computer chips is offered on the black market two batches of 10 000 chips each. According to the seller, in one batch about 50% of the chips are defective, while this percentage is about 10% in the other batch. Our dealer is only interested in this last batch. Unfortunately the seller cannot tell the two batches apart. To help him to make up his mind, the seller offers our dealer one batch, from which he is allowed to select and test 10 chips. After selecting 10 chips arbitrarily, it turns out that only the second one is defective. Our dealer at once decides to buy this batch. Is this a wise decision?

With the batch where 50% of the chips are defective it is *more likely* that defective chips will appear, whereas with the other batch one would expect hardly any defective chip. Clearly, our dealer chooses the batch for which it is *most likely* that only one chip is defective. This is also the guiding idea behind the maximum likelihood principle.

> THE MAXIMUM LIKELIHOOD PRINCIPLE. Given a dataset, choose the parameter(s) of interest in such a way that the data are most likely.

Set $R_i = 1$ in case the ith tested chip was defective and $R_i = 0$ in case it was operational, where $i = 1, \ldots, 10$. Then $R_1, \ldots, R_{10}$ are ten independent $Ber(p)$ distributed random variables, where p is the probability that a randomly selected chip is defective. The probability that the observed data occur is equal to

$$P(R_1 = 0, R_2 = 1, R_3 = 0, \ldots, R_{10} = 0) = p(1 - p)^9.$$

For the batch where about 10% of the chips are defective we find that

$$P(R_1 = 0, R_2 = 1, R_3 = 0, \ldots, R_{10} = 0) = \frac{1}{10}\left(\frac{9}{10}\right)^9 = 0.039,$$

whereas for the other batch

$$P(R_1 = 0, R_2 = 1, R_3 = 0, \ldots, R_{10} = 0) = \frac{1}{2}\left(\frac{1}{2}\right)^9 = 0.00098.$$

So the probability for the batch with only 10% defective chips is about 40 times larger than the probability for the other batch. Given the data, our dealer made a sound decision.

QUICK EXERCISE 21.1 Which batch should the dealer choose if only the first three chips are defective?

Returning to the example of the number of cycles up to pregnancy, denoting X_i as the number of cycles up to pregnancy of the ith smoker, recall that

$$P(X_i = k) = (1 - p)^{k-1}p$$

and

$$P(X_i > 12) = P(\text{no success in cycle 1 to 12}) = (1 - p)^{12};$$

cf. Quick exercise 4.6. From Table 21.1 we see that there are 29 smokers for which $X_i = 1$, that there are 16 for which $X_i = 2$, etc. Since we model the data as a random sample from a geometric distribution, the probability of the data—as a function of p—is given by

$$\begin{aligned}
L(p) &= C \cdot P(X_i = 1)^{29} \cdot P(X_i = 2)^{16} \cdots P(X_i = 12)^3 \cdot P(X_i > 12)^7 \\
&= C \cdot p^{29} \cdot ((1-p)p)^{16} \cdots ((1-p)^{11}p)^3 \cdot ((1-p)^{12})^7 \\
&= C \cdot p^{93} \cdot (1-p)^{322}.
\end{aligned}$$

Here C is the number of ways we can assign 29 ones, 16 twos, $\ldots$, 3 twelves, and 7 numbers larger than 12 to 100 smokers.[1] According to the *maximum likelihood principle* we now choose p, with $0 \le p \le 1$, in such a way, that $L(p)$

---
[1] $C = 31165702882281944145184268216785480009626362520835911650443115348728076083\allowbreak 2000000000.$

is maximal. Since C does not depend on p, we do not need to know the value of C explicitly to find for which p the function $L(p)$ is maximal.

Differentiating $L(p)$ with respect to p yields that

$$\begin{aligned}
L'(p) &= C\left[93p^{92}(1-p)^{322} - 322p^{93}(1-p)^{321}\right] \\
&= Cp^{92}(1-p)^{321}\left[93(1-p) - 322p\right] \\
&= Cp^{92}(1-p)^{321}(93 - 415p).
\end{aligned}$$

Now $L'(p) = 0$ if $p = 0$, $p = 1$, or $p = 93/415 = 0.224$, and $L(p)$ attains its unique maximum in this last point (check this!). We say that $93/415 = 0.224$ is the *maximum likelihood estimate* of p for the smokers. Note that this estimate is quite a lot smaller than the estimate 0.29 for the smokers we found in the previous section, and the estimate 0.2809 you obtained in Exercise 17.5.

QUICK EXERCISE 21.2 Check that for the nonsmokers the probability of the data is given by
$$L(p) = \text{constant} \cdot p^{474}(1-p)^{955}.$$
Compute the maximum likelihood estimate for p.

> **Remark 21.1 (Some history).** The method of maximum likelihood estimation was propounded by Ronald Aylmer Fisher in a highly influential paper. In fact, this paper does not contain the original statement of the method, which was published by Fisher in 1912 [9], nor does it contain the original definition of *likelihood*, which appeared in 1921 (see [10]). The roots of the maximum likelihood method date back as far as 1713, when Jacob Bernoulli's *Ars Conjectandi* ([1]) was posthumously published. In the eighteenth century other important contributions were by Daniel Bernoulli, Lambert, and Lagrange (see also [2], [16], and [17]). It is interesting to remark that another giant of statistics, Karl Pearson, had not understood Fisher's method. Fisher was hurt by Pearson's lack of understanding, which eventually led to a violent confrontation.

21.3 Likelihood and loglikelihood

Suppose we have a dataset $x_1, x_2, \ldots, x_n$, modeled as a realization of a random sample from a distribution characterized by a parameter θ. To stress the dependence of the distribution on θ, we write

$$p_\theta(x)$$

for the probability mass function in case we have a sample from a discrete distribution and

$$f_\theta(x)$$

for the probability density function when we have a sample from a continuous distribution.

For a dataset $x_1, x_2, \ldots, x_n$ modeled as the realization of a random sample $X_1, \ldots, X_n$ from a *discrete* distribution, the maximum likelihood principle now tells us to estimate θ by that value, for which the function $L(\theta)$, given by

$$L(\theta) = \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = p_\theta(x_1) \cdots p_\theta(x_n)$$

is maximal. This value is called the maximum likelihood estimate of θ. The function $L(\theta)$ is called the *likelihood function*. This is a function of θ, determined by the numbers $x_1, x_2, \ldots, x_n$.

In case the sample is from a *continuous* distribution we clearly need to define the likelihood function $L(\theta)$ in a way different from the discrete case (if we would define $L(\theta)$ as in the discrete case, one always would have that $L(\theta) = 0$). For a reasonable definition of the likelihood function we have the following motivation. Let f_θ be the probability density function of X, and let $\varepsilon > 0$ be some fixed, small number. It is sensible to choose θ in such a way, that the probability $\mathrm{P}(x_1 - \varepsilon \le X_1 \le x_1 + \varepsilon, \ldots, x_n - \varepsilon \le X_n \le x_n + \varepsilon)$ is maximal. Since the X_i are independent, we find that

$$\begin{aligned}
&\mathrm{P}(x_1 - \varepsilon \le X_1 \le x_1 + \varepsilon, \ldots, x_n - \varepsilon \le X_n \le x_n + \varepsilon) \\
&= \mathrm{P}(x_1 - \varepsilon \le X_1 \le x_1 + \varepsilon) \cdots \mathrm{P}(x_n - \varepsilon \le X_n \le x_n + \varepsilon) \quad (21.1) \\
&\approx f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n)(2\varepsilon)^n,
\end{aligned}$$

where in the last step we used that (see also Equation (5.1))

$$\mathrm{P}(x_i - \varepsilon \le X_i \le x_i + \varepsilon) = \int_{x_i - \varepsilon}^{x_i + \varepsilon} f_\theta(x)\, \mathrm{d}x \approx 2\varepsilon f_\theta(x_i).$$

Note that the right-hand side of (21.1) is maximal whenever the function $f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n)$ is maximal, irrespective of the value of ε. In view of this, given a dataset $x_1, x_2, \ldots, x_n$, the likelihood function $L(\theta)$ is defined by

$$L(\theta) = f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n)$$

in the continuous case.

> MAXIMUM LIKELIHOOD ESTIMATES. The *maximum likelihood estimate* of θ is the value $t = h(x_1, x_2, \ldots, x_n)$ that maximizes the likelihood function $L(\theta)$. The corresponding random variable
>
> $$T = h(X_1, X_2, \ldots, X_n)$$
>
> is called the *maximum likelihood estimator* for θ.

As an example, suppose we have a dataset $x_1, x_2, \ldots, x_n$ modeled as a realization of a random sample from an $Exp(\lambda)$ distribution, with probability density function given by $f_\lambda(x) = 0$ if $x < 0$ and

$$f_\lambda(x) = \lambda e^{-\lambda x} \qquad \text{for} \quad x \geq 0.$$

Then the likelihood is given by

$$
\begin{aligned}
L(\lambda) &= f_\lambda(x_1) f_\lambda(x_2) \cdots f_\lambda(x_n) \\
&= \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdots \lambda e^{-\lambda x_n} \\
&= \lambda^n \cdot e^{-\lambda(x_1 + x_2 + \cdots + x_n)}.
\end{aligned}
$$

To obtain the maximum likelihood estimate of λ it is enough to find the maximum of $L(\lambda)$. To do so, we determine the derivative of $L(\lambda)$:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda} L(\lambda) &= n\lambda^{n-1} e^{-\lambda \sum_{i=1}^n x_i} - \lambda^n \left(\sum_{i=1}^n x_i \right) e^{-\lambda \sum_{i=1}^n x_i} \\
&= n \left(\lambda^{n-1} e^{-\lambda \sum_{i=1}^n x_i} \left(1 - \frac{\lambda}{n} \sum_{i=1}^n x_i \right) \right).
\end{aligned}
$$

We see that $\mathrm{d}\,(L(\lambda))\,/\mathrm{d}\lambda = 0$ if and only if

$$1 - \lambda \bar{x}_n = 0,$$

i.e., if $\lambda = 1/\bar{x}_n$. Check that for this value of λ the likelihood function $L(\lambda)$ attains a maximum! So the maximum likelihood estimator for λ is $1/\bar{X}_n$.

In the example of the number of cycles up to pregnancy of smoking women, we have seen that $L(p) = C \cdot p^{93} \cdot (1-p)^{322}$. The maximum likelihood estimate of p was found by differentiating $L(p)$. Differentiating is not always possible, as the following example shows.

Estimating the upper endpoint of a uniform distribution

Suppose the dataset $x_1 = 0.98$, $x_2 = 1.57$, and $x_3 = 0.31$ is the realization of a random sample from a $U(0, \theta)$ distribution with $\theta > 0$ unknown. The probability density function of each X_i is now given by $f_\theta(x) = 0$ if x is not in $[0, \theta]$ and

$$f_\theta(x) = \frac{1}{\theta} \qquad \text{for} \quad 0 \leq x \leq \theta.$$

The likelihood $L(\theta)$ is zero if θ is smaller than at least one of the x_i, and equals $1/\theta^3$ if θ is greater than or equal to each of the three x_i, i.e.,

$$
L(\theta) = f_\theta(x_1) f_\theta(x_2) f_\theta(x_3) =
\begin{cases}
\frac{1}{\theta^3} & \text{if} \quad \theta \geq \max(x_1, x_2, x_3) = 1.57 \\
0 & \text{if} \quad \theta < \max(x_1, x_2, x_3) = 1.57.
\end{cases}
$$
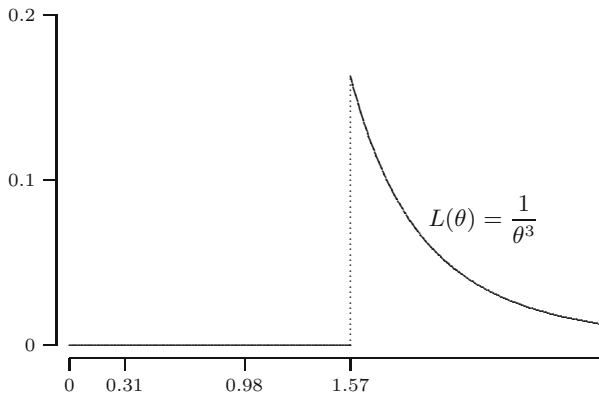
Fig. 21.1. Likelihood function $L(\theta)$ of a sample from a $U(0,\theta)$ distribution.

Figure 21.1 depicts this likelihood function. One glance at this figure is enough to realize that $L(\theta)$ attains its maximum at $\max(x_1, x_2, x_3) = 1.57$.

In general, given a dataset $x_1, x_2, \ldots, x_n$ originating from a $U(0,\theta)$ distribution, we see that $L(\theta) = 0$ if θ is smaller than at least one of the x_i and that $L(\theta) = 1/\theta^n$ if θ is greater than or equal to the largest of the x_i. We conclude that the maximum likelihood estimator of θ is given by $\max\{X_1, X_2, \ldots, X_n\}$.

Loglikelihood

In the preceding example it was easy to find the value of the parameter for which the likelihood is maximal. Usually one can find the maximum by differentiating the likelihood function $L(\theta)$. The calculation of the derivative of $L(\theta)$ may be tedious, because $L(\theta)$ is a product of terms, all involving θ (see also Quick exercise 21.3). To differentiate $L(\theta)$ we have to apply the product rule from calculus. Considering the logarithm of $L(\theta)$ changes the product of the terms involving θ into a *sum* of logarithms of these terms, which makes the process of differentiating easier. Moreover, because the logarithm is an increasing function, the likelihood function $L(\theta)$ and the *loglikelihood function* $\ell(\theta)$, defined by

$$\ell(\theta) = \ln(L(\theta)),$$

attain their extreme values for the same values of θ. In particular, $L(\theta)$ is maximal if and only if $\ell(\theta)$ is maximal. This is illustrated in Figure 21.2 by the likelihood function $L(p) = Cp^{93}(1-p)^{322}$ and the loglikelihood function $\ell(p) = \ln(C) + 93\ln(p) + 322\ln(1-p)$ for the smokers.

In the situation that we have a dataset $x_1, x_2, \ldots, x_n$ modeled as a realization of a random sample from an $Exp(\lambda)$ distribution, we found as likelihood function $L(\lambda) = \lambda^n \cdot e^{-\lambda(x_1+x_2+\cdots+x_n)}$. Therefore, the loglikelihood function is given by

$$\ell(\lambda) = n\ln(\lambda) - \lambda(x_1 + x_2 + \cdots + x_n).$$
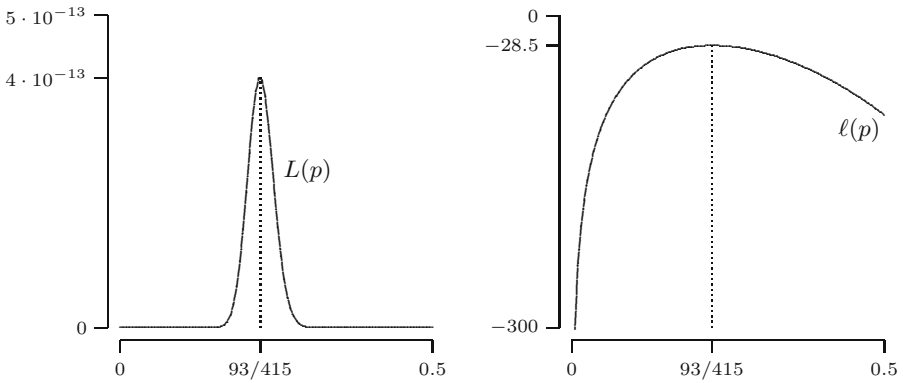
Fig. 21.2. The graphs of the likelihood function $L(p)$ and the loglikelihood function $\ell(p)$ for the smokers.

QUICK EXERCISE 21.3 In this example, use the loglikelihood function $\ell(\lambda)$ to show that the maximum likelihood estimate of λ equals $1/\bar{x}_n$.

Estimating the parameters of the normal distribution

Suppose that the dataset $x_1, x_2, \ldots, x_n$ is a realization of a random sample from an $N(\mu, \sigma^2)$ distribution, with μ and σ unknown. What are the maximum likelihood estimates for μ and σ?

In this case θ is the vector (μ, σ), and therefore the likelihood function is a function of two variables:

$$L(\mu, \sigma) = f_{\mu,\sigma}(x_1) f_{\mu,\sigma}(x_2) \cdots f_{\mu,\sigma}(x_n),$$

where each $f_{\mu,\sigma}(x)$ is the $N(\mu, \sigma^2)$ probability density function:

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

Since

$$\ln\left(f_{\mu,\sigma}(x)\right) = -\ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2,$$

one finds that

$$\ell(\mu, \sigma) = \ln\left(f_{\mu,\sigma}(x_1)\right) + \cdots + \ln\left(f_{\mu,\sigma}(x_n)\right)$$
$$= -n\ln(\sigma) - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\left((x_1 - \mu)^2 + \cdots + (x_n - \mu)^2\right).$$

The partial derivatives of ℓ are

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \left((x_1 - \mu) + (x_2 - \mu) + \cdots + (x_n - \mu) \right) = \frac{n}{\sigma^2} (\bar{x}_n - \mu)$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \left((x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2 \right)$$

$$= -\frac{n}{\sigma^3} \left(\sigma^2 - \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \right).$$

Solving $\dfrac{\partial \ell}{\partial \mu} = 0$ and $\dfrac{\partial \ell}{\partial \sigma} = 0$ yields

$$\mu = \bar{x}_n \quad \text{and} \quad \sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2}.$$

It is not hard to show that for these values of μ and σ the likelihood function $L(\mu, \sigma)$ attains a maximum. We find that $\bar{x}_n$ is the maximum likelihood estimate for μ and that

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

is the maximum likelihood estimate for σ.

21.4 Properties of maximum likelihood estimators

Apart from the fact that the maximum likelihood principle provides a general principle to construct estimators, one can also show that maximum likelihood estimators have several desirable properties.

Invariance principle

In the previous example, we saw that

$$D_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$$

is the maximum likelihood estimator for the parameter σ of an $N(\mu, \sigma^2)$ distribution. Does this imply that D_n^2 is the maximum likelihood estimator for σ^2? This is indeed the case! In general one can show that if T is the maximum likelihood estimator of a parameter θ and $g(\theta)$ is an invertible function of θ, then $g(T)$ is the maximum likelihood estimator for $g(\theta)$.

Asymptotic unbiasedness

The maximum likelihood estimator T may be biased. For example, because $D_n^2 = \frac{n-1}{n}S_n^2$, for the previously mentioned maximum likelihood estimator D_n^2 of the parameter σ^2 of an $N(\mu, \sigma^2)$ distribution, it follows from Section 19.4 that

$$\mathrm{E}\left[D_n^2\right] = \mathrm{E}\left[\frac{n-1}{n}S_n^2\right] = \frac{n-1}{n}\mathrm{E}\left[S_n^2\right] = \frac{n-1}{n}\sigma^2.$$

We see that D_n^2 is a biased estimator for σ^2, but also that as n goes to infinity, the expected value of D_n^2 converges to σ^2. This holds more generally. Under mild conditions on the distribution of the random variables X_i under consideration (see, e.g., [36]), one can show that asymptotically (that is, as the size n of the dataset goes to infinity) maximum likelihood estimators are unbiased. By this we mean that if $T_n = h(X_1, X_2, \ldots, X_n)$ is the maximum likelihood estimator for a parameter θ, then

$$\lim_{n\to\infty} \mathrm{E}\left[T_n\right] = \theta.$$

Asymptotic minimum variance

The variance of an unbiased estimator for a parameter θ is always larger than or equal to a certain positive number, known as the Cramér-Rao lower bound (see Remark 20.2). Again under mild conditions one can show that maximum likelihood estimators have asymptotically the smallest variance among unbiased estimators. That is, asymptotically the variance of the maximum likelihood estimator for a parameter θ attains the Cramér-Rao lower bound.

21.5 Solutions to the quick exercises

21.1 In the case that only the first three chips are defective, the probability that the observed data occur is equal to

$$\mathrm{P}(R_1 = 1, R_2 = 1, R_3 = 1, R_4 = 0, \ldots, R_{10} = 0) = p^3(1-p)^7.$$

For the batch where about 10% of the chips are defective we find that

$$\mathrm{P}(R_1 = 1, R_2 = 1, R_3 = 1, R_4 = 0, \ldots, R_{10} = 0) = \left(\frac{1}{10}\right)^3 \left(\frac{9}{10}\right)^7 = 0.00048,$$

whereas for the other batch this probability is equal to $\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 = 0.00098$. So the probability for the batch with about 50% defective chips is about 2 times larger than the probability for the other batch. In view of this, it would be reasonable to choose the other batch, not the tested one.

21.2 From Table 21.1 we derive

$$L(p) = \text{constant} \cdot \text{P}(X_i = 1)^{198} \, \text{P}(X_i = 2)^{107} \cdots \text{P}(X_i = 12)^6 \, \text{P}(X_i > 12)^{12}$$
$$= \text{constant} \cdot p^{198} \cdot \left[(1-p)p\right]^{107} \cdots \left[(1-p)^{11}p\right]^6 \cdot \left[(1-p)^{12}\right]^{12}$$
$$= \text{constant} \cdot p^{474} \cdot (1-p)^{955}.$$

Here the constant is the number of ways we can assign 198 ones, 107 twos, ..., 6 twelves, and 12 numbers larger than 12 to 486 nonsmokers. Differentiating $L(p)$ with respect to p yields that

$$L'(p) = \text{constant} \cdot \left[474p^{473}(1-p)^{955} - 955p^{474}(1-p)^{954}\right]$$
$$= \text{constant} \cdot p^{473}(1-p)^{954}\left[474(1-p) - 955p\right]$$
$$= \text{constant} \cdot p^{473}(1-p)^{954}(474 - 1429p).$$

Now $L'(p) = 0$ if $p = 0$, $p = 1$, or $p = 474/1429 = 0.33$, and $L(p)$ attains its unique maximum in this last point.

21.3 The loglikelihood function $L(\lambda)$ has derivative

$$\ell'(\lambda) = \frac{n}{\lambda} - (x_1 + x_2 + \cdots + x_n) = n\left(\frac{1}{\lambda} - \bar{x}_n\right).$$

One finds that $\ell'(\lambda) = 0$ if and only if $\lambda = 1/\bar{x}_n$ and that this is a maximum. The maximum likelihood estimate for λ is therefore $1/\bar{x}_n$.

21.6 Exercises

21.1 ⊞ Consider the following situation. Suppose we have two fair dice, D_1 with 5 red sides and 1 white side and D_2 with 1 red side and 5 white sides. We pick one of the dice randomly, and throw it repeatedly until *red* comes up for the first time. With the same die this experiment is repeated two more times. Suppose the following happens:

First experiment:	first red appears in 3rd throw
Second experiment:	first red appears in 5th throw
Third experiment:	first red appears in 4th throw.

Show that for die D_1 this happens with probability $5.7424 \cdot 10^{-8}$, and for die D_2 the probability with which this happens is $8.9725 \cdot 10^{-4}$. Given these probabilities, which die do you think we picked?

21.2 ⊡ We throw an unfair coin repeatedly until heads comes up for the first time. We repeat this experiment three times (with the same coin) and obtain the following data:

First experiment: heads first comes up in   3rd throw
Second experiment: heads first comes up in   5th throw
Third experiment: heads first comes up in   4th throw.

Let p be the probability that heads comes up in a throw with this coin. Determine the maximum likelihood estimate $\hat{p}$ of p.

21.3 In Exercise 17.4 we modeled the hits of London by flying bombs by a Poisson distribution with parameter μ.

a. Use the data from Exercise 17.4 to find the maximum likelihood estimate of μ.

b. Suppose the summarized data from Exercise 17.4 got corrupted in the following way:

Number of hits	0 or 1	2	3	4	5	6	7
Number of squares	440	93	35	7	0	0	1

Using this new data, what is the maximum likelihood estimate of μ?

21.4 ⊞ In Section 19.1, we considered the arrivals of packages at a network server, where we modeled the number of arrivals per minute by a $Pois(\mu)$ distribution. Let $x_1, x_2, \ldots, x_n$ be a realization of a random sample from a $Pois(\mu)$ distribution. We saw on page 286 that a natural estimate of the probability of zeros in the dataset is given by

$$\frac{\text{number of } x_i \text{ equal to zero}}{n}.$$

a. Show that the likelihood $L(\mu)$ is given by

$$L(\mu) = \frac{e^{-n\mu}}{x_1! \cdots x_n!} \mu^{x_1 + x_2 + \cdots + x_n}.$$

b. Determine the loglikelihood $\ell(\mu)$ and the formula of the maximum likelihood estimate for μ.

c. What is the maximum likelihood estimate for the probability $e^{-\mu}$ of zero arrivals?

21.5 ⊡ Suppose that $x_1, x_2, \ldots, x_n$ is a dataset, which is a realization of a random sample from a normal distribution.

a. Let the probability density of this normal distribution be given by

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} \quad \text{for } -\infty < x < \infty.$$

Determine the maximum likelihood estimate for μ.

b. Now suppose that the density of this normal distribution is given by

$$f_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}x^2/\sigma^2} \quad \text{for } -\infty < x < \infty.$$

Determine the maximum likelihood estimate for σ.

21.6 Let $x_1, x_2, \ldots, x_n$ be a dataset that is a realization of a random sample from a distribution with probability density $f_\delta(x)$ given by

$$f_\delta(x) = \begin{cases} e^{-(x-\delta)} & \text{for } x \geq \delta \\ 0 & \text{for } x < \delta. \end{cases}$$

a. Draw the likelihood $L(\delta)$.

b. Determine the maximum likelihood estimate for δ.

21.7 □ Suppose that $x_1, x_2, \ldots, x_n$ is a dataset, which is a realization of a random sample from a Rayleigh distribution, which is a continuous distribution with probability density function given by

$$f_\theta(x) = \frac{x}{\theta^2}e^{-\frac{1}{2}x^2/\theta^2} \quad \text{for } x \geq 0.$$

In this case what is the maximum likelihood estimate for θ?

21.8 ⊞ (Exercises 19.7 and 20.7 continued) A certain type of plant can be divided into four types: starchy-green, starchy-white, sugary-green, and sugary-white. The following table lists the counts of the various types among 3839 leaves.

Type	Count
Starchy-green	1997
Sugary-white	32
Starchy-white	906
Sugary-green	904

Setting

$$X = \begin{cases} 1 & \text{if the observed leave is of type starchy-green} \\ 2 & \text{if the observed leave is of type sugary-white} \\ 3 & \text{if the observed leave is of type starchy-white} \\ 4 & \text{if the observed leave is of type sugary-green,} \end{cases}$$

the probability mass function p of X is given by

a	1	2	3	4
$p(a)$	$\frac{1}{4}(2+\theta)$	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$

and $p(a) = 0$ for all other a. Here $0 < \theta < 1$ is an unknown parameter, which was estimated in Exercise 19.7. We want to find a maximum likelihood estimate of θ.

a. Use the data to find the likelihood $L(\theta)$ and the loglikelihood $\ell(\theta)$.

b. What is the maximum likelihood estimate of θ using the data from the preceding table?

c. Suppose that we have the counts of n different leaves: n_1 of type starchy-green, n_2 of type sugary-white, n_3 of type starchy-white, and n_4 of type sugary-green (so $n = n_1 + n_2 + n_3 + n_4$). Determine the general formula for the maximum likelihood estimate of θ.

21.9 ☐ Let $x_1, x_2, \ldots, x_n$ be a dataset that is a realization of a random sample from a $U(\alpha, \beta)$ distribution (with α and β unknown, $\alpha < \beta$). Determine the maximum likelihood estimates for α and β.

21.10 Let $x_1, x_2, \ldots, x_n$ be a dataset, which is a realization of a random sample from a $Par(\alpha)$ distribution. What is the maximum likelihood estimate for α?

21.11 ⊞ In Exercise 4.13 we considered the situation where we have a box containing an unknown number—say N—of identical bolts. In order to get an idea of the size of N we introduced three random variables X, Y, and Z. Here we will use X and Y, and in the next exercise Z, to find maximum likelihood estimates of N.

a. Suppose that $x_1, x_2, \ldots, x_n$ is a dataset, which is a realization of a random sample from a $Geo(1/N)$ distribution. Determine the maximum likelihood estimate for N.

b. Suppose that $y_1, y_2, \ldots, y_n$ is a dataset, which is a realization of a random sample from a discrete uniform distribution on $1, 2, \ldots, N$. Determine the maximum likelihood estimate for N.

21.12 (Exercise 21.11 continued.) Suppose that m bolts in the box were marked and then r bolts were selected from the box; Z is the number of marked bolts in the sample. (Recall that it was shown in Exercise 4.13 **c** that Z has a hypergeometric distribution, with parameters m, N, and r.) Suppose that k bolts in the sample were marked. Show that the likelihood $L(N)$ is given by

$$L(N) = \frac{\binom{m}{k}\binom{N-m}{r-k}}{\binom{N}{r}}.$$

Next show that $L(N)$ increases for $N < mr/k$ and decreases for $N > mr/k$, and conclude that mr/k is the maximum likelihood estimate for N.

21.13 Often one can model the times that customers arrive at a shop rather well by a Poisson process with (unknown) rate λ (customers/hour). On a certain day, one of the attendants noticed that between noon and 12.45 p.m.

two customers arrived, and another attendant noticed that on the same day one customer arrived between 12.15 and 1 p.m. Use the observations of the attendants to determine the maximum likelihood estimate of λ.

21.14 A very inexperienced archer shoots n times an arrow at a disc of (unknown) radius θ. The disc is hit every time, but at completely random places. Let $r_1, r_2, \ldots, r_n$ be the distances of the various hits to the center of the disc. Determine the maximum likelihood estimate for θ.

21.15 On January 28, 1986, the main fuel tank of the space shuttle *Challenger* exploded shortly after takeoff. Essential in this accident was the leakage of some of the six O-rings of the *Challenger*. In Section 1.4 the probability of failure of an O-ring is given by

$$p(t) = \frac{e^{a+b \cdot t}}{1 + e^{a+b \cdot t}},$$

where t is the temperature at launch in degrees Fahrenheit. In Table 21.2 the temperature t (in °F, rounded to the nearest integer) and the number of failures N for 23 missions are given, ordered according to increasing temperatures. (See also Figure 1.3, where these data are graphically depicted.) Give the likelihood $L(a, b)$ and the loglikelihood $\ell(a, b)$.

Table 21.2. Space shuttle failure data of pre-*Challenger* missions.

t	53	57	58	63	66	67	67	67
N	2	1	1	1	0	0	0	0

t	68	69	70	70	70	70	72	73
N	0	0	0	0	1	1	0	0

t	75	75	76	76	78	79	81
N	0	2	0	0	0	0	0

21.16 In the 18th century Georges-Louis Leclerc, Comte de Buffon (1707–1788) found an amusing way to approximate the number π using probability theory and statistics. Buffon had the following idea: take a needle and a large sheet of paper, and draw horizontal lines that are a needle-length apart. Throw the needle a number of times (say n times) on the sheet, and count how often it hits one of the horizontal lines. Say this number is s_n, then s_n is the realization of a $Bin(n, p)$ distributed random variable S_n. Here p is the probability that the needle hits one of the horizontal lines. In Exercise 9.20 you found that $p = 2/\pi$. Show that

$$T = \frac{2n}{S_n}$$

is the maximum likelihood estimator for π.

The method of least squares

The maximum likelihood principle provides a way to estimate parameters. The applicability of the method is quite general but not universal. For example, in the simple linear regression model, introduced in Section 17.4, we need to know the distribution of the response variable in order to find the maximum likelihood estimates for the parameters involved. In this chapter we will see how these parameters can be estimated using the method of least squares. Furthermore, the relation between least squares and maximum likelihood will be investigated in the case of normally distributed errors.

22.1 Least squares estimation and regression

Recall from Section 17.4 the simple linear regression model for a bivariate dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. In this model $x_1, x_2, \ldots, x_n$ are non-random and $y_1, y_2, \ldots, y_n$ are realizations of random variables $Y_1, Y_2, \ldots, Y_n$ satisfying

$$Y_i = \alpha + \beta x_i + U_i \qquad \text{for} \quad i = 1, 2, \ldots, n,$$

where $U_1, U_2, \ldots, U_n$ are independent random variables with zero expectation and variance σ^2. How can one obtain estimates for the parameters α, β, and σ^2 in this model?

Note that we cannot find maximum likelihood estimates for these parameters, simply because we have no further knowledge about the distribution of the U_i (and consequently of the Y_i). We want to choose α and β in such a way that we obtain a line that fits the data best. A classical approach to do this is to consider the sum of squared distances between the observed values y_i and the values $\alpha + \beta x_i$ on the regression line $y = \alpha + \beta x$. See Figure 22.1, where these distances are indicated. The *method of least squares* prescribes to choose α and β such that the sum of squares

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$
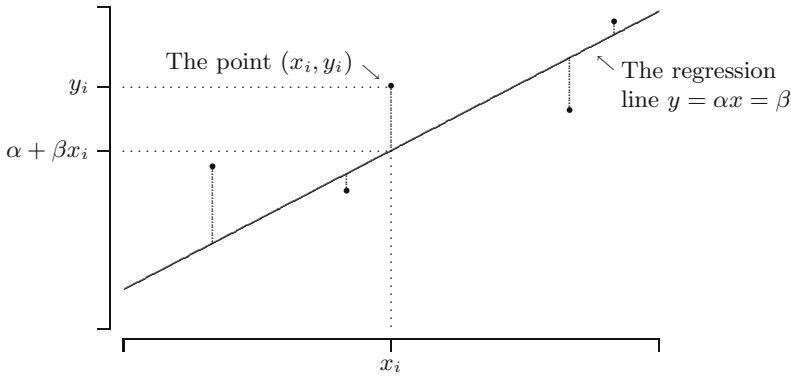
Fig. 22.1. The observed value y_i corresponding to x_i and the value $\alpha + \beta x_i$ on the regression line $y = \alpha + \beta x$.

is minimal. The ith term in the sum is the squared distance in the vertical direction from (x_i, y_i) to the line $y = \alpha + \beta x$. To find these so-called *least squares estimates*, we differentiate $S(\alpha, \beta)$ with respect to α and β, and we set the derivatives equal to 0:

$$\frac{\partial}{\partial \alpha} S(\alpha, \beta) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial}{\partial \beta} S(\alpha, \beta) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0.$$

This is equivalent to

$$n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

For example, for the timber data from Table 15.5 we would obtain

$$36\,\alpha + 1646.4\,\beta = 52\,901$$

$$1646.4\,\alpha + 81750.02\,\beta = 2\,790\,525.$$

These are two equations with two unknowns α and β. Solving for α and β yields the solutions $\hat{\alpha} = -1160.5$ and $\hat{\beta} = 57.51$. In Figure 22.2 a scatterplot of the timber dataset, together with the estimated regression line $y = -1160.5 + 57.51x$, is depicted.

QUICK EXERCISE 22.1 Suppose you are given a piece of Australian timber with density 65. What would you choose as an estimate for the Janka hardness?
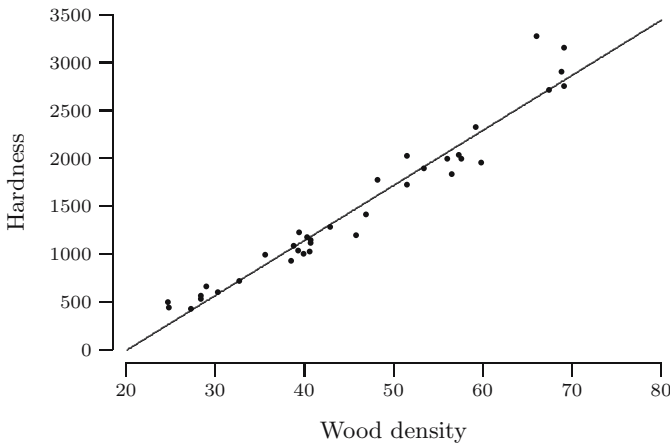
Fig. 22.2. Scatterplot and estimated regression line for the timber data.

In general, writing $\sum$ instead of $\sum_{i=1}^{n}$, we find the following formulas for the estimates $\hat{\alpha}$ (the *intercept*) and $\hat{\beta}$ (the *slope*):

$$\hat{\beta} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} \tag{22.1}$$

$$\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n. \tag{22.2}$$

Since $S(\alpha, \beta)$ is an elliptic paraboloid (a "vase"), it follows that $(\hat{\alpha}, \hat{\beta})$ is the unique minimum of $S(\alpha, \beta)$ (except when all x_i are equal).

QUICK EXERCISE 22.2 Check that the line $y = \hat{\alpha} + \hat{\beta}x$ always passes through the "center of gravity" $(\bar{x}_n, \bar{y}_n)$.

Least squares estimators are unbiased

We denote the least squares *estimates* by $\hat{\alpha}$ and $\hat{\beta}$. It is quite common to also denote the least squares *estimators* by $\hat{\alpha}$ and $\hat{\beta}$:

$$\hat{\alpha} = \bar{Y}_n - \hat{\beta}\bar{x}_n, \qquad \hat{\beta} = \frac{n \sum x_i Y_i - (\sum x_i)(\sum Y_i)}{n \sum x_i^2 - (\sum x_i)^2}.$$

In Exercise 22.12 it is shown that $\hat{\beta}$ is an unbiased estimator for β. Using this and the fact that $\mathrm{E}[Y_i] = \alpha + \beta x_i$ (see page 258), we find for $\hat{\alpha}$:

$$\mathrm{E}[\hat{\alpha}] = \mathrm{E}\left[\bar{Y}_n\right] - \bar{x}_n \mathrm{E}\left[\hat{\beta}\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[Y_i] - \bar{x}_n \beta$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) - \bar{x}_n \beta = \alpha + \beta\bar{x}_n - \bar{x}_n\beta$$

$$= \alpha.$$

We see that $\hat{\alpha}$ is an unbiased estimator for α.

An unbiased estimator for σ^2

In the simple linear regression model the assumptions imply that the random variables Y_i are independent with variance σ^2. Unfortunately, one cannot apply the usual estimator $(1/(n-1))\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)^2$ for the variance of the Y_i (see Section 19.4), because different Y_i have different expectations. What would be a reasonable estimator for σ^2? The following quick exercise suggests a candidate.

QUICK EXERCISE 22.3 Let $U_1, U_2, \ldots, U_n$ be independent random variables, each with expected value zero and variance σ^2. Show that

$$T = \frac{1}{n}\sum_{i=1}^{n} U_i^2$$

is an unbiased estimator for σ^2.

At first sight one might be tempted to think that the unbiased estimator T from this quick exercise is a useful tool to estimate σ^2. Unfortunately, we only observe the x_i and Y_i, not the U_i. However, from the fact that $U_i = Y_i - \alpha - \beta x_i$, it seems reasonable to try

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \tag{22.3}$$

as an estimator for σ^2. Tedious calculations show that the expected value of this random variable equals $\frac{n-2}{n}\sigma^2$. But then we can easily turn it into an unbiased estimator for σ^2.

> AN UNBIASED ESTIMATOR FOR σ^2. In the simple linear regression model the random variable
>
> $$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$
>
> is an unbiased estimator for σ^2.

22.2 Residuals

A way to explore whether the simple linear regression model is appropriate to model a given bivariate dataset is to inspect a scatterplot of the so-called *residuals* r_i against the x_i. The ith residual r_i is defined as the vertical distance between the ith point and the estimated regression line:

$$r_i = y_i - \hat{\alpha} - \hat{\beta}x_i, \qquad i = 1, 2, \ldots, n.$$

When a linear model is appropriate, the scatterplot of the residuals r_i against the x_i should show truly random fluctuations around zero, in the sense that it should not exhibit any trend or pattern. This seems to be the case in Figure 22.3, which shows the residuals for the black cherry tree data from Exercise 17.9.
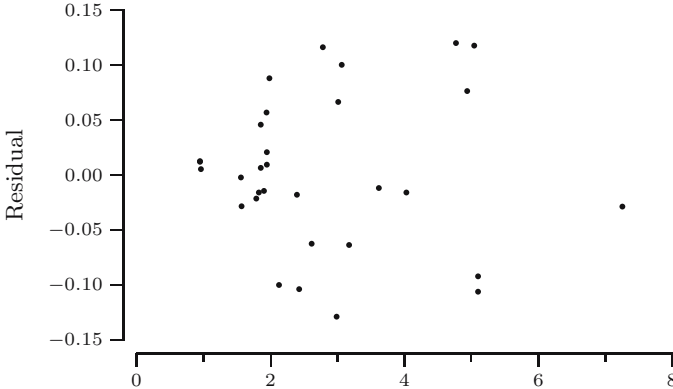
Fig. 22.3. Scatterplot of r_i versus x_i for the black cherry tree data.

QUICK EXERCISE 22.4 Recall from Quick exercise 22.2 that $(\bar{x}_n, \bar{y}_n)$ is on the regression line $y = \hat{\alpha} + \hat{\beta}x$, i.e., that $\bar{y}_n = \hat{\alpha} + \hat{\beta}\bar{x}_n$. Use this to show that $\sum_{i=1}^{n} r_i = 0$, i.e., that the sum of the residuals is zero.

In Figure 22.4 we depicted r_i versus x_i for the timber dataset. In this case a slight parabolic pattern can be observed. Figures 22.2 and 22.4 suggest that
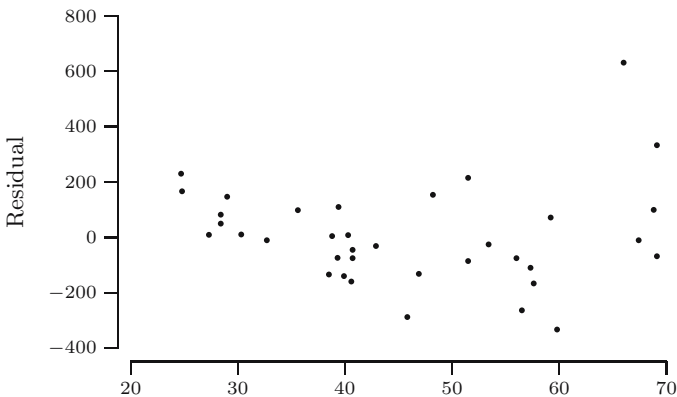
Fig. 22.4. Scatterplot of r_i versus x_i for the timber data with the simple linear regression model $Y_i = \alpha + \beta x_i + U_i$.

for the timber dataset a better model might be

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + U_i \quad \text{for} \quad i = 1, 2, \ldots, n.$$

In this new model the residuals are

$$r_i = y_i - \hat{\alpha} - \hat{\beta} x_i - \hat{\gamma} x_i^2,$$

where $\hat{\alpha}, \hat{\beta}$, and $\hat{\gamma}$ are the least squares estimates obtained by minimizing

$$\sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i - \gamma x_i^2 \right)^2.$$

In Figure 22.5 we depicted r_i versus x_i. The residuals display no trend or pattern, except that they "fan out"—an example of a phenomenon called *heteroscedasticity*.
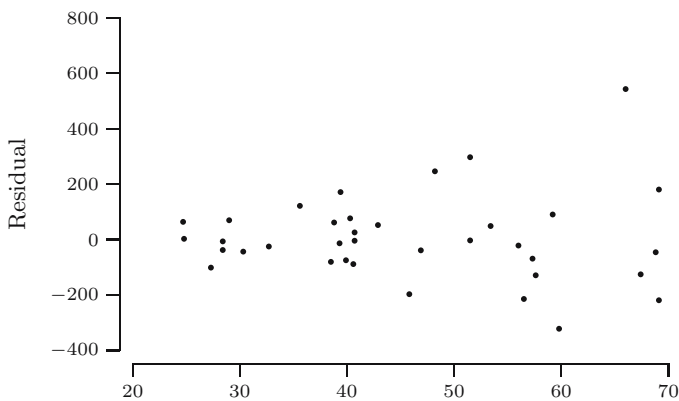
Fig. 22.5. Scatterplot of r_i versus x_i for the timber data with the model $Y_i = \alpha + \beta x_i + \gamma x_i^2 + U_i$.

Heteroscedasticity

The assumption of equal variance of the U_i (and therefore of the Y_i) is called *homoscedasticity*. In case the variance of Y_i depends on the value of x_i, we speak of *heteroscedasticity*. For instance, heteroscedasticity occurs when Y_i with a large expected value have a larger variance than those with small expected values. This produces a "fanning out" effect, which can be observed in Figure 22.5. This figure strongly suggests that the timber data are heteroscedastic. Possible ways out of this problem are a technique called weighted least squares or the use of variance-stabilizing transformations.

22.3 Relation with maximum likelihood

To apply the method of least squares no assumption is needed about the type of distribution of the U_i. In case the type of distribution of the U_i is known, the maximum likelihood principle can be applied. Consider, for instance, the classical situation where the U_i are independent with an $N(0, \sigma^2)$ distribution. What are the maximum likelihood estimates for α and β?

In this case the Y_i are independent, and Y_i has an $N(\alpha + \beta x_i, \sigma^2)$ distribution. Under these assumptions and assuming that the linear model is appropriate to model a given bivariate dataset, the r_i should look like the realization of a random sample from a normal distribution. As an example a histogram of the residuals r_i of the cherry tree data of Exercise 17.9 is depicted in Figure 22.6.
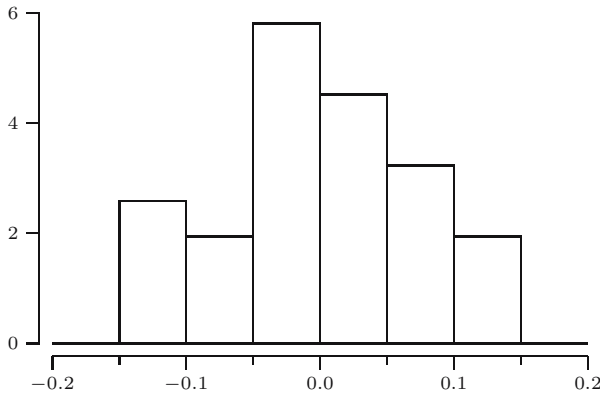
Fig. 22.6. Histogram of the residuals r_i for the black cherry tree data.

The data do not exhibit strong evidence against the assumption of normality. When Y_i has an $N(\alpha + \beta x_i, \sigma^2)$ distribution, the probability density of Y_i is given by

$$f_i(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\alpha-\beta x_i)^2/(2\sigma^2)} \qquad \text{for} \quad -\infty < y < \infty.$$

Since

$$\ln(f_i(y_i)) = -\ln(\sigma) - \ln(\sqrt{2\pi}) - \frac{1}{2}\left(\frac{y_i - \alpha - \beta x_i}{\sigma}\right)^2,$$

the loglikelihood is:

$$\ell(\alpha, \beta, \sigma) = \ln(f_1(y_1)) + \cdots + \ln(f_n(y_n))$$
$$= -n\ln(\sigma) - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

Note that for any fixed $\sigma > 0$, the loglikelihood $\ell(\alpha, \beta, \sigma)$ attains its maximum precisely when $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ is minimal. Hence, in case the U_i are independent with an $N(0, \sigma^2)$ distribution, the maximum likelihood principle and the least squares method yield the *same* estimators.

To find the maximum likelihood estimate of σ we differentiate $\ell(\alpha, \beta, \sigma)$ with respect to σ:

$$\frac{\partial}{\partial \sigma}\ell(\alpha, \beta, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

It follows (from the invariance principle on page 321) that the maximum likelihood estimator of σ^2 is given by

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2,$$

which is the estimator from (22.3).

22.4 Solutions to the quick exercises

22.1 We can use the estimated regression line $y = -1160.5 + 57.51x$ to predict the Janka hardness. For density $x = 65$ we find as a prediction for the Janka hardness $y = 2577.65$.

22.2 Rewriting $\hat{\alpha} = \bar{y}_n - \hat{\beta}$, it follows that $\bar{y}_n = \hat{\alpha} + \hat{\beta}\bar{x}_n$, which means that $(\bar{x}_n, \bar{y}_n)$ is a point on the estimated regression line $y = \hat{\alpha} + \hat{\beta}x$.

22.3 We need to show that $E[T] = \sigma^2$. Since $E[U_i] = 0$, $\text{Var}(U_i) = E[U_i^2]$, so that:

$$E[T] = E\left[\frac{1}{n}\sum_{i=1}^{n}U_i^2\right] = \frac{1}{n}\sum_{i=1}^{n}E[U_i^2] = \frac{1}{n}\sum_{i=1}^{n}\text{Var}(U_i) = \sigma^2.$$

22.4 Since $r_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$ for $i = 1, 2, \ldots, n$, it follows that the sum of the residuals equals

$$\sum r_i = \sum y_i - \left(n\hat{\alpha} + \hat{\beta}\sum x_i\right)$$
$$= n\bar{y}_n - \left(n\hat{\alpha} + n\hat{\beta}\bar{x}_n\right) = n\left(\bar{y}_n - (\hat{\alpha} + \hat{\beta}\bar{x}_n)\right) = 0,$$

because $\bar{y}_n = \hat{\alpha} + \hat{\beta}\bar{x}_n$, according to Quick exercise 22.2.

22.5 Exercises

22.1 ⊞ Consider the following bivariate dataset:

$$(1, 2) \quad (3, 1.8) \quad (5, 1).$$

a. Determine the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters of the regression line $y = \alpha + \beta x$.

b. Determine the residuals r_1, r_2, and r_3 and check that they add up to 0.

c. Draw in one figure the scatterplot of the data and the estimated regression line $y = \hat{\alpha} + \hat{\beta} x$.

22.2 Adding one point may dramatically change the estimates of α and β. Suppose one extra datapoint is added to the dataset of the previous exercise and that we have as dataset:

$$(0, 0) \quad (1, 2) \quad (3, 1.8) \quad (5, 1).$$

Determine the least squares estimate of $\hat{\beta}$. A point such as $(0, 0)$, which dramatically changes the estimates for α and β, is called a *leverage point*.

22.3 Suppose we have the following bivariate dataset:

$$(1, 3.1) \quad (1.7, 3.9) \quad (2.1, 3.8) \quad (2.5, 4.7) \quad (2.7, 4.5).$$

a. Determine the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters of the regression line $y = \alpha + \beta x$. You may use that $\sum x_i = 10$, $\sum y_i = 20$, $\sum x_i^2 = 21.84$, and $\sum x_i y_i = 41.61$.

b. Draw in one figure the scatterplot of the data and the estimated regression line $y = \hat{\alpha} + \hat{\beta} x$.

22.4 We are given a bivariate dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_{100}, y_{100})$. For this bivariate dataset it is known that $\sum x_i = 231.7$, $\sum x_i^2 = 2400.8$, $\sum y_i = 321$, and $\sum x_i y_i = 5189$. What are the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters of the regression line $y = \alpha + \beta x$?

22.5 ⊞ For the timber dataset it seems reasonable to leave out the intercept α ("no hardness without density"). The model then becomes

$$Y_i = \beta x_i + U_i \quad \text{for} \quad i = 1, 2, \ldots, n.$$

Show that the least squares estimator $\hat{\beta}$ of β is now given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$$

by minimizing the appropriate sum of squares.

22.6 ☐ (Quick exercise 22.1 and Exercise 22.5 continued). Suppose we are given a piece of Australian timber with density 65. What would you choose as an estimate for the Janka hardness, based on the regression model with no intercept? Recall that $\sum x_i y_i = 2790525$ and $\sum x_i^2 = 81750.02$ (see also Section 22.1).

22.7 Consider the dataset

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

where $x_1, x_2, \ldots, x_n$ are nonrandom and $y_1, y_2, \ldots, y_n$ are realizations of random variables $Y_1, Y_2, \ldots, Y_n$, satisfying

$$Y_i = e^{\alpha + \beta x_i} + U_i \qquad \text{for} \quad i = 1, 2, \ldots, n.$$

Here $U_1, U_2, \ldots, U_n$ are independent random variables with zero expectation and variance σ^2. What are the least squares estimates for the parameters α and β in this model?

22.8 ☐ Which simple regression model has the larger *residual sum of squares* $\sum_{i=1}^n r_i^2$, the model with intercept or the one without?

22.9 For some datasets it seems reasonable to leave out the slope β. For example, in the jury example from Section 6.3 it was assumed that the score that juror i assigns when the performance deserves a score g is $Y_i = g + Z_i$, where Z_i is a random variable with values around zero. In general, when the slope β is left out, the model becomes

$$Y_i = \alpha + U_i \quad \text{for } i = 1, 2, \ldots, n.$$

Show that $\bar{Y}_n$ is the least squares estimator $\hat{\alpha}$ of α.

22.10 ☐ In the method of least squares we choose α and β in such a way that the sum of squared residuals $S(\alpha, \beta)$ is minimal. Since the ith term in this sum is the squared vertical distance from (x_i, y_i) to the regression line $y = \alpha + \beta x$, one might also wonder whether it is a good idea to replace this squared distance simply by the distance. So, given a bivariate dataset

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

choose α and β in such a way that the sum

$$A(\alpha, \beta) = \sum_{i=1}^n |y_i - \alpha - \beta x_i|$$

is minimal. We will investigate this by a simple example. Consider the following bivariate dataset:

$$(0, 2), (1, 2), (2, 0).$$

a. Determine the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$, and draw in one figure the scatterplot of the data and the estimated regression line $y = \hat{\alpha} + \hat{\beta}x$. Finally, determine $A(\hat{\alpha}, \hat{\beta})$.

b. One might wonder whether $\hat{\alpha}$ and $\hat{\beta}$ also minimize $A(\alpha, \beta)$. To investigate this, choose $\beta = -1$ and find α's for which $A(\alpha, -1) < A(\hat{\alpha}, \hat{\beta})$. For which α is $A(\alpha, -1)$ minimal?

c. Find α and β for which $A(\alpha, \beta)$ is minimal.

22.11 Consider the dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where the x_i are nonrandom and the y_i are realizations of random variables $Y_1, Y_2, \ldots, Y_n$ satisfying
$$Y_i = g(x_i) + U_i \quad \text{for } i = 1, 2, \ldots, n,$$
where $U_1, U_2, \ldots, U_n$ are independent random variables with zero expectation and variance σ^2. Visual inspection of the scatterplot of our dataset in
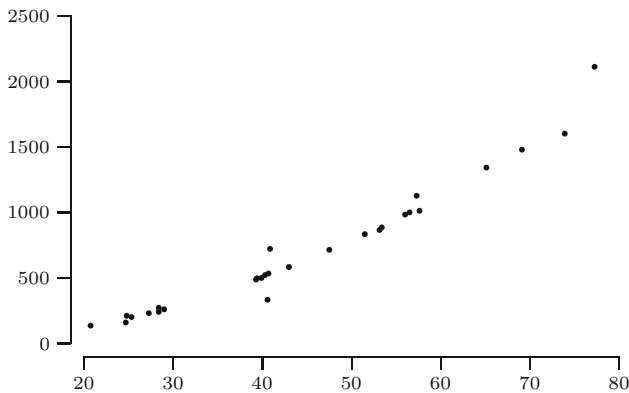
Fig. 22.7. Scatterplot of y_i versus x_i.

Figure 22.7 suggests that we should model the Y_i by
$$Y_i = \beta x_i + \gamma x_i^2 + U_i \quad \text{for } i = 1, 2, \ldots, n.$$

a. Show that the least squares estimators $\hat{\beta}$ and $\hat{\gamma}$ satisfy
$$\beta \sum x_i^2 + \gamma \sum x_i^3 = \sum x_i y_i,$$
$$\beta \sum x_i^3 + \gamma \sum x_i^4 = \sum x_i^2 y_i.$$

b. Infer from **a**—for instance, by using linear algebra—that the estimators $\hat{\beta}$ and $\hat{\gamma}$ are given by
$$\hat{\beta} = \frac{(\sum x_i Y_i)(\sum x_i^4) - (\sum x_i^3)(\sum x_i^2 Y_i)}{(\sum x_i^2)(\sum x_i^4) - (\sum x_i^3)^2}$$

and
$$\hat{\gamma} = \frac{(\sum x_i^2)(\sum x_i^2 Y_i) - (\sum x_i^3)(\sum x_i Y_i)}{(\sum x_i^2)(\sum x_i^4) - (\sum x_i^3)^2}.$$

22.12 ⊞ The least square estimator $\hat{\beta}$ from (22.1) is an unbiased estimator for β. You can show this in four steps.

a. First show that
$$\text{E}\left[\hat{\beta}\right] = \frac{n \sum x_i \text{E}[Y_i] - (\sum x_i)(\sum \text{E}[Y_i])}{n \sum x_i^2 - (\sum x_i)^2}.$$

b. Next use that $\text{E}[Y_i] = \alpha + \beta x_i$, to obtain that
$$\text{E}\left[\hat{\beta}\right] = \frac{n \sum x_i(\alpha + \beta x_i) - (\sum x_i)[n\alpha + \beta \sum x_i]}{n \sum x_i^2 - (\sum x_i)^2}.$$

c. Simplify this last expression to find
$$\text{E}\left[\hat{\beta}\right] = \frac{n\alpha \sum x_i + n\beta \sum x_i^2 - n\alpha \sum x_i - \beta(\sum x_i)^2}{n \sum x_i^2 - (\sum x_i)^2}.$$

d. Finally, conclude that $\hat{\beta}$ is an unbiased estimator for β.

23

Confidence intervals for the mean

Sometimes, a *range* of plausible values for an unknown parameter is preferred to a single estimate. We shall discuss how to turn data into what are called *confidence intervals* and show that this can be done in such a manner that definite statements can be made about *how* confident we are that the true parameter value is in the reported interval. This level of confidence is something you can choose. We start this chapter with the general principle of confidence intervals. We continue with confidence intervals for the mean, the common way to refer to confidence intervals made for the expected value of the model distribution. Depending on the situation, one of the four methods presented will apply.

23.1 General principle

In previous chapters we have encountered sample statistics as estimators for distribution features. This started somewhat informally in Chapter 17, where it was claimed, for example, that the sample mean and the sample variance are usually close to μ and σ^2 of the underlying distribution. Bias and MSE of estimators, discussed in Chapters 19 and 20, are used to judge the quality of estimators. If we have at our disposal an estimator T for an unknown parameter θ, we use its realization t as our estimate for θ. For example, when collecting data on the speed of light, as Michelson did (see Section 13.1), the unknown speed of light would be the parameter θ, our estimator T could be the sample mean, and Michelson's data then yield an estimate t for θ of 299 852.4 km/sec. We call this number a *point estimate*: if we are required to select *one* number, this is it. Had the measurements started a day earlier, however, the whole experiment would in essence be the same, but the results might have been different. Hence, we cannot say that the estimate *equals* the speed of light but rather that it is *close to* the true speed of light. For example, we could say something like: "we have great confidence that the true speed of

light is somewhere between ... and ... ." In addition to providing an interval of plausible values for θ we would want to add a specific statement about *how* confident we are that the true θ is among them.

In this chapter we shall present methods to make *confidence statements* about unknown parameters, based on knowledge of the sampling distributions of corresponding estimators. To illustrate the main idea, suppose the estimator T is unbiased for the speed of light θ. For the moment, also suppose that T has standard deviation $\sigma_T = 100$ km/sec (we shall drop this unrealistic assumption shortly). Then, applying formula (13.1), which was derived from Chebyshev's inequality (see Section 13.2), we find

$$P(|T - \theta| < 2\sigma_T) \geq \tfrac{3}{4}. \tag{23.1}$$

In words this reads: with probability at least 75%, the estimator T is within $2\sigma_T = 200$ of the true speed of light θ. We could rephrase this as

$$T \in (\theta - 200, \theta + 200) \quad \text{with probability at least 75\%.}$$

However, if I am near the city of Paris, then the city of Paris is near me: the statement "T is within 200 of θ" is the same as "θ is within 200 of T," and we could equally well rephrase (23.1) as

$$\theta \in (T - 200, T + 200) \quad \text{with probability at least 75\%.}$$

Note that of the last two equations the first is a statement about a *random variable T* being in a *fixed interval*, whereas in the second equation the *interval is random* and the statement is about the probability that the random interval covers the *fixed* but unknown θ. The interval $(T - 200, T + 200)$ is sometimes called an *interval estimator*, and its realization is an *interval estimate*.

Evaluating T for the Michelson data we find as its realization $t = 299\,852.4$, and this yields the statement

$$\theta \in (299\,652.4, 300\,052.4). \tag{23.2}$$

Because we substituted the realization for the random variable, we cannot claim that (23.2) holds with probability at least 75%: either the true speed of light θ belongs to the interval or it does not; the statement we make is either true or false, we just do not know which. However, because the procedure guarantees a probability of at least 75% of getting a "right" statement, we say:

$$\theta \in (299\,652.4, 300\,052.4) \quad \text{with confidence at least 75\%.} \tag{23.3}$$

The construction of this *confidence interval* only involved an unbiased estimator and knowledge of its standard deviation. When more information on the sampling distribution of the estimator is available, more refined statements can be made, as we shall see shortly.

QUICK EXERCISE 23.1 Repeat the preceding derivation, starting from the statement $P(|T - \theta| < 3\sigma_T) \geq 8/9$ (check that this follows from Chebyshev's inequality). What is the resulting confidence interval for the speed of light, and what is the corresponding confidence?

A general definition

Many confidence intervals are of the form[1]

$$(t - c \cdot \sigma_T,\ t + c \cdot \sigma_T)$$

we just encountered, where c is a number near 2 or 3. The corresponding confidence is often much higher than in the preceding example. Because there are many other ways confidence intervals can (or have to) be constructed, the general definition looks a bit different.

CONFIDENCE INTERVALS. Suppose a dataset $x_1, \ldots, x_n$ is given, modeled as realization of random variables $X_1, \ldots, X_n$. Let θ be the parameter of interest, and γ a number between 0 and 1. If there exist sample statistics $L_n = g(X_1, \ldots, X_n)$ and $U_n = h(X_1, \ldots, X_n)$ such that

$$P(L_n < \theta < U_n) = \gamma$$

for every value of θ, then

$$(l_n, u_n),$$

where $l_n = g(x_1, \ldots, x_n)$ and $u_n = h(x_1, \ldots, x_n)$, is called a $100\gamma\%$ *confidence interval for θ*. The number γ is called the *confidence level*.

Sometimes sample statistics L_n and U_n as required in the definition do not exist, but one *can* find L_n and U_n that satisfy

$$P(L_n < \theta < U_n) \geq \gamma.$$

The resulting confidence interval (l_n, u_n) is called a *conservative* $100\gamma\%$ confidence interval for θ: the actual confidence level might be higher. For example, the interval in (23.2) is a conservative 75% confidence interval.

QUICK EXERCISE 23.2 Why is the interval in (23.2) a *conservative* 75% confidence interval?

There is no way of knowing whether an individual confidence interval is correct, in the sense that it indeed *does* cover θ. The procedure guarantees that each time we make a confidence interval we have probability γ of covering θ. What this means in practice can easily be illustrated with an example, using simulation:

---

[1] Another form is, for example, $(c_1 t, c_2 t)$.

Generate $x_1, \ldots, x_{20}$ from an $N(0,1)$ distribution. Next, pretend that it is known that the data are from a normal distribution but that both μ and σ are unknown. Construct the 90% confidence interval for the expectation μ using the method described in the next section, which says to use (l_n, u_n) with

$$l_n = \bar{x}_{20} - 1.729 \frac{s_{20}}{\sqrt{20}} \qquad u_n = \bar{x}_{20} + 1.729 \frac{s_{20}}{\sqrt{20}},$$

where $\bar{x}_{20}$ and s_{20} are the sample mean and standard deviation. Finally, check whether the "true μ," in this case 0, is in the confidence interval.

We repeated the whole procedure 50 times, making 50 confidence intervals for μ. Each confidence interval is based on a fresh independently generated set of data. The 50 intervals are plotted in Figure 23.1 as horizontal line
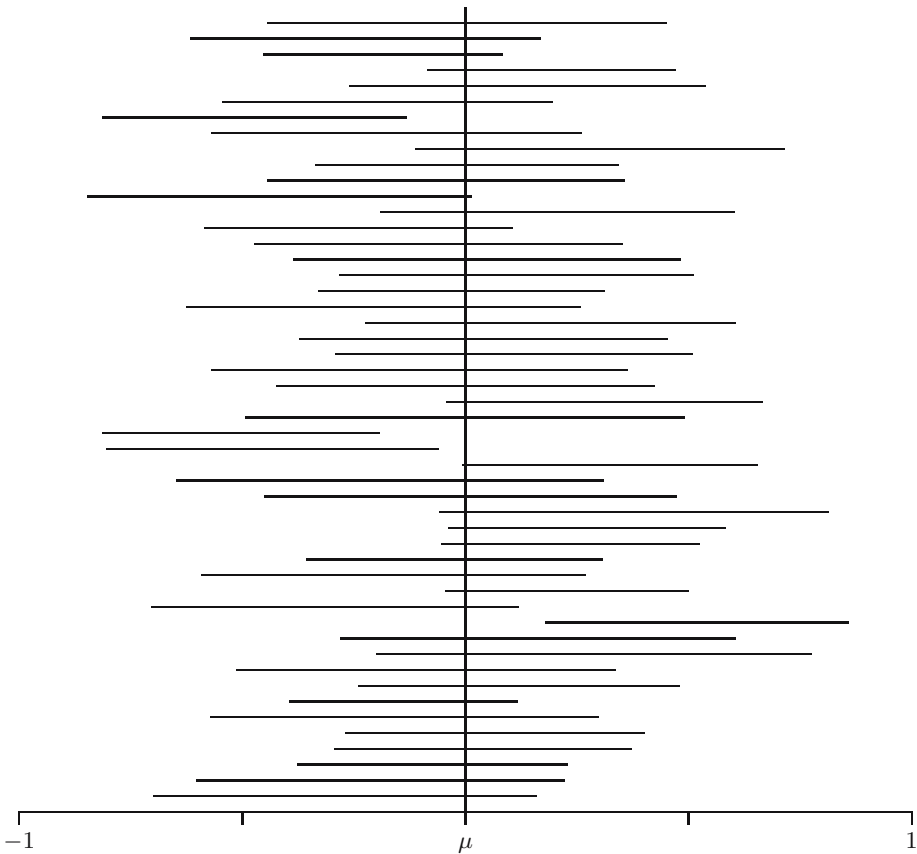
Fig. 23.1. Fifty 90% confidence intervals for $\mu = 0$.

segments, and at μ (0!) a vertical line is drawn. We count 46 "hits": only four intervals do not contain the true μ.

QUICK EXERCISE 23.3 Suppose you were to make 40 confidence intervals with confidence level 95%. About how many of them should you expect to be "wrong"? Should you be surprised if 10 of them are wrong?

In the remainder of this chapter we consider *confidence intervals for the mean*: confidence intervals for the unknown expectation μ of the distribution from which the sample originates. We start with the situation where it is known that the data originate from a normal distribution, first with known variance, then with unknown variance. Then we drop the normal assumption, first use the bootstrap, and finally show how, for very large samples, confidence intervals based on the central limit theorem are made.

23.2 Normal data

Suppose the data can be seen as the realization of a sample $X_1, \ldots, X_n$ from an $N(\mu, \sigma^2)$ distribution and μ is the (unknown) parameter of interest. If the variance σ^2 is known, confidence intervals are easily derived. Before we do this, some preparation has to be done.

Critical values

We shall need so-called critical values for the standard normal distribution. The *critical value* z_p of an $N(0, 1)$ distribution is the number that has right tail probability p. It is defined by

$$P(Z \geq z_p) = p,$$

where Z is an $N(0, 1)$ random variable. For example, from Table B.1 we read $P(Z \geq 1.96) = 0.025$, so $z_{0.025} = 1.96$. In fact, z_p is the $(1 - p)$th quantile of the standard normal distribution:

$$\Phi(z_p) = P(Z \leq z_p) = 1 - p.$$

By the symmetry of the standard normal density, $P(Z \leq -z_p) = P(Z \geq z_p) = p$, so $P(Z \geq -z_p) = 1 - p$ and therefore

$$z_{1-p} = -z_p.$$

For example, $z_{0.975} = -z_{0.025} = -1.96$. All this is illustrated in Figure 23.2.

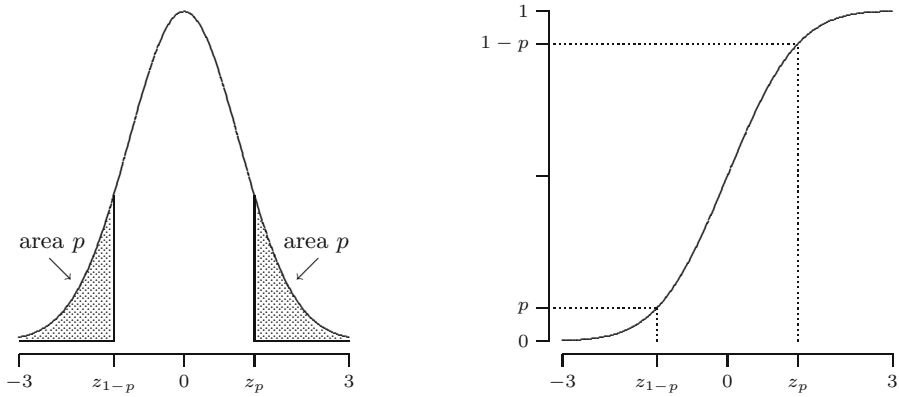QUICK EXERCISE 23.4 Determine $z_{0.01}$ and $z_{0.95}$ from Table B.1.

Fig. 23.2. Critical values of the standard normal distribution.

Variance known

If $X_1, \ldots, X_n$ is a random sample from an $N(\mu, \sigma^2)$ distribution, then $\bar{X}_n$ has an $N(\mu, \sigma^2/n)$ distribution, and from the properties of the normal distribution (see page 106), we know that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \quad \text{has an } N(0,1) \text{ distribution.}$$

If c_l and c_u are chosen such that $P(c_l < Z < c_u) = \gamma$ for an $N(0,1)$ distributed random variable Z, then

$$\gamma = P\left(c_l < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < c_u\right)$$

$$= P\left(c_l \frac{\sigma}{\sqrt{n}} < \bar{X}_n - \mu < c_u \frac{\sigma}{\sqrt{n}}\right)$$

$$= P\left(\bar{X}_n - c_u \frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n - c_l \frac{\sigma}{\sqrt{n}}\right).$$

We have found that

$$L_n = \bar{X}_n - c_u \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad U_n = \bar{X}_n - c_l \frac{\sigma}{\sqrt{n}}$$

satisfy the confidence interval definition: the interval (L_n, U_n) covers μ with probability γ. Therefore

$$\left(\bar{x}_n - c_u \frac{\sigma}{\sqrt{n}}, \ \bar{x}_n - c_l \frac{\sigma}{\sqrt{n}}\right)$$

is a $100\gamma\%$ confidence interval for μ. A common choice is to divide $\alpha = 1 - \gamma$ evenly between the tails,[2] that is, solve c_l and c_u from

---

[2] Here this choice could be motivated by the fact that it leads to the shortest confidence interval; in other examples the shortest interval requires an *asymmetric*

$$P(Z \geq c_u) = \alpha/2 \quad \text{and} \quad P(Z \leq c_l) = \alpha/2,$$

so that $c_u = z_{\alpha/2}$ and $c_l = z_{1-\alpha/2} = -z_{\alpha/2}$. Summarizing, the $100(1-\alpha)\%$ confidence interval for μ is:

$$\left(\bar{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}},\ \bar{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right).$$

For example, if $\alpha = 0.05$, we use $z_{0.025} = 1.96$ and the 95% confidence interval is

$$\left(\bar{x}_n - 1.96 \frac{\sigma}{\sqrt{n}},\ \bar{x}_n + 1.96 \frac{\sigma}{\sqrt{n}} \right).$$

Example: gross calorific content of coal

When a shipment of coal is traded, a number of its properties should be known accurately, because the value of the shipment is determined by them. An important example is the so-called gross calorific value, which characterizes the heat content and is a numerical value in megajoules per kilogram (MJ/kg). The International Organization of Standardization (ISO) issues standard procedures for the determination of these properties. For the gross calorific value, there is a method known as ISO 1928. When the procedure is carried out properly, resulting measurement errors are known to be approximately normal, with a standard deviation of about 0.1 MJ/kg. Laboratories that operate according to standard procedures receive ISO certificates. In Table 23.1, a number of such ISO 1928 measurements is given for a shipment of Osterfeld coal coded 262DE27.

Table 23.1. Gross calorific value measurements for Osterfeld 262DE27.

23.870	23.730	23.712	23.760	23.640	23.850	23.840	23.860
23.940	23.830	23.877	23.700	23.796	23.727	23.778	23.740
23.890	23.780	23.678	23.771	23.860	23.690	23.800	

Source: A.M.H. van der Veen and A.J.M. Broos. Interlaboratory study programme "ILS coal characterization"—reported data. Technical report, NMi Van Swinden Laboratorium B.V., The Netherlands, 1996.

We want to combine these values into a confidence statement about the "true" gross calorific content of Osterfeld 262DE27. From the data, we compute $\bar{x}_n = 23.788$. Using the given $\sigma = 0.1$ and $\alpha = 0.05$, we find the 95% confidence interval

$$\left(23.788 - 1.96 \frac{0.1}{\sqrt{23}},\ 23.788 + 1.96 \frac{0.1}{\sqrt{23}} \right) = (23.747, 23.829) \quad \text{MJ/kg}.$$

division of α. If you are only concerned with the left or right boundary of the confidence interval, see the next chapter.

Variance unknown

When σ is unknown, the fact that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution has become useless, as it involves this un-known σ, which would subsequently appear in the confidence interval. How-ever, if we substitute the estimator S_n for σ, the resulting random variable

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

has a distribution that only depends on n and *not* on μ or σ. Moreover, its density can be given explicitly.

> DEFINITION. A continuous random variable has a *t-distribution with parameter m*, where $m \geq 1$ is an integer, if its probability density is given by
>
> $$f(x) = k_m \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}} \qquad \text{for } -\infty < x < \infty,$$
>
> where $k_m = \Gamma\left(\frac{m+1}{2}\right) / \left(\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}\right)$. This distribution is denoted by $t(m)$ and is referred to as the t-distribution with m *degrees of freedom*.

The normalizing constant k_m is given in terms of the gamma function, which was defined on page 157. For $m = 1$, it evaluates to $k_1 = 1/\pi$, and the resulting density is that of the standard Cauchy distribution (see page 161). If X has a $t(m)$ distribution, then $\mathrm{E}[X] = 0$ for $m \geq 2$ and $\mathrm{Var}(X) = m/(m-2)$ for $m \geq 3$. Densities of t-distributions look like that of the standard normal distribution: they are also symmetric around 0 and bell-shaped. As m goes to infinity the limit of the $t(m)$ density is the standard normal density. The distinguishing feature is that densities of t-distributions have heavier tails: $f(x)$ goes to zero as x goes to $+\infty$ or $-\infty$, but more slowly than the density $\phi(x)$ of the standard normal distribution. These properties are illustrated in Figure 23.3, which shows the densities and distribution functions of the $t(1)$, $t(2)$, and $t(5)$ distribution as well as those of the standard normal.

We will also need critical values for the $t(m)$ distribution: the critical value $t_{m,p}$ is the number satisfying

$$\mathrm{P}(T \geq t_{m,p}) = p,$$

where T is a $t(m)$ distributed random variable. Because the t-distribution is symmetric around zero, using the same reasoning as for the critical values of the standard normal distribution, we find:
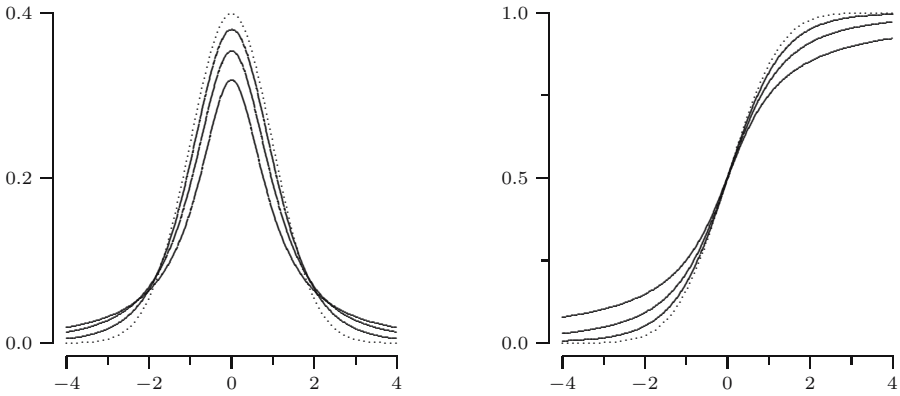
Fig. 23.3. Three t-distributions and the standard normal distribution. The dotted line corresponds to the standard normal. The other distributions depicted are the $t(1)$, $t(2)$, and $t(5)$, which in that order resemble the standard normal more and more.

$$t_{m,1-p} = -t_{m,p}.$$

For example, in Table B.2 we read $t_{10,0.01} = 2.764$, and from this we deduce that $t_{10,0.99} = -2.764$.

QUICK EXERCISE 23.5 Determine $t_{3,0.01}$ and $t_{35,0.9975}$ from Table B.2.

We now return to the distribution of

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

and construct a confidence interval for μ.

> THE STUDENTIZED MEAN OF A NORMAL RANDOM SAMPLE. For a random sample $X_1, \ldots, X_n$ from an $N(\mu, \sigma^2)$ distribution, the *studentized mean*
> $$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$
> has a $t(n-1)$ distribution, regardless of the values of μ and σ.

From this fact and using critical values of the t-distribution, we derive that

$$\mathrm{P}\left(-t_{n-1,\alpha/2} < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < t_{n-1,\alpha/2}\right) = 1 - \alpha, \tag{23.4}$$

and in the same way as when σ is known it now follows that a $100(1 - \alpha)\%$ confidence interval for μ is given by:

$$\left(\bar{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}},\ \bar{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}\right).$$

Returning to the coal example, there was another shipment, of Daw Mill 258GB41 coal, where there were actually some doubts whether the stated accuracy of the ISO 1928 method was attained. We therefore prefer to consider σ unknown and estimate it from the data, which are given in Table 23.2.

Table 23.2. Gross calorific value measurements for Daw Mill 258GB41.

30.990	31.030	31.060	30.921	30.920	30.990	31.024	30.929
31.050	30.991	31.208	30.830	31.330	30.810	31.060	30.800
31.091	31.170	31.026	31.020	30.880	31.125		

Source: A.M.H. van der Veen and A.J.M. Broos. Interlaboratory study programme "ILS coal characterization"—reported data. Technical report, NMi Van Swinden Laboratorium B.V., The Netherlands, 1996.

Doing this, we find $\bar{x}_n = 31.012$ and $s_n = 0.1294$. Because $n = 22$, for a 95% confidence interval we use $t_{21,0.025} = 2.080$ and obtain

$$\left(31.012 - 2.080\frac{0.1294}{\sqrt{22}},\ 31.012 + 2.080\frac{0.1294}{\sqrt{22}}\right) = (30.954, 31.069).$$

Note that this confidence interval is (50%!) wider than the one we made for the Osterfeld coal, with almost the same sample size. There are two reasons for this; one is that $\sigma = 0.1$ is replaced by the (larger) estimate $s_n = 0.1294$, and the second is that the critical value $z_{0.025} = 1.96$ is replaced by the larger $t_{21,0.025} = 2.080$. The differences in the method and the ingredients seem minor, but they matter, especially for small samples.

23.3 Bootstrap confidence intervals

It is not uncommon that the methods of the previous section are used even when the normal distribution is *not* a good model for the data. In some cases this is not a big problem: with small deviations from normality the actual confidence level of a constructed confidence interval may deviate only a few percent from the intended confidence level. For large datasets the central limit theorem in fact ensures that this method provides confidence intervals with approximately correct confidence levels, as we shall see in the next section.

If we doubt the normality of the data and we do *not* have a large sample, usually the best thing to do is to bootstrap. Suppose we have a dataset $x_1, \ldots, x_n$, modeled as a realization of a random sample from some distribution F, and we want to construct a confidence interval for its (unknown) expectation μ.

In the previous section we saw that it suffices to find numbers c_l and c_u such that

$$P\left(c_l < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < c_u\right) = 1 - \alpha.$$

The $100(1 - \alpha)\%$ confidence interval would then be

$$\left(\bar{x}_n - c_u\frac{s_n}{\sqrt{n}}, \ \bar{x}_n - c_l\frac{s_n}{\sqrt{n}}\right),$$

where, of course, $\bar{x}_n$ and s_n are the sample mean and the sample standard deviation. To find c_l and c_u we need to know the distribution of the studentized mean

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}.$$

We apply the bootstrap principle. From the data $x_1, \ldots, x_n$ we determine an estimate $\hat{F}$ of F. Let $X_1^*, \ldots, X_n^*$ be a random sample from $\hat{F}$, with $\mu^* = \mathrm{E}[X_i^*]$, and consider

$$T^* = \frac{\bar{X}_n^* - \mu^*}{S_n^*/\sqrt{n}}.$$

The distribution of T^* is now used as an approximation to the distribution of T. If we use $\hat{F} = F_n$, we get the following.

EMPIRICAL BOOTSTRAP SIMULATION FOR THE STUDENTIZED MEAN.
Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F. The expectation corresponding to F_n is $\mu^* = \bar{x}_n$.
1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the studentized mean for the bootstrap dataset:

$$t^* = \frac{\bar{x}_n^* - \bar{x}_n}{s_n^*/\sqrt{n}},$$

where $\bar{x}_n^*$ and s_n^* are the sample mean and sample standard deviation of $x_1^*, x_2^*, \ldots, x_n^*$.
Repeat steps 1 and 2 many times.

From the bootstrap experiment we can determine c_l^* and c_u^* such that

$$P\left(c_l^* < \frac{\bar{X}_n^* - \mu^*}{S_n^*/\sqrt{n}} < c_u^*\right) \approx 1 - \alpha.$$

By the bootstrap principle we may transfer this statement about the distribution of T^* to the distribution of T. That is, we may use these estimated critical values as bootstrap approximations to c_l and c_u:

$$c_l \approx c_l^* \quad \text{and} \quad c_u \approx c_u^*,$$

Therefore, we call

$$\left(\bar{x}_n - c_u^* \frac{s_n}{\sqrt{n}}, \ \bar{x}_n - c_l^* \frac{s_n}{\sqrt{n}}\right)$$

a $100(1-\alpha)\%$ *bootstrap confidence interval for* μ.

Example: the software data

Recall the software data, a dataset of interfailure times (see Section 17.3). From the nature of the data—failure times are positive numbers—and the histogram (Figure 17.5), we know that they should not be modeled as a realization of a random sample from a normal distribution. From the data we know $\bar{x}_n = 656.88$, $s_n = 1037.3$, and $n = 135$. We generate one thousand bootstrap datasets, and for each dataset we compute t^* as in step 2 of the procedure. The histogram and empirical distribution function made from these one thousand values are estimates of the density and the distribution function, respectively, of the bootstrap sample statistic T^*; see Figure 23.4.
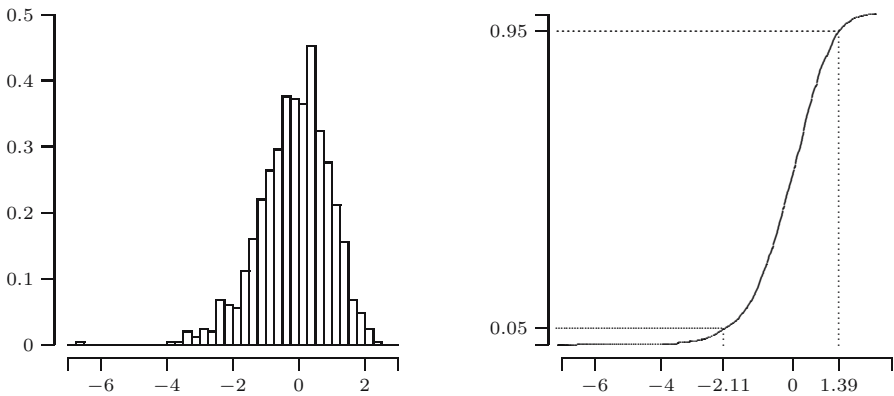
Fig. 23.4. Histogram and empirical distribution function of the studentized bootstrap simulation results for the software data.

We want to make a 90% bootstrap confidence interval, so we need c_l^* and c_u^*, or the 0.05th and 0.95th quantile from the empirical distribution function in Figure 23.4. The 50th order statistic of the one thousand t^* values is -2.107. This means that 50 out of the one thousand values, or 5%, are smaller than or equal to this value, and so $c_l^* = -2.107$. Similarly, from the 951st order statistic, 1.389, we obtain[3] $c_u^* = 1.389$. Inserting these values, we find the following 90% bootstrap confidence interval for μ:

---

[3] These results deviate slightly from the definition of empirical quantiles as given in Section 16.3. That method is a little more accurate.

$$\left(656.88 - 1.389\frac{1037.3}{\sqrt{135}},\ 656.88 - (-2.107)\frac{1037.3}{\sqrt{135}}\right) = (532.9, 845.0).$$

QUICK EXERCISE 23.6 The 25th and 976th order statistic from the preceding bootstrap results are -2.443 and 1.713, respectively. Use these numbers to construct a confidence interval for μ. What is the corresponding confidence level?

Why the bootstrap may be better

The reason to use the bootstrap is that it should lead to a more accurate approximation of the distribution of the studentized mean than the $t(n-1)$ distribution that follows from *assuming* normality. If, in the previous example, we would think we had normal data, we would use critical values from the $t(134)$ distribution: $t_{134,0.05} = 1.656$. The result would be

$$\left(656.88 - 1.656\frac{1037.3}{\sqrt{135}},\ 656.88 + 1.656\frac{1037.3}{\sqrt{135}}\right) = (509.0, 804.7).$$

Comparing the intervals, we see that here the bootstrap interval is a little larger and, as opposed to the t-interval, not centered around the sample mean but *skewed* to the right side. This is one of the features of the bootstrap: if the distribution from which the data originate is *skewed*, this is reflected in the confidence interval. Looking at the histogram of the software data (Figure 17.5), we see that is it skewed to the right: it has a long tail on the right, but not on the left, so the same most likely holds for the distribution from which these data originate. The skewness is reflected in the confidence interval, which extends more to the right of $\bar{x}_n$ than to the left. In some sense, the bootstrap adapts to the shape of the distribution, and in this way it leads to more accurate confidence statements than using the method for normal data. What we mean by this is that, for example, with the normal method only 90% of the 95% confidence statements would actually cover the true value, whereas for the bootstrap intervals this percentage would be close(r) to 95%.

23.4 Large samples

A variant of the central limit theorem states that as n goes to infinity, the distribution of the studentized mean

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

approaches the standard normal distribution. This fact is the basis for so-called *large sample confidence intervals*. Suppose $X_1, \ldots, X_n$ is a random

sample from some distribution F with expectation μ. If n is large enough, we may use

$$P\left(-z_{\alpha/2} < \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha. \qquad (23.5)$$

This implies that if $x_1, \ldots, x_n$ can be seen as a realization of a random sample from some unknown distribution with expectation μ and if n is large enough, then

$$\left(\bar{x}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}}, \ \bar{x}_n + z_{\alpha/2}\frac{s_n}{\sqrt{n}}\right)$$

is an approximate $100(1 - \alpha)\%$ confidence interval for μ.

Just as earlier with the central limit theorem, a key question is "how big should n be?" Again, there is no easy answer. To give you some idea, we have listed in Table 23.3 the results of a small simulation experiment. For each of the distributions, sample sizes, and confidence levels listed, we constructed 10 000 confidence intervals with the large sample method; the numbers listed in the table are the confidence levels as estimated from the simulation, the *coverage probabilities*. The chosen Pareto distribution is very skewed, and this shows; the coverage probabilities for the exponential are just a few percent off.

Table 23.3. Estimated coverage probabilities for large sample confidence intervals for non-normal data.

Distribution	n	γ 0.900	0.950
$Exp(1)$	20	0.851	0.899
$Exp(1)$	100	0.890	0.938
$Par(2.1)$	20	0.727	0.774
$Par(2.1)$	100	0.798	0.849

In the case of simulation one can often quite easily generate a very large number of independent repetitions, and then this question poses no problem. In other cases there may be nothing better to do than hope that the dataset is large enough. We give an example where (we believe!) this is definitely the case.

In an article published in 1910 ([28]), Rutherford and Geiger reported their observations on the radioactive decay of the element polonium. Using a small disk coated with polonium they counted the number of emitted alpha-particles during 2608 intervals of 7.5 seconds each. The dataset consists of the counted number of alpha-particles for each of the 2608 intervals and can be summarized as in Table 23.4.

Table 23.4. Alpha-particle counts for 2608 intervals of 7.5 seconds.

Count	0	1	2	3	4
Frequency	57	203	383	525	532
Count	5	6	7	8	9
Frequency	408	273	139	45	27
Count	10	11	12	13	14
Frequency	10	4	0	1	1

Source: E. Rutherford and H. Geiger (with a note by H. Bateman), The proba-
bility variations in the distribution of α particles, *Phil. Mag.*, 6: 698–704, 1910;
the table on page 701.

The total number of counted alpha-particles is $10\,097$, the average number
per interval is therefore 3.8715. The sample standard deviation can also
be computed from the table; it is 1.9225. So we know of the actual data
$x_1, x_2, \ldots, x_{2608}$ (where the counts x_i are between 0 and 14) that $\bar{x}_n = 3.8715$
and $s_n = 1.9225$. We construct a 98% confidence interval for the expected
number of particles per interval. As $z_{0.01} = 2.33$ this results in

$$\left(3.8715 - 2.33\frac{1.9225}{\sqrt{2608}},\ 3.8715 + 2.33\frac{1.9225}{\sqrt{2608}}\right) = (3.784, 3.959).$$

23.5 Solutions to the quick exercises

23.1 From the probability statement, we derive, using $\sigma_T = 100$ and $8/9 =$
0.889:

$$\theta \in (T - 300, T + 300) \quad \text{with probability at least 88\%.}$$

With $t = 299\,852.4$, this becomes

$$\theta \in (299\,552.4,\ 300\,152.4) \quad \text{with confidence at least 88\%.}$$

23.2 Chebyshev's inequality only gives an upper bound. The actual value
of $P(|T - \theta| < 2\sigma_T)$ could be higher than 3/4, depending on the distribution
of T. For example, in Quick exercise 13.2 we saw that in case of an exponen-
tial distribution this probability is 0.865. For other distributions, even higher
values are attained; see Exercise 13.1.

23.3 For each of the confidence intervals we have a 5% probability that
it is wrong. Therefore, the number of wrong confidence intervals has a
$Bin(40, 0.05)$ distribution, and we would expect about $40 \cdot 0.05 = 2$ to be
wrong. The standard deviation of this distribution is $\sqrt{40 \cdot 0.05 \cdot 0.95} = 1.38$.
The outcome "10 confidence intervals wrong" is $(10 - 2)/1.38 = 5.8$ standard
deviations from the expectation and would be a surprising outcome indeed.
(The probability of 10 or more wrong is 0.00002.)

23.4 We need to solve $P(Z \geq a) = 0.01$. In Table B.1 we find $P(Z \geq 2.33) = 0.0099 \approx 0.01$, so $z_{0.01} \approx 2.33$. For $z_{0.95}$ we need to solve $P(Z \geq a) = 0.95$, and because this is in the left tail of the distribution, we use $z_{0.95} = -z_{0.05}$. In the table we read $P(Z \geq 1.64) = 0.0505$ and $P(Z \geq 1.65) = 0.0495$, from which we conclude $z_{0.05} \approx (1.64 + 1.65)/2 = 1.645$ and $z_{0.95} \approx -1.645$.

23.5 In Table B.1 we find $P(T_3 \geq 4.541) = 0.01$, so $t_{3,0.01} = 4.541$. For $t_{35,0.9975}$, we need to use $t_{35,0.9975} = -t_{35,0.0025}$. In the table we find $t_{30,0.0025} = 3.030$ and $t_{40,0.0025} = 2.971$, and by interpolation $t_{35,0.0025} \approx (3.030 + 2.971)/2 = 3.0005$. Hence, $t_{35,0.9975} \approx -3.000$.

23.6 The order statistics are estimates for $c^*_{0.025}$ and $c^*_{0.975}$, respectively. So the corresponding α is 0.05, and the 95% bootstrap confidence interval for μ is:

$$\left(656.88 - 1.713 \frac{1037.3}{\sqrt{135}}, \ 656.88 - (-2.443) \frac{1037.3}{\sqrt{135}} \right) = (504.0, 875.0).$$

23.6 Exercises

23.1 ⊡ A bottling machine is known to fill wine bottles with amounts that follow an $N(\mu, \sigma^2)$ distribution, with $\sigma = 5$ (ml). In a sample of 16 bottles, $\bar{x} = 743$ (ml) was found. Construct a 95% confidence interval for μ.

23.2 ⊡ You are given a dataset that may be considered a realization of a normal random sample. The size of the dataset is 34, the average is 3.54, and the sample standard deviation is 0.13. Construct a 98% confidence interval for the unknown expectation μ.

23.3 You have ordered 10 bags of cement, which are supposed to weigh 94 kg each. The average weight of the 10 bags is 93.5 kg. Assuming that the 10 weights can be viewed as a realization of a random sample from a normal distribution with unknown parameters, construct a 95% confidence interval for the expected weight of a bag. The sample standard deviation of the 10 weights is 0.75.

23.4 A new type of car tire is launched by a tire manufacturer. The automobile association performs a durability test on a random sample of 18 of these tires. For each tire the durability is expressed as a percentage: a score of 100 (%) means that the tire lasted exactly as long as the average standard tire, an accepted comparison standard. From the multitude of factors that influence the durability of individual tires the assumption is warranted that the durability of an arbitrary tire follows an $N(\mu, \sigma^2)$ distribution. The parameters μ and σ^2 characterize the tire *type*, and μ could be called the durability index for this type of tire. The automobile association found for the tested tires: $\bar{x}_{18} = 195.3$ and $s_{18} = 16.7$. Construct a 95% confidence interval for μ.

23.5 ⊞ During the 2002 Winter Olympic Games in Salt Lake City a newspaper article mentioned the alleged advantage speed-skaters have in the 1500 m race if they start in the outer lane. In the men's 1500 m, there were 24 races, but in race 13 (really!) someone fell and did not finish. The results in seconds of the remaining 23 races are listed in Table 23.5. You should know that who races against whom, in which race, and who starts in the outer lane are all determined by a fair lottery.

Table 23.5. Speed-skating results in seconds, men's 1500 m (except race 13), 2002 Winter Olympic Games.

Race number	Inner lane	Outer lane	Difference
1	107.04	105.98	1.06
2	109.24	108.20	1.04
3	111.02	108.40	2.62
4	108.02	108.58	−0.56
5	107.83	105.51	2.32
6	109.50	112.01	−2.51
7	111.81	112.87	−1.06
8	111.02	106.40	4.62
9	106.04	104.57	1.47
10	110.15	110.70	−0.55
11	109.42	109.45	−0.03
12	108.13	109.57	−1.44
14	105.86	105.97	−0.11
15	108.27	105.63	2.64
16	107.63	105.41	2.22
17	107.72	110.26	−2.54
18	106.38	105.82	0.56
19	107.78	106.29	1.49
20	108.57	107.26	1.31
21	106.99	103.95	3.04
22	107.21	106.00	1.21
23	105.34	105.26	0.08
24	108.76	106.75	2.01
Mean	108.25	107.43	0.82
St.dev.	1.70	2.42	1.78

a. As a consequence of the lottery and the fact that many different factors contribute to the actual time difference "inner lane minus outer lane" the assumption of a normal distribution for the difference is warranted. The numbers in the last column can be seen as realizations from an $N(\delta, \sigma^2)$

distribution, where δ is the expected outer lane advantage. Construct a 95% confidence interval for δ. N.B. $n = 23$, not 24!

b. You decide to make a bootstrap confidence interval instead. Describe the appropriate bootstrap experiment.

c. The bootstrap experiment was performed with one thousand repetitions. Part of the bootstrap outcomes are listed in the following table. From the *ordered* list of results, numbers 21 to 60 and 941 to 980 are given. Use these to construct a 95% bootstrap confidence interval for δ.

21–25	−2.202	−2.164	−2.111	−2.109	−2.101
26–30	−2.099	−2.006	−1.985	−1.967	−1.929
31–35	−1.917	−1.898	−1.864	−1.830	−1.808
36–40	−1.800	−1.799	−1.774	−1.773	−1.756
41–45	−1.736	−1.732	−1.731	−1.717	−1.716
46–50	−1.699	−1.692	−1.691	−1.683	−1.666
51–55	−1.661	−1.644	−1.638	−1.637	−1.620
56–60	−1.611	−1.611	−1.601	−1.600	−1.593
941–945	1.648	1.667	1.669	1.689	1.696
946–950	1.708	1.722	1.726	1.735	1.814
951–955	1.816	1.825	1.856	1.862	1.864
956–960	1.875	1.877	1.897	1.905	1.917
961–965	1.923	1.948	1.961	1.987	2.001
966–970	2.015	2.015	2.017	2.018	2.034
971–975	2.035	2.037	2.039	2.053	2.060
976–980	2.088	2.092	2.101	2.129	2.143

23.6 ⊞ A dataset $x_1, x_2, \ldots, x_n$ is given, modeled as realization of a sample $X_1, X_2, \ldots, X_n$ from an $N(\mu, 1)$ distribution. Suppose there are sample statistics $L_n = g(X_1, \ldots, X_n)$ and $U_n = h(X_1, \ldots, X_n)$ such that

$$P(L_n < \mu < U_n) = 0.95$$

for every value of μ. Suppose that the corresponding 95% confidence interval derived from the data is $(l_n, u_n) = (-2, 5)$.

a. Suppose $\theta = 3\mu + 7$. Let $\tilde{L}_n = 3L_n + 7$ and $\tilde{U}_n = 3U_n + 7$. Show that $P\left(\tilde{L}_n < \theta < \tilde{U}_n\right) = 0.95$.

b. Write the 95% confidence interval for θ in terms of l_n and u_n.

c. Suppose $\theta = 1 - \mu$. Again, find $\tilde{L}_n$ and $\tilde{U}_n$, as well as the confidence interval for θ.

d. Suppose $\theta = \mu^2$. Can you construct a confidence interval for θ?

23.7 ☐ A 95% confidence interval for the parameter μ of a $Pois(\mu)$ distribution is given: $(2,3)$. Let X be a random variable with this distribution. Construct a 95% confidence interval for $P(X=0)=e^{-\mu}$.

23.8 Suppose that in Exercise 23.1 the content of the bottles has to be determined by weighing. It is known that the wine bottles involved weigh on average 250 grams, with a standard deviation of 15 grams, and the weights follow a normal distribution. For a sample of 16 bottles, an average weight of 998 grams was found. You may assume that 1 ml of wine weighs 1 gram, and that the filling amount is independent of the bottle weight. Construct a 95% confidence interval for the expected amount of wine per bottle, μ.

23.9 Consider the alpha-particle counts discussed in Section 23.4; the data are given in Table 23.4. We want to bootstrap in order to make a bootstrap confidence interval for the expected number of particles in a 7.5-second interval.

a. Describe in detail how you would perform the bootstrap simulation.

b. The bootstrap experiment was performed with one thousand repetitions. Part of the (ordered) bootstrap t^*'s are given in the following table. Construct the 95% bootstrap confidence interval for the expected number of particles in a 7.5-second interval.

1–5	−2.996	−2.942	−2.831	−2.663	−2.570
6–10	−2.537	−2.505	−2.290	−2.273	−2.228
11–15	−2.193	−2.112	−2.092	−2.086	−2.045
16–20	−1.983	−1.980	−1.978	−1.950	−1.931
21–25	−1.920	−1.910	−1.893	−1.889	−1.888
26–30	−1.865	−1.864	−1.832	−1.817	−1.815
31–35	−1.755	−1.751	−1.749	−1.746	−1.744
36–40	−1.734	−1.723	−1.710	−1.708	−1.705
41–45	−1.703	−1.700	−1.696	−1.692	−1.691
46–50	−1.691	−1.675	−1.660	−1.656	−1.650
951–955	1.635	1.638	1.643	1.648	1.661
956–960	1.666	1.668	1.678	1.681	1.686
961–965	1.692	1.719	1.721	1.753	1.772
966–970	1.773	1.777	1.806	1.814	1.821
971–975	1.824	1.826	1.837	1.838	1.845
976–980	1.862	1.877	1.881	1.883	1.956
981–985	1.971	1.992	2.060	2.063	2.083
986–990	2.089	2.177	2.181	2.186	2.224
991–995	2.234	2.264	2.273	2.310	2.348
996–1000	2.483	2.556	2.870	2.890	3.546

c. Answer this without doing any calculations: if we made the 98% bootstrap confidence interval, would it be smaller or larger than the interval constructed in Section 23.4?

23.10 In a report you encounter a 95% confidence interval $(1.6, 7.8)$ for the parameter μ of an $N(\mu, \sigma^2)$ distribution. The interval is based on 16 observations, constructed according to the studentized mean procedure.

a. What is the mean of the (unknown) dataset?

b. You prefer to have a 99% confidence interval for μ. Construct it.

23.11 ⊞ A 95% confidence interval for the unknown expectation of some distribution contains the number 0.

a. We construct the corresponding 98% confidence interval, using the same data. Will it contain the number 0?

b. The confidence interval in fact is a bootstrap confidence interval. We repeat the bootstrap experiment (using the same data) and construct a new 95% confidence interval based on the results. Will it contain the number 0?

c. We collect new data, resulting in a dataset of the same size. With this data, we construct a 95% confidence interval for the unknown expectation. Will the interval contain 0?

23.12 Let $Z_1, \ldots, Z_n$ be a random sample from an $N(0, 1)$ distribution. Define $X_i = \mu + \sigma Z_i$ for $i = 1, \ldots, n$ and $\sigma > 0$. Let $\bar{Z}$, $\bar{X}$ denote the sample averages and S_Z and S_X the sample standard deviations, of the Z_i and X_i, respectively.

a. Show that $X_1, \ldots, X_n$ is a random sample from an $N(\mu, \sigma^2)$ distribution.

b. Express $\bar{X}$ and S_X in terms of $\bar{Z}$, S_Z, μ, and σ.

c. Verify that

$$\frac{\bar{X} - \mu}{S_X/\sqrt{n}} = \frac{\bar{Z}}{S_Z/\sqrt{n}},$$

and explain why this shows that the distribution of the studentized mean does not depend on μ and σ.

More on confidence intervals

While in Chapter 23 we were solely concerned with confidence intervals for expectations, in this chapter we treat a variety of topics. First, we focus on confidence intervals for the parameter p of the binomial distribution. Then, based on an example, we briefly discuss a general method to construct confidence intervals. One-sided confidence intervals, or upper and lower confidence bounds, are discussed next. At the end of the chapter we investigate the question of how to determine the sample size when a confidence interval of a certain width is desired.

24.1 The probability of success

A common situation is that we observe a random variable X with a $Bin(n, p)$ distribution and use X to estimate p. For example, if we want to estimate the proportion of voters that support candidate G in an election, we take a sample from the voter population and determine the proportion in the sample that supports G. If n individuals are selected at random from the population, where a proportion p supports candidate G, the number of supporters X in the sample is modeled by a $Bin(n, p)$ distribution; we count the supporters of candidate G as "successes." Usually, the sample proportion X/n is taken as an estimator for p.

If we want to make a confidence interval for p, based on the number of successes X in the sample, we need to find statistics L and U (see the definition of confidence intervals on page 343) such that

$$P(L < p < U) = 1 - \alpha,$$

where L and U are to be based on X only. In general, this problem does not have a solution. However, the method for large n described next, sometimes called "the Wilson method" (see [40]), yields confidence intervals with

confidence level approximately $100(1 - \alpha)\%$. (How close the true confidence level is to $100(1 - \alpha)\%$ depends on the (unknown) p, though it is known that for p near 0 and 1 it is too low. For some details and an alternative for this situation, see Remark 24.1.)

Recall the normal approximation to the binomial distribution, a consequence of the central limit theorem (see page 201 and Exercise 14.5): for large n, the distribution of X is approximately normal and

$$\frac{X - np}{\sqrt{np(1-p)}}$$

is approximately standard normal. By dividing by n in both the numerator and the denominator, we see that this equals:

$$\frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}.$$

Therefore, for large n

$$\mathrm{P}\left(-z_{\alpha/2} < \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) \approx 1 - \alpha.$$

Note that the event

$$-z_{\alpha/2} < \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}$$

is the same as

$$\left(\frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}\right)^2 < \left(z_{\alpha/2}\right)^2$$

or

$$\left(\frac{X}{n} - p\right)^2 - \left(z_{\alpha/2}\right)^2 \frac{p(1-p)}{n} < 0.$$

To derive expressions for L and U we can rewrite the inequality in this statement to obtain the form $L < p < U$, but the resulting formulas are rather awkward. To obtain the confidence interval, we instead substitute the data values directly and then solve for p, which yields the desired result.

Suppose, in a sample of 125 voters, 78 support one candidate. What is the 95% confidence interval for the population proportion p supporting that candidate? The realization of X is $x = 78$ and $n = 125$. We substitute this, together with $z_{\alpha/2} = z_{0.025} = 1.96$, in the last inequality:

$$\left(\frac{78}{125} - p\right)^2 - \frac{(1.96)^2}{125} p(1-p) < 0,$$
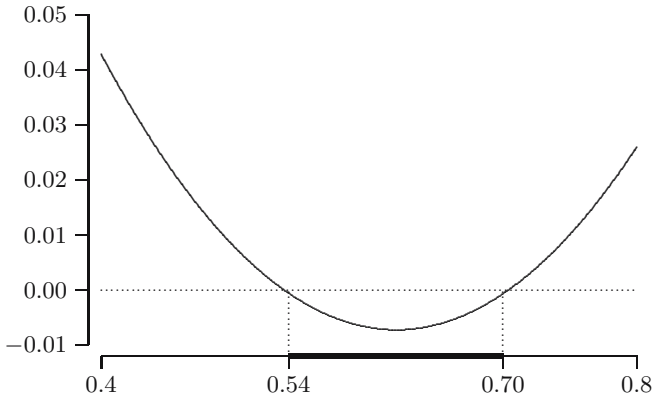
Fig. 24.1. The parabola $1.0307\,p^2 - 1.2787\,p + 0.3894$ and the resulting confidence interval.

or, working out squares and products and grouping terms:

$$1.0307\,p^2 - 1.2787\,p + 0.3894 < 0.$$

This quadratic form describes a parabola, which is depicted in Figure 24.1. Also, for other values of n and x there always results a quadratic inequality like this, with a positive coefficient for p^2 and a similar picture. For the confidence interval we need to find the values where the parabola intersects the horizontal axis. The solutions we find are:

$$p_{1,2} = \frac{-(-1.2787) \pm \sqrt{(-1.2787)^2 - 4 \cdot 1.0307 \cdot 0.3894}}{2 \cdot 1.0307} = 0.6203 \pm 0.0835;$$

hence, $l = 0.54$ and $u = 0.70$, so the resulting confidence interval is $(0.54, 0.70)$.

QUICK EXERCISE 24.1 Suppose in another election we find 80 supporters in a sample of 200. Suppose we use $\alpha = 0.0456$ for which $z_{\alpha/2} = 2$. Construct the corresponding confidence interval for p.

> **Remark 24.1 (Coverage probabilities and an alternative method).**
> Because of the discrete nature of the binomial distribution, the probability that the confidence interval covers the true parameter value depends on p. As a function of p it typically oscillates in a sawtooth-like manner around $1 - \alpha$, being too high for some values and too low for others. This is something that cannot be escaped from; the phenomenon is present in every method. In an average sense, the method treated in the text yields coverage probabilities close to $1 - \alpha$, though for arbitrarily high values of n it is possible to find p's for which the actual coverage is several percentage points too low. The low coverage occurs for p's near 0 and 1.

An alternative is the method proposed by Agresti and Coull, which overall is more conservative than the Wilson method (in fact, the Agresti-Coull interval contains the Wilson interval as a proper subset). Especially for p near 0 or 1 this method yields conservative confidence intervals. Define

$$\tilde{X} = X + \frac{(z_{\alpha/2})^2}{2} \quad \text{and} \quad \tilde{n} = n + (z_{\alpha/2})^2,$$

and $\tilde{p} = \tilde{X}/\tilde{n}$. The approximate $100(1 - \alpha)\%$ confidence interval is then given by

$$\left(\tilde{p} - z_{\alpha/2}\sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}}, \ \tilde{p} + z_{\alpha/2}\sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}} \right).$$

For a clear survey paper on confidence intervals for p we recommend Brown et al. [4].

24.2 Is there a general method?

We have now seen a number of examples of confidence intervals, and while it should be clear to you that in each of these cases the resulting intervals *are* valid confidence intervals, you may wonder how we go about finding confidence intervals in new situations. One could ask: is there a general method? We first consider an example.

A confidence interval for the minimum lifetime

Suppose we have a random sample $X_1, \ldots, X_n$ from a *shifted* exponential distribution, that is, $X_i = \delta + Y_i$, where $Y_1, \ldots, Y_n$ are a random sample from an *Exp*(1) distribution. This type of random variable is sometimes used to model lifetimes; a minimum lifetime is guaranteed, but otherwise the lifetime has an exponential distribution. The unknown parameter δ represents the minimum lifetime, and the probability density of the X_i is positive only for values greater than δ.

To derive information about δ it is natural to use the smallest observed value $T = \min\{X_1, \ldots, X_n\}$. This is also the maximum likelihood estimator for δ; see Exercise 21.6. Writing

$$T = \min\{\delta + Y_1, \ldots, \delta + Y_n\} = \delta + \min\{Y_1, \ldots, Y_n\}$$

and observing that $M = \min\{Y_1, \ldots, Y_n\}$ has an *Exp*(n) distribution (see Exercise 8.18), we find for the distribution function of T: $F_T(a) = 0$ for $a < \delta$ and

$$
\begin{aligned}
F_T(a) = \mathrm{P}(T \leq a) = \mathrm{P}(\delta + M \leq a) &= \mathrm{P}(M \leq a - \delta) \\
&= 1 - \mathrm{e}^{-n(a-\delta)} \quad \text{for } a \geq \delta.
\end{aligned}
\tag{24.1}
$$

Next, we solve

$$P(c_l < T < c_u) = 1 - \alpha$$

by requiring

$$P(T \le c_l) = P(T \ge c_u) = \tfrac{1}{2}\alpha.$$

Using (24.1) we find the following equations:

$$1 - e^{-n(c_l - \delta)} = \tfrac{1}{2}\alpha \quad \text{and} \quad e^{-n(c_u - \delta)} = \tfrac{1}{2}\alpha$$

whose solutions are

$$c_l = \delta - \frac{1}{n} \ln\left(1 - \tfrac{1}{2}\alpha\right) \quad \text{and} \quad c_u = \delta - \frac{1}{n} \ln\left(\tfrac{1}{2}\alpha\right).$$

Both c_l and c_u are values larger than δ, because the logarithms are negative. We have found that, whatever the value of δ:

$$P\left(\delta - \frac{1}{n} \ln\left(1 - \tfrac{1}{2}\alpha\right) < T < \delta - \frac{1}{n} \ln\left(\tfrac{1}{2}\alpha\right)\right) = 1 - \alpha.$$

By rearranging the inequalities, we see this is equivalent to

$$P\left(T + \frac{1}{n} \ln\left(\tfrac{1}{2}\alpha\right) < \delta < T + \frac{1}{n} \ln\left(1 - \tfrac{1}{2}\alpha\right)\right) = 1 - \alpha,$$

and therefore a $100(1 - \alpha)\%$ confidence interval for δ is given by

$$\left(t + \frac{1}{n} \ln\left(\tfrac{1}{2}\alpha\right), \ t + \frac{1}{n} \ln\left(1 - \tfrac{1}{2}\alpha\right)\right). \tag{24.2}$$

For $\alpha = 0.05$ this becomes:

$$\left(t - \frac{3.69}{n}, \ t - \frac{0.0253}{n}\right).$$

QUICK EXERCISE 24.2 Suppose you have a dataset of size 15 from a shifted $Exp(1)$ distribution, whose minimum value is 23.5. What is the 99% confidence interval for δ?

Looking back at the example, we see that the confidence interval could be constructed because we know that $T - \delta = M$ has an exponential distribution. There are many more examples of this type: some function $g(T, \theta)$ of a sample statistic T and the unknown parameter θ has a known distribution. However, this still does not cover all the ways to construct confidence intervals (see also the following remark).

Remark 24.2 (About a general method). Suppose $X_1, \ldots, X_n$ is a random sample from some distribution depending on some unknown parameter θ and let T be a sample statistic. One possible choice is to select a T that is an estimator for θ, but this is not necessary. In each case, the

distribution of T depends on θ, just as that of $X_1, \ldots, X_n$ does. In some cases it might be possible to find functions $g(\theta)$ and $h(\theta)$ such that

$$P(g(\theta) < T < h(\theta)) = 1 - \alpha \quad \text{for every value of } \theta. \qquad (24.3)$$

If this is so, then confidence statements about θ can be made. In more special cases, for example if g and h are strictly increasing, the inequalities $g(\theta) < T < h(\theta)$ can be rewritten as

$$h^{-1}(T) < \theta < g^{-1}(T),$$

and then (24.3) is equivalent to

$$P(h^{-1}(T) < \theta < g^{-1}(T)) = 1 - \alpha \quad \text{for every value of } \theta.$$

Checking with the confidence interval definition, we see that the last statement implies that $(h^{-1}(t), g^{-1}(t))$ is a $100(1-\alpha)\%$ confidence interval for θ.

24.3 One-sided confidence intervals

Suppose you are in charge of a power plant that generates and sells electricity, and you are about to buy a shipment of coal, say a shipment of the Daw Mill coal identified as 258GB41 earlier. You plan to buy the shipment if you are confident that the gross calorific content exceeds 31.00 MJ/kg. At the end of Section 23.2 we obtained for the gross calorific content the 95% confidence interval $(30.946, 31.067)$: based on the data we are 95% confident that the gross calorific content is higher than 30.946 and lower than 31.067.

In the present situation, however, we are *only* interested in the lower bound: we would prefer a confidence statement of the type "we are 95% confident that the gross calorific content exceeds 31.00." Modifying equation (23.4) we find

$$P\left(\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < t_{n-1,\alpha}\right) = 1 - \alpha,$$

which is equivalent to

$$P\left(\bar{X}_n - t_{n-1,\alpha}\frac{S_n}{\sqrt{n}} < \mu\right) = 1 - \alpha.$$

We conclude that

$$\left(\bar{x}_n - t_{n-1,\alpha}\frac{s_n}{\sqrt{n}}, \infty\right)$$

is a $100(1-\alpha)\%$ *one-sided confidence interval* for μ. For the Daw Mill coal, using $\alpha = 0.05$, with $t_{21,0.05} = 1.721$ this results in:

$$\left(31.012 - 1.721\frac{0.1294}{\sqrt{22}}, \infty\right) = (30.964, \infty).$$

We see that because "all uncertainty may be put on one side," the lower bound in the one-sided interval is higher than that in the two-sided one, though still below 31.00. Other situations may require a confidence *upper bound*. For example, if the calorific value is below a certain number you can try to negotiate a lower the price.

The definition of confidence intervals (page 343) can be extended to include one-sided confidence intervals as well. If we have a sample statistic L_n such that

$$\mathrm{P}(L_n < \theta) = \gamma$$

for every value of the parameter of interest θ, then

$$(l_n, \infty)$$

is called a $100\gamma\%$ *one-sided confidence interval for* θ. The number l_n is sometimes called a $100\gamma\%$ *lower confidence bound for* θ. Similary, U_n with $\mathrm{P}(\theta < U_n) = \gamma$ for every value of θ, yields the one-sided confidence interval $(-\infty, u_n)$, and u_n is called a $100\gamma\%$ *upper confidence bound*.

QUICK EXERCISE 24.3 Determine the 99% upper confidence bound for the gross calorific value of the Daw Mill coal.

24.4 Determining the sample size

The narrower the confidence interval the better (why?). As a general principle, we know that more accurate statements can be made if we have more measurements. Sometimes, an accuracy requirement is set, even before data are collected, and the corresponding sample size is to be determined. We provide an example of how to do this and note that this generally can be done, but the actual computation varies with the type of confidence interval.

Consider the question of the calorific content of coal once more. We have a shipment of coal to test and we want to obtain a 95% confidence interval, but it should not be wider than 0.05 MJ/kg, i.e., the lower and upper bound should not differ more than 0.05. How many measurements do we need?

We answer this question for the case when ISO method 1928 is used, whence we may assume that measurements are normally distributed with standard deviation $\sigma = 0.1$. When the desired confidence level is $1 - \alpha$, the width of the confidence interval will be

$$2 \cdot z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Requiring that this is at most w means finding the smallest n that satisfies

$$2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq w$$

or
$$n \geq \left(\frac{2z_{\alpha/2}\sigma}{w}\right)^2.$$

For the example: $w = 0.05$, $\sigma = 0.1$, and $z_{0.025} = 1.96$; so
$$n \geq \left(\frac{2 \cdot 1.96 \cdot 0.1}{0.05}\right)^2 = 61.4,$$

that is, we should perform at least 62 measurements.

In case σ is unknown, we somehow have to estimate it, and then the method can only give an indication of the required sample size. The standard deviation as we (afterwards) estimate it from the data may turn out to be quite different, and the obtained confidence interval may be smaller or larger than intended.

QUICK EXERCISE 24.4 What is the required sample size if we want the 99% confidence interval to be 0.05 MJ/kg wide?

24.5 Solutions to the quick exercises

24.1 We need to solve
$$\left(\frac{80}{200} - p\right)^2 - \frac{(2)^2}{200}p(1-p) < 0, \quad \text{or} \quad 1.02\,p^2 - 0.82p + 0.16 < 0.$$

The solutions are:
$$p_{1,2} = \frac{-(-0.82) \pm \sqrt{(-0.82)^2 - 4 \cdot 1.02 \cdot 0.16}}{2 \cdot 1.02} = 0.4020 \pm 0.0686,$$

so the confidence interval is $(0.33, 0.47)$.

24.2 We should substitute $n = 15$, $t = 23.5$, and $\alpha = 0.01$ into:
$$\left(t + \frac{1}{n}\ln\left(\tfrac{1}{2}\alpha\right), \ t + \frac{1}{n}\ln\left(1 - \tfrac{1}{2}\alpha\right)\right),$$

which yields
$$\left(23.5 - \frac{5.30}{15}, \ 23.5 - \frac{0.0050}{15}\right) = (23.1467, 23.4997).$$

24.3 The upper confidence bound is given by
$$u_n = \bar{x}_n + t_{21,0.01}\frac{s_n}{\sqrt{22}},$$

where $\bar{x}_n = 31.012$, $t_{21,0.01} = 2.518$, and $s_n = 0.1294$. Substitution yields $u_n = 31.081$.

24.4 The confidence level changes to 99%, so we use $z_{0.005} = 2.576$ instead of 1.96 in the computation:

$$n \geq \left(\frac{2 \cdot 2.576 \cdot 0.1}{0.05} \right)^2 = 106.2,$$

so we need at least 107 measurements.

24.6 Exercises

24.1 ⊡ Of a series of 100 (independent and identical) chemical experiments, 70 were concluded succesfully. Construct a 90% confidence interval for the success probability of this type of experiment.

24.2 In January 2002 the Euro was introduced and soon after stories started to circulate that some of the Euro coins would not be fair coins, because the "national side" of some coins would be too heavy or too light (see, for example, the *New Scientist* of January 4, 2002, but also national newspapers of that date).

a. A French 1 Euro coin was tossed six times, resulting in 1 heads and 5 tails. Is it reasonable to use the Wilson method, introduced in Section 24.1, to construct a confidence interval for p?

b. A Belgian 1 Euro coin was tossed 250 times: 140 heads and 110 tails. Construct a 95% confidence interval for the probability of getting heads with this coin.

24.3 In Exercise 23.1, what sample size is needed if we want a 99% confidence interval for μ at most 1 ml wide?

24.4 ⊡ Recall Exercise 23.3 and the 10 bags of cement that should each weigh 94 kg. The average weight was 93.5 kg, with sample standard deviation 0.75.

a. Based on these data, how many bags would you need to sample to make a 90% confidence interval that is 0.1 kg wide?

b. Suppose you actually do measure the required number of bags and construct a new confidence interval. Is it guaranteed to be at most 0.1 kg wide?

24.5 Suppose we want to make a 95% confidence interval for the probability of getting heads with a Dutch 1 Euro coin, and it should be at most 0.01 wide. To determine the required sample size, we note that the probability of getting heads is about 0.5. Furthermore, if X has a $Bin(n, p)$ distribution, with n large and $p \approx 0.5$, then

$$\frac{X - np}{\sqrt{n/4}} \quad \text{is approximately standard normal.}$$

a. Use this statement to derive that the width of the 95% confidence interval for p is approximately

$$\frac{z_{0.025}}{\sqrt{n}}.$$

Use this width to determine how large n should be.

b. The coin is thrown the number of times just computed, resulting in $19\,477$ times heads. Construct the 95% confidence interval and check whether the required accuracy is attained.

24.6 ⊞ Environmentalists have taken 16 samples from the wastewater of a chemical plant and measured the concentration of a certain carcinogenic substance. They found $\bar{x}_{16} = 2.24$ (ppm) and $s_{16}^2 = 1.12$, and want to use these data in a lawsuit against the plant. It may be assumed that the data are a realization of a normal random sample.

a. Construct the 97.5% one-sided confidence interval that the environmentalists made to convince the judge that the concentration exceeds legal limits.

b. The plant management uses the same data to construct a 97.5% one-sided confidence interval to show that concentrations are not too high. Construct this interval as well.

24.7 Consider once more the Rutherford-Geiger data as given in Section 23.4. Knowing that the number of α-particle emissions during an interval has a Poisson distribution, we may see the data as observations from a $Pois(\mu)$ distribution. The central limit theorem tells us that the average $\bar{X}_n$ of a large number of independent $Pois(\mu)$ approximately has a normal distribution and hence that

$$\frac{\bar{X}_n - \mu}{\sqrt{\mu}/\sqrt{n}}$$

has a distribution that is approximately $N(0,1)$.

a. Show that the large sample 95% confidence interval contains those values of μ for which

$$(\bar{x}_n - \mu)^2 \leq (1.96)^2 \frac{\mu}{n}.$$

b. Use the result from **a** to construct the large sample 95% confidence interval based on the Rutherford-Geiger data.

c. Compare the result with that of Exercise 23.9 **b**. Is this surprising?

24.8 ▢ Recall Exercise 23.5 about the $1500\,\mathrm{m}$ speed-skating results in the 2002 Winter Olympic Games. If there were no outer lane advantage, the number

out of the 23 completed races won by skaters starting in the outer lane would have a $Bin(23, p)$ distribution with $p = 1/2$, because of the lane assignment by lottery.

a. Of the 23 races, 15 were won by the skater starting in the outer lane. Use this information to construct a 95% confidence interval for p by means of the Wilson method. If you think that $n = 23$ is probably too small to use a method based on the central limit theorem, we agree. We should be careful with conclusions we draw from this confidence interval.

b. The question posed earlier "Is there an outer lane advantage?" implies that a one-sided confidence interval is more suitable. Construct the appropriate 95% one-sided confidence interval for p by first constructing a 90% two-sided confidence interval.

24.9 ⊞ Suppose we have a dataset $x_1, \ldots, x_{12}$ that may be modeled as the realization of a random sample $X_1, \ldots, X_{12}$ from a $U(0, \theta)$ distribution, with θ unknown. Let $M = \max\{X_1, \ldots, X_{12}\}$.

a. Show that for $0 \le t \le 1$

$$P\left(\frac{M}{\theta} \le t\right) = t^{12}.$$

b. Use $\alpha = 0.1$ and solve

$$P\left(\frac{M}{\theta} \le c_l\right) = P\left(\frac{M}{\theta} \le c_u\right) = \tfrac{1}{2}\alpha.$$

c. Suppose the realization of M is $m = 3$. Construct the 90% confidence interval for θ.

d. Derive the general expression for a confidence interval of level $1 - \alpha$ based on a sample of size n.

24.10 Suppose we have a dataset $x_1, \ldots, x_n$ that may be modeled as the realization of a random sample $X_1, \ldots, X_n$ from an $Exp(\lambda)$ distribution, where λ is unknown. Let $S_n = X_1 + \cdots + X_n$.

a. Check that λS_n has a $Gam(n, 1)$ distribution.

b. The following quantiles of the $Gam(20, 1)$ distribution are given: $q_{0.05} = 13.25$ and $q_{0.95} = 27.88$. Use these to construct a 90% confidence interval for λ when $n = 20$.

25

Testing hypotheses: essentials

The statistical methods that we have discussed until now have been developed to infer knowledge about certain features of the model distribution that represent our quantities of interest. These inferences often take the form of numerical estimates, as either single numbers or confidence intervals. However, sometimes the conclusion to be drawn is *not* expressed numerically, but is concerned with choosing between two conflicting theories, or *hypotheses*. For instance, one has to assess whether the lifetime of a certain type of ball bearing deviates or does not deviate from the lifetime guaranteed by the manufacturer of the bearings; an engineer wants to know whether dry drilling is faster or the same as wet drilling; a gynecologist wants to find out whether smoking affects or does not affect the probability of getting pregnant; the Allied Forces want to know whether the German war production is equal to or smaller than what Allied intelligence agencies reported. The process of formulating the possible conclusions one can draw from an experiment and choosing between two alternatives is known as *hypothesis testing*. In this chapter we start to explore this statistical methodology.

25.1 Null hypothesis and test statistic

We will introduce the basic concepts of hypothesis testing with an example. Let us return to the analysis of German war equipment. During World War II the Allied Forces received reports by the Allied intelligence agencies on German war production. The numbers of produced tires, tanks, and other equipment, as claimed in these reports, were a lot higher than indicated by the observed serial numbers. The objective was to decide whether the actual produced quantities were smaller than the ones reported.

For simplicity suppose that we have observed tanks with (recoded) serial numbers

$$61 \quad 19 \quad 56 \quad 24 \quad 16.$$

Furthermore, suppose that the Allied intelligence agencies report a production of 350 tanks.[1] This is a lot more than we would surmise from the observed data. We want to choose between the proposition that the total number of tanks is 350 and the proposition that the total number is smaller than 350. The two competing propositions are called *null hypothesis*, denoted by H_0, and *alternative hypothesis*, denoted by H_1. The way we go about choosing between H_0 and H_1 is conceptually similar to the way a jury deliberates in a court trial. The null hypothesis corresponds to the position of the defendant: just as he is presumed to be innocent until proven guilty, so is the null hypothesis presumed to be true until the data provide convincing evidence against it. The alternative hypothesis corresponds to the charges brought against the defendant.

To decide whether H_0 is false we use a statistical model. As argued in Chapter 20 the (recoded) serial numbers are modeled as a realization of random variables $X_1, X_2, \ldots, X_5$ representing five draws *without replacement* from the numbers $1, 2, \ldots, N$. The parameter N represents the total number of tanks. The two hypotheses in question are

$$H_0 : N = 350$$
$$H_1 : N < 350.$$

If we reject the null hypothesis we will accept H_1; we speak of *rejecting H_0 in favor of H_1*. Usually, the alternative hypothesis represents the theory or belief that we would like to accept if we do reject H_0. This means that we must carefully choose H_1 in relation with our interests in the problem at hand. In our example we are particularly interested in whether the number of tanks is *less* than 350; so we test the null hypothesis against $H_1 : N < 350$. If we would be interested in whether the number of tanks *differs* from 350, or is *greater* than 350, we would test against $H_1 : N \neq 350$ or $H_1 : N > 350$.

QUICK EXERCISE 25.1 In the drilling example from Sections 15.5 and 16.4 the data on drill times for dry drilling are modeled as a realization of a random sample from a distribution with expectation μ_1, and similarly the data for wet drilling correspond to a distribution with expectation μ_2. We want to know whether dry drilling is faster than wet drilling. To this end we test the null hypothesis $H_0 : \mu_1 = \mu_2$ (the drill time is the same for both methods). What would you choose for H_1?

The next step is to select a criterion based on $X_1, X_2, \ldots, X_5$ that provides an indication about whether H_0 is false. Such a criterion involves a test statistic.

---

[1] This may seem ridiculous. However, when after the war official German production statistics became available, the average monthly production of tanks during the period 1940–1943 was 342. During the war this number was estimated at 327, whereas Allied intelligence reported 1550! (see [27]).

> TEST STATISTIC. Suppose the dataset is modeled as the realization of random variables $X_1, X_2, \ldots, X_n$. A *test statistic* is any sample statistic $T = h(X_1, X_2, \ldots, X_n)$, whose numerical value is used to decide whether we reject H_0.

In the tank example we use the test statistic

$$T = \max\{X_1, X_2, \ldots, X_5\}.$$

Having chosen a test statistic T, we investigate what sort of values T can attain. These values can be viewed on a credibility scale for H_0, and we must determine which of these values provide evidence in favor of H_0, and which provide evidence in favor of H_1. First of all note that if we find a value of T larger than 350, we immediately know that H_0 as well as H_1 is false. If this happens, we actually should be considering another testing problem, but for the current problem of testing $H_0 : N = 350$ against $H_1 : N < 350$ such values are irrelevant. Hence the possible values of T that are of interest to us are the integers from 5 to 350.

If H_0 is true, then what is a typical value for T and what is not? Remember from Section 20.1 that, because $n = 5$, the expectation of T is $\mathrm{E}\,[T] = \frac{5}{6}(N+1)$. This means that the distribution of T is centered around $\frac{5}{6}(N+1)$. Hence, if H_0 is true, then typical values of T are in the neighborhood of $\frac{5}{6} \cdot 351 = 292.5$. Values of T that deviate a lot from 292.5 are evidence *against* H_0. Values that are much greater than 292.5 are evidence against H_0 but provide even stronger evidence against H_1. For such values we will *not reject* H_0 in favor of H_1. Also values a little smaller than 292.5 are grounds *not to reject* H_0, because we are committed to giving H_0 the benefit of the doubt. On the other hand, values of T very close to 5 should be considered as strong evidence *against* the null hypothesis and are *in favor* of H_1, hence they lead to a decision to *reject* H_0. This is summarized in Figure 25.1.
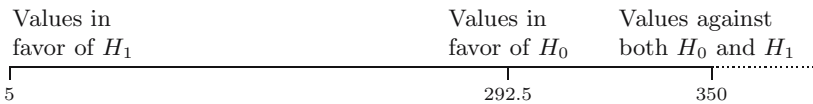
Fig. 25.1. Values of the test statistic T.

QUICK EXERCISE 25.2 Another possible test statistic would be $\bar{X}_5$. If we use its values as a credibility scale for H_0, then what are the possible values of $\bar{X}_5$, which values of $\bar{X}_5$ are in favor of $H_1 : N < 350$, and which values are in favor of $H_0 : N = 350$?

For the data we find

$$t = \max\{61, 19, 56, 24, 16\} = 61$$

as the realization of the test statistic. How do we use this to decide on H_0?

25.2 Tail probabilities

As we have just seen, if H_0 is true, then typical values of T are in the neighborhood of $\frac{5}{6} \cdot 351 = 292.5$. In view of Figure 25.1, the more a value of T is to the left, the stronger evidence it provides in favor of H_1. The value 61 is in the left region of Figure 25.1. Can we now reject H_0 and conclude that N is smaller than 350, or can the fact that we observe 61 as maximum be attributed to chance? In courtroom terminology: can we reach the conclusion that the null hypothesis is *false beyond reasonable doubt*? One way to investigate this is to examine how likely it is that one would observe a value of T that provides even stronger evidence against H_0 than 61, in the situation that $N = 350$. If this is very unlikely, then 61 already bears strong evidence against H_0.

Values of T that provide stronger evidence against H_0 than 61 are to the left of 61. Therefore we compute $P(T \leq 61)$. In the situation that $N = 350$, the test statistic T is the maximum of 5 numbers drawn without replacement from $1, 2, \ldots, 350$. We find that

$$P(T \leq 61) = P(\max\{X_1, X_2, \ldots, X_5\} \leq 61)$$
$$= \frac{61}{350} \cdot \frac{60}{349} \cdots \frac{57}{346} = 0.00014.$$

This probability is so small that we view the value 61 as strong evidence against the null hypothesis. Indeed, if the null hypothesis would be true, then values of T that would provide the same or even stronger evidence against H_0 than 61 are *very unlikely* to occur, i.e., they occur with probability 0.00014! In other words, the observed value 61 is *exceptionally small* in case H_0 is true.

At this point we can do two things: either we believe that H_0 is true and that something very unlikely has happened, or we believe that events with such a small probability do not happen in practice, so that $T \leq 61$ could only have occurred because H_0 is false. We choose to believe that things happening with probability 0.00014 are so exceptional that we *reject* the null hypothesis $H_0 : N = 350$ in favor of the alternative hypothesis $H_1 : N < 350$. In courtroom terminology: we say that a value of T smaller than or equal to 61 implies that the null hypothesis is false *beyond reasonable doubt*.

P-values

In our example, the more a value of T is to the left, the stronger evidence it provides against H_0. For this reason we computed the *left tail probability*

$P(T \leq 61)$. In other situations, the direction in which values of T provide stronger evidence against H_0 may be to the right of the observed value t, in which case one would compute a *right tail probability* $P(T \geq t)$. In both cases the tail probability expresses how likely it is to obtain a value of the test statistic T *at least as extreme as* the value t observed for the data. Such a probability is called a *p-value*. In a way, the size of the *p*-value reflects how much evidence the observed value t provides against H_0. The *smaller* the *p*-value, the *stronger evidence* the observed value t bears against H_0.

The phrase "at least as extreme as the observed value t" refers to a particular direction, namely the direction in which values of T provide stronger evidence against H_0 and in favor of H_1. In our example, this was to the left of 61, and the *p*-value corresponding to 61 was $P(T \leq 61) = 0.00014$. In this case it is clear what is meant by "at least as extreme as t" and which tail probability corresponds to the *p*-value. However, in some testing problems one can deviate from H_0 in *both* directions. In such cases it may not be clear what values of T are at least as extreme as the observed value, and it may be unclear how the *p*-value should be computed. One approach to a solution in this case is to simply compute the *one-tailed p-value* that corresponds to the direction in which t deviates from H_0.

QUICK EXERCISE 25.3 Suppose that the Allied intelligence agencies had reported a production of 80 tanks, so that we would test $H_0 : N = 80$ against $H_1 : N < 80$. Compute the *p*-value corresponding to 61. Would you conclude H_0 is false beyond reasonable doubt?

25.3 Type I and type II errors

Suppose that the maximum is 200 instead of 61. This is also to the left of the expected value 292.5 of T. Is it far enough to the left to reject the null hypothesis? In this case the *p*-value is equal to

$$P(T \leq 200) = P(\max\{X_1, X_2, \ldots, X_5\} \leq 200)$$
$$= \frac{200}{350} \cdot \frac{199}{349} \cdots \frac{196}{346} = 0.0596.$$

This means that *if* the total number of produced tanks is 350, then in 5.96% of all cases we would observe a value of T that is at least as extreme as the value 200. Before we decide whether 0.0596 is small enough to reject the null hypothesis let us explore in more detail what the preceding probability stands for.

It is important to distinguish between (1) the *true state of nature*: H_0 is true or H_1 is true and (2) *our decision*: we reject or do not reject H_0 *on the basis of the data*. In our example the possibilities for the true state of nature are:

• H_0 is true, i.e., there are 350 tanks produced.
• H_1 is true, i.e., the number of tanks produced is less than 350.

We do not know in which situation we are. There are two possible decisions:

- We reject H_0 in favor of H_1.
- We do not reject H_0.

This leads to four possible situations, which are summarized in Figure 25.2.

		True state of nature	
		H_0 is true	H_1 is true
Our decision on the basis of the data	Reject H_0	*Type I error*	Correct decision
	Not reject H_0	Correct decision	*Type II error*

Fig. 25.2. Four situations when deciding about H_0.

There are two situations in which the decision made on the basis of the data is wrong. The null hypothesis H_0 may be true, whereas the data lead to rejection of H_0. On the other hand, the alternative hypothesis H_1 may be true, whereas we do not reject H_0 on the basis of the data. These wrong decisions are called type I and type II errors.

> TYPE I AND II ERRORS.   A *type I error* occurs if we falsely reject H_0. A *type II error* occurs if we falsely do not reject H_0.

In courtroom terminology, a type I error corresponds to convicting an innocent defendant, whereas a type II error corresponds to acquitting a criminal.

If $H_0 : N = 350$ is true, then the decision to reject H_0 is a type I error. We will never know whether we make a type I error. However, given a particular decision rule, we can say something about the probability of committing a type I error. Suppose the decision rule would be "reject $H_0 : N = 350$ whenever $T \leq 200$." With this decision rule the probability of committing a type I error is $P(T \leq 200) = 0.0596$. If we are willing to run the risk of committing a type I error with probability 0.0596, we could adopt this decision rule. This would also mean that on the basis of an observed maximum of 200 we would reject H_0 in favor of $H_1 : N < 350$.

QUICK EXERCISE 25.4 Suppose we adopt the following decision rule about the null hypothesis: "reject $H_0 : N = 350$ whenever $T \leq 250$." Using this decision rule, what is the probability of committing a type I error?

The question remains what amount of risk one is willing to take to falsely reject H_0, or in courtroom terminology: how small should the p-value be to reach a conclusion that is "beyond reasonable doubt"? In many situations, as a rule of thumb 0.05 is used as the level where reasonable doubt begins. Something happening with probability less than or equal to 0.05 is then viewed as being too exceptional. However, there is no general rule that specifies how small the p-value must be to reject H_0. There is no way to argue that this probability *should be* below 0.10 or 0.18 or 0.009—or anything else.

A possible solution is to solely report the p-value corresponding to the observed value of the test statistic. This is objective and does not have the arbitrariness of a preselected level such as 0.05. An investigator who reports the p-value conveys the maximum amount of information contained in the dataset and permits all decision makers to choose their own level and make their own decision about the null hypothesis. This is especially important when there is no justifiable reason for preselecting a particular value for such a level.

25.4 Solutions to the quick exercises

25.1 One is interested in whether dry drilling is *faster* than wet drilling. Hence if we reject $H_0 : \mu_1 = \mu_2$, we would like to conclude that the drill time is *smaller* for dry drilling than for wet drilling. Since μ_1 and μ_2 represent the drill time for dry and wet drilling, we should choose $H_1 : \mu_1 < \mu_2$.

25.2 The value of $\bar{X}_5$ is at least 3 and if we find a value of $\bar{X}_5$ that is larger than 348, then at least one of the five numbers must be greater than 350, so that we immediately know that H_0 as well as H_1 is false. Hence the possible values of $\bar{X}_5$ that are relevant for our testing problem are between 3 and 348. We know from Section 20.1 that $2\bar{X}_5 - 1$ is an unbiased estimator for N, no matter what the value of N is. This implies that values of $\bar{X}_5$ itself are centered around $(N + 1)/2$. Hence values close to $351/2{=}175.5$ are in favor of H_0, whereas values close to 3 are in favor of H_1. Values close to 348 are against H_0, but also against H_1. See Figure 25.3.

Values in favor of H_1	Values in favor of H_0	Values against both H_0 and H_1
3	175.5	348
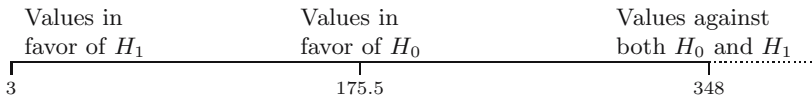
Fig. 25.3. Values of the test statistic $\bar{X}_5$.

25.3 The p-value corresponding to 61 is now equal to

$$P(T \le 61) = \tfrac{61}{80} \cdot \tfrac{60}{79} \cdots \tfrac{57}{76} = 0.2475.$$

If H_0 is true, then in 24.75% of the time one will observe a value T less than or equal to 61. Such values are not exceptionally small for T under H_0, and therefore the evidence that the value 61 bears against H_0 is pretty weak. We cannot reject H_0 beyond reasonable doubt.

25.4 The type I error associated with the decision rule occurs if $N = 350$ (H_0 is true) and $t \leq 250$ (reject H_0). The probability that this happens is $P(T \leq 250) = \frac{250}{350} \cdot \frac{249}{349} \cdots \frac{246}{346} = 0.1838$.

25.5 Exercises

25.1 In a study about train delays in The Netherlands one was interested in whether arrival delays of trains exhibit more variation during rush hours than during quiet hours. The observed arrival delays during rush hours are modeled as realizations of a random sample from a distribution with variance σ_1^2, and similarly the observed arrival delays during quiet hours correspond to a distribution with variance σ_2^2. One tests the null hypothesis $H_0 : \sigma_1 = \sigma_2$. What do you choose as the alternative hypothesis?

25.2 ⊡ On average, the number of babies born in Cleveland, Ohio, in the month of September is 1472. On January 26, 1977, the city was immobilized by a blizzard. Nine months later, in September 1977, the recorded number of births was 1718. Can the increase of 246 be attributed to chance? To investigate this, the number of births in the month of September is modeled by a Poisson random variable with parameter μ, and we test $H_0 : \mu = 1472$. What would you choose as the alternative hypothesis?

25.3 Recall Exercise 17.9 about black cherry trees. The scatterplot of y (volume) versus $x = d^2h$ (squared diameter times height) seems to indicate that the regression line $y = \alpha + \beta x$ runs through the origin. One wants to investigate whether this is true by means of a testing problem. Formulate a null hypothesis and alternative hypothesis in terms of (one of) the parameters α and β.

25.4 ⊞ Consider the example from Section 4.4 about the number of cycles up to pregnancy of smoking and nonsmoking women. Suppose the observed number of cycles are modeled as realizations of random samples from geometric distributions. Let p_1 be the parameter of the geometric distribution corresponding to smoking women and p_2 be the parameter for the nonsmoking women. We are interested in whether p_1 is different from p_2, and we investigate this by testing $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$.

  a. If the data are as given in Exercise 17.5, what would you choose as a test statistic?

b. What would you choose as a test statistic, if you were given the extra knowledge as in Table 21.1?

c. Suppose we are interested in whether smoking women are less likely to get pregnant than nonsmoking women. What is the appropriate alternative hypothesis in this case?

25.5 ☐ Suppose a dataset is a realization of a random sample $X_1, X_2, \ldots, X_n$ from a uniform distribution on $[0, \theta]$, for some (unknown) $\theta > 0$. We test $H_0 : \theta = 5$ versus $H_1 : \theta \neq 5$.

a. We take $T_1 = \max\{X_1, X_2, \ldots, X_n\}$ as our test statistic. Specify what the (relevant) possible values are for T and which are in favor of H_0 and which are in favor of H_1. For instance, make a picture like Figure 25.1.

b. Same as **a**, but now for test statistic $T_2 = |2\bar{X}_n - 5|$.

25.6 ☐ To test a certain null hypothesis H_0 one uses a test statistic T with a continuous sampling distribution. One agrees that H_0 is rejected if one observes a value t of the test statistic for which (under H_0) the right tail probability $P(T \geq t)$ is smaller than or equal to 0.05. Given below are different values t and a corresponding left or right tail probability (under H_0). Specify for each case what the p-value is, if possible, and whether we should reject H_0.

a. $t = 2.34$ and $P(T \geq 2.34) = 0.23$.

b. $t = 2.34$ and $P(T \leq 2.34) = 0.23$.

c. $t = 0.03$ and $P(T \geq 0.03) = 0.968$.

d. $t = 1.07$ and $P(T \leq 1.07) = 0.981$.

e. $t = 1.07$ and $P(T \leq 2.34) = 0.01$.

f. $t = 2.34$ and $P(T \leq 1.07) = 0.981$.

g. $t = 2.34$ and $P(T \leq 1.07) = 0.800$.

25.7 (Exercise 25.2 continued). The number of births in September is modeled by a Poisson random variable T with parameter μ, which represents the expected number of births. Suppose that one uses T to test the null hypothesis $H_0 : \mu = 1472$ and that one decides to reject H_0 on the basis of observing the value $t = 1718$.

a. In which direction do values of T provide evidence against H_0 (and in favor of H_1)?

b. Compute the p-value corresponding to $t = 1718$, where you may use the fact that the distribution of T can be approximated by an $N(\mu, \mu)$ distribution.

25.8 Suppose we want to test the null hypothesis that our dataset is a realization of a random sample from a standard normal distribution. As test statistic we use the Kolmogorov-Smirnov distance between the empirical distribution

function F_n of the data and the distribution function Φ of the standard normal:

$$T = \sup_{a \in \mathbb{R}} |F_n(a) - \Phi(a)|.$$

What are the possible values of T and in which direction do values of T deviate from the null hypothesis?

25.9 Recall the example from Section 18.3, where we investigated whether the software data are exponential by means of the Kolmogorov-Smirnov distance between the empirical distribution function F_n of the data and the estimated exponential distribution function:

$$T_{\mathrm{ks}} = \sup_{a \in \mathbb{R}} |F_n(a) - (1 - \mathrm{e}^{-\hat{\Lambda}a})|.$$

For the data we found $t_{\mathrm{ks}} = 0.176$. By means of a new parametric bootstrap we simulated $100\,000$ realizations of T_{ks} and found that all of them are smaller than 0.176. What can you say about the p-value corresponding to 0.176?

25.10 ⊞ Consider the coal data from Table 23.1, where 23 gross calorific value measurements are listed for Osterfeld coal coded 262DE27. We modeled this dataset as a realization of a random sample from a normal distribution with expectation μ unknown and standard deviation 0.1 MJ/kg. We are planning to buy a shipment if the gross calorific value exceeds 23.75 MJ/kg. In order to decide whether this is sensible, we test the null hypothesis $H_0 : \mu = 23.75$ with test statistic $\bar{X}_n$.

 a. What would you choose as the alternative hypothesis?
 b. For the dataset $\bar{x}_n$ is 23.788. Compute the corresponding p-value, using that $\bar{X}_n$ has an $N(23.75, (0.1)^2/23)$ distribution under the null hypothesis.

25.11 ⊞ One is given a number t, which is the realization of a random variable T with an $N(\mu, 1)$ distribution. To test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$, one uses T as the test statistic. One decides to reject H_0 in favor of H_1 if $|t| \geq 2$. Compute the probability of committing a type I error.

Testing hypotheses: elaboration

In the previous chapter we introduced the setup for testing a null hypothesis against an alternative hypothesis using a test statistic T. The notions of type I error and type II error were introduced. A type I error occurs when we falsely reject H_0 on the basis of the observed value of T, whereas a type II error occurs when we falsely do not reject H_0. The decision to reject H_0 or not was based on the size of the p-value. In this chapter we continue the introduction of basic concepts of testing hypotheses, such as *significance level* and *critical region*, and investigate the probability of committing a type II error.

26.1 Significance level

As mentioned in the previous chapter, there is no general rule that specifies a level below which the p-value is considered exceptionally small. However, there are situations where this level is set *a priori*, and the question is: which values of the test statistic should then lead to rejection of H_0? To illustrate this, consider the following example. The speed limit on freeways in The Netherlands is 120 kilometers per hour. A device next to freeway A2 between Amsterdam and Utrecht measures the speed of passing vehicles. Suppose that the device is designed in such a way that it conducts three measurements of the speed of a passing vehicle, modeled by a random sample X_1, X_2, X_3. On the basis of the value of the average $\bar{X}_3$, the driver is either fined for speeding or not. For what values of $\bar{X}_3$ should we fine the driver, if we allow that 5% of the drivers are fined unjustly?

Let us rephrase things in terms of a testing problem. Each measurement can be thought of as

$$\text{measurement} = \text{true speed} + \text{measurement error}.$$

Suppose for the moment that the measuring device is carefully calibrated, so that the measurement error is modeled by a random variable with mean zero

and known variance σ^2, say $\sigma^2 = 4$. Moreover, in physical experiments such as this one, the measurement error is often modeled by a random variable with a normal distribution. In that case, the measurements X_1, X_2, X_3 are modeled by a random sample from an $N(\mu, 4)$ distribution, where the parameter μ represents the true speed of the passing vehicle. Our testing problem can now be formulated as testing

$$H_0 : \mu = 120 \quad \text{against} \quad H_1 : \mu > 120,$$

with test statistic

$$T = \frac{X_1 + X_2 + X_3}{3} = \bar{X}_3.$$

Since sums of independent normal random variables again have a normal distribution (see Remark 11.2), it follows that $\bar{X}_3$ has an $N(\mu, 4/3)$ distribution. In particular, the distribution of $T = \bar{X}_3$ is centered around μ no matter what the value of μ is. Values of T close to 120 are therefore in favor of H_0. Values of T that are far from 120 are considered as strong evidence against H_0. Values much larger than 120 suggest that $\mu > 120$ and are therefore in favor of H_1. Values much smaller than 120 suggest that $\mu < 120$. They also constitute evidence against H_0, but even stronger evidence against H_1. Thus we reject H_0 in favor of H_1 *only* for values of T larger than 120. See also Figure 26.1.

Fig. 26.1. Possible values of $T = \bar{X}_3$.

Rejection of H_0 in favor of H_1 corresponds to fining the driver for speeding. Unjustly fining a driver corresponds to falsely rejecting H_0, i.e., committing a type I error. Since we allow 5% of the drivers to be fined unjustly, we are dealing with a testing problem where the probability of committing a type I error is set a priori at 0.05. The question is: for which values of T should we reject H_0? The decision rule for rejecting H_0 should be such that the corresponding probability of committing a type I error is 0.05. The value 0.05 is called the significance level.

> SIGNIFICANCE LEVEL. The *significance level* is the largest acceptable probability of committing a type I error and is denoted by α, where $0 < \alpha < 1$.

We speak of "performing the test at level α," as well as "rejecting H_0 in favor of H_1 at level α." In our example we are testing $H_0 : \mu = 120$ against $H_1 : \mu > 120$ at level 0.05.

QUICK EXERCISE 26.1 Suppose that in the freeway example $H_0 : \mu = 120$ is rejected in favor of $H_1 : \mu > 120$ at level $\alpha = 0.05$. Will it necessarily be rejected at level $\alpha = 0.01$? On the other hand, suppose that $H_0 : \mu = 120$ is rejected in favor of $H_1 : \mu > 120$ at level $\alpha = 0.01$. Will it necessarily be rejected at level $\alpha = 0.05$?

Let us continue with our example and determine for which values of $T = \bar{X}_3$ we should reject H_0 at level $\alpha = 0.05$ in favor of $H_1 : \mu > 120$. Suppose we decide to fine each driver whose recorded average speed is 121 or more, i.e., we reject H_0 whenever $T \geq 121$. Then how large is the probability of a type I error $P(T \geq 121)$? When $H_0 : \mu = 120$ is true, then $T = \bar{X}_3$ has an $N(120, 4/3)$ distribution, so that by the change-of-units rule for the normal distribution (see page 106), the random variable

$$Z = \frac{T - 120}{2/\sqrt{3}}$$

has an $N(0, 1)$ distribution. This implies that

$$P(T \geq 121) = P\left(\frac{T - 120}{2/\sqrt{3}} \geq \frac{121 - 120}{2/\sqrt{3}}\right) = P(Z \geq 0.87).$$

From Table B.1, we find $P(Z \geq 0.87) = 0.1922$, which means that the probability of a type I error is greater than the significance level $\alpha = 0.05$. Since this level was defined as the largest acceptable probability of a type I error, we do not reject H_0. Similarly, if we decide to reject H_0 whenever we record an average of 122 or more, the probability of a type I error equals 0.0416 (check this). This is smaller than $\alpha = 0.05$, so in that case we reject H_0. The boundary case is the value c that satisfies $P(T \geq c) = 0.05$. To find c, we must solve

$$P\left(Z \geq \frac{c - 120}{2/\sqrt{3}}\right) = 0.05.$$

From Table B.2 we have that $z_{0.05} = t_{\infty, 0.05} = 1.645$, so that we find

$$\frac{c - 120}{2/\sqrt{3}} = 1.645,$$

which leads to

$$c = 120 + 1.645 \cdot \frac{2}{\sqrt{3}} = 121.9.$$

Hence, if we set the significance level α at 0.05, we should reject $H_0 : \mu = 120$ in favor of $H_1 : \mu > 120$ whenever $T \geq 121.9$. For our freeway example this means that if the average recorded speed of a passing vehicle is greater than or equal to 121.9, then the driver is fined for speeding. With this decision rule, at most 5% of the drivers get fined unjustly.

In connection with p-values: the significance level is the level below which the p-value is sufficiently small to reject H_0. Indeed, for any observed value $t \geq 121.9$ we reject H_0, and the p-value for such a t is at most 0.05:

$$P(T \geq t) \leq P(T \geq 121.9) = 0.05.$$

We will see more about this relation in the next section.

26.2 Critical region and critical values

In the freeway example the significance level 0.05 corresponds to the decision rule "reject $H_0 : \mu = 120$ in favor $H_1 : \mu > 120$ whenever $T \geq 121.9$." The set $K = [121.9, \infty)$ consisting of values of the test statistic T for which we reject H_0 is called critical region. The value 121.9, which is the boundary case between rejecting and not rejecting H_0, is called the critical value.

> CRITICAL REGION AND CRITICAL VALUES. Suppose we test H_0 against H_1 at significance level α by means of a test statistic T. The set $K \subset \mathbb{R}$ that corresponds to all values of T for which we reject H_0 in favor of H_1 is called the *critical region*. Values on the boundary of the critical region are called *critical values*.

The precise shape of the critical region depends on both the chosen significance level α and the test statistic T that is used. But it will always be such that the probability that $T \in K$ satisfies

$$P(T \in K) \leq \alpha \quad \text{in the case that } H_0 \text{ is true.}$$

At this point it becomes important to emphasize whether probabilities are computed under the assumption that H_0 is true. With a slight abuse of notation, we briefly write $P(T \in K \mid H_0)$ for the probability.

Relation with p-values

If we record average speed $t = 124$, then this value falls in the critical region $K = [121.9, \infty)$, so that $H_0 : \mu = 120$ is rejected in favor $H_1 : \mu > 120$. On the other hand we can also compute the p-value corresponding to the observed value 124. Since values of T to the right provide stronger evidence against H_0, the p-value is the following right tail probability

$$P(T \geq 124 \mid H_0) = P\left(\frac{T - 120}{2/\sqrt{3}} \geq \frac{124 - 120}{2/\sqrt{3}}\right) = P(Z \geq 3.46) = 0.0003,$$

which is smaller than the significance level 0.05. This is no coincidence.

In general, suppose that we perform a test at level α using test statistic T and that we have observed t as the value of our test statistic. Then

$t \in K \quad \Leftrightarrow \quad$ the p-value corresponding to t is less than or equal to α.

Figure 26.2 illustrates this for a testing problem where values of T to the right provide evidence against H_0 and in favor of H_1. In that case, the p-value corresponds to the right tail probability $P(T \geq t \mid H_0)$. The shaded area to the right of c_α corresponds to $\alpha = P(T \geq c_\alpha \mid H_0)$, whereas the more intensely shaded area to the right of t represents the p-value. We see that deciding whether to reject H_0 at a given significance level α can be done by comparing either t with c_α or the p-value with α. For this reason the p-value is sometimes called the *observed significance level*.
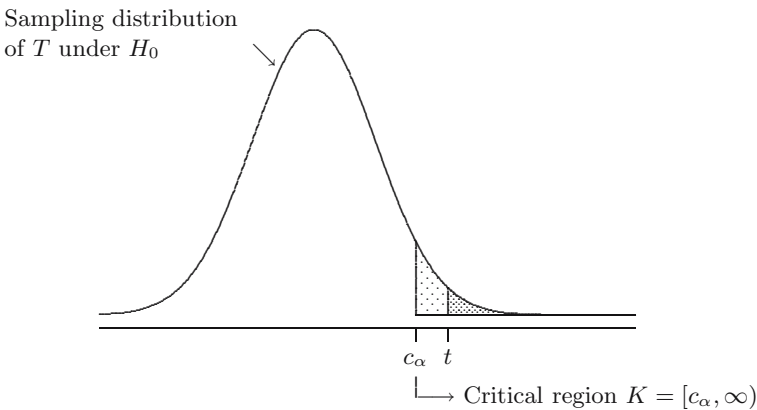
Fig. 26.2. *P*-value and critical value.

The concepts of critical value and p-value have their own merit. The critical region and the corresponding critical values specify exactly what values of T lead to rejection of H_0 at a given level α. This can be done even without obtaining a dataset and computing the value t of the test statistic. The p-value, on the other hand, represents the strength of the evidence the observed value t bears against H_0. But it does not specify all values of T that lead to rejection of H_0 at a given level α.

QUICK EXERCISE 26.2 In our freeway example, we have already computed the relevant tail probability to decide whether a person with recorded average speed $t = 124$ gets fined if we set the significance level at 0.05. Suppose the significance level is set at $\alpha = 0.01$ (we allow 1% of the drivers to get fined unjustly). Determine whether a person with recorded average speed $t = 124$ gets fined ($H_0 : \mu = 120$ is rejected). Furthermore, determine the critical region in this case.

Sometimes the critical region K can be constructed such that $P(T \in K \mid H_0)$ is exactly equal to α, as in the freeway example. However, when the distribution of T is discrete, this is not always possible. This is illustrated by the next example.

After the introduction of the Euro, Polish mathematicians claimed that the Belgian 1 Euro coin is not a fair coin (see, for instance, the *New Scientist*, January 4, 2002). Suppose we put a 1 Euro coin to the test. We will throw it ten times and record X, the number of heads. Then X has a $Bin(10, p)$ distribution, where p denotes the probability of heads. We like to find out whether p differs from $1/2$. Therefore we test

$$H_0 : p = \frac{1}{2} \text{ (the coin is fair)} \quad \text{against} \quad H_1 : p \neq \frac{1}{2} \text{ (the coin is not fair).}$$

We use X as the test statistic. When we set the significance level α at 0.05, for what values of X will we reject H_0 and conclude that the coin is not fair? Let us first find out what values of X are in favor of H_1. If $H_0 : p = 1/2$ is true, then $E[X] = 10 \cdot \frac{1}{2} = 5$, so that values of X close to 5 are in favor H_0. Values close to 10 suggest that $p > 1/2$ and values close to 0 suggest that $p < 1/2$. Hence, both values close to 0 and values close to 10 are in favor of $H_1 : p \neq 1/2$.
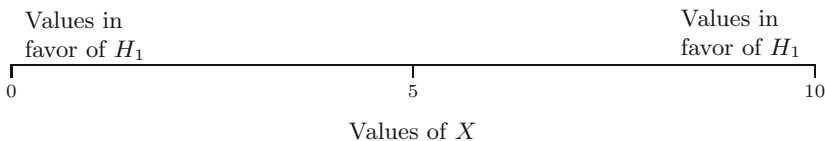
This means that we will reject H_0 in favor of H_1 whenever $X \leq c_l$ or $X \geq c_u$. Therefore, the critical region is the set

$$K = \{0, 1, \dots, c_l\} \cup \{c_u, \dots, 9, 10\}.$$

The boundary values c_l and c_u are called *left* and *right* critical values. They must be chosen such that the critical region K is as large as possible and still satisfies

$$P(X \in K \mid H_0) = P\big(X \leq c_l \mid p = \tfrac{1}{2}\big) + P\big(X \geq c_u \mid p = \tfrac{1}{2}\big) \leq 0.05.$$

Here $P\big(X \geq c_u \mid p = \tfrac{1}{2}\big)$ denotes the probability $P(X \geq c_u)$ computed with X having a $Bin(10, \tfrac{1}{2})$ distribution. Since we have no preference for rejecting H_0 for values close to 0 or close to 10, we divide 0.05 over the two sides, and we choose c_l as large as possible and c_u as small as possible such that

$$P\big(X \leq c_l \mid p = \tfrac{1}{2}\big) \leq 0.025 \quad \text{and} \quad P\big(X \geq c_u \mid p = \tfrac{1}{2}\big) \leq 0.025.$$

Table 26.1. Left tail probabilities of the $Bin(10, \frac{1}{2})$ distribution.

k	$P(X \leq k)$	k	$P(X \leq k)$
0	0.00098	6	0.82813
1	0.01074	7	0.94531
2	0.05469	8	0.98926
3	0.17188	9	0.99902
4	0.37696	10	1.00000
5	0.62305		

The left tail probabilities of the $Bin(10, \frac{1}{2})$ distribution are listed in Table 26.1. We immediately see that $c_l = 1$ is the largest value such that $P(X \leq c_l \mid p = 1/2) \leq 0.025$. Similarly, $c_u = 9$ is the smallest value such that $P(X \geq c_u \mid p = 1/2) \leq 0.025$. Indeed, when X has a $Bin(10, \frac{1}{2})$ distribution,

$$P(X \geq 9) = 1 - P(X \leq 8) = 1 - 0.98926 = 0.01074,$$
$$P(X \geq 8) = 1 - P(X \leq 7) = 1 - 0.94531 = 0.05469.$$

Hence, if we test $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$ at level $\alpha = 0.05$, the critical region is the set $K = \{0, 1, 9, 10\}$. The corresponding type I error is

$$P(X \in K) = P(X \leq 1) + P(X \geq 9) = 0.01074 + 0.01074 = 0.02148,$$

which is smaller than the significance level. You may perform ten throws with your favorite coin and see whether the number of heads falls in the critical region.

QUICK EXERCISE 26.3 Recall the tank example where we tested $H_0 : N = 350$ against $H_1 : N < 350$ by means of the test statistic $T = \max X_i$. Suppose that we perform the test at level 0.05. Deduce the critical region K corresponding to level 0.05 from the left tail probabilities given here:

k	195	194	193	192	191
$P(T \leq k \mid H_0)$	0.0525	0.0511	0.0498	0.0485	0.0472

Is $P(T \in K \mid H_0) = 0.05$?

One- and two-tailed p-values

In the Euro coin example, we deviate from $H_0 : p = 1/2$ in *two* directions: values of X both far to the right and far to the left of 5 are evidence against H_0. Suppose that in ten throws with the 1 Euro coin we recorded x heads. What would the p-value be corresponding to x? The problem is that the direction in which values of X are *at least as extreme as* the observed value x depends on whether x lies to the right or to the left of 5.

At this point there are two natural solutions. One may report the appropriate left or right tail probability, which corresponds to the direction in which x deviates from H_0. For instance, if x lies to the right of 5, we compute $P(X \geq x \mid H_0)$. This is called a *one-tailed p-value*. The disadvantage of one-tailed p-values is that they are somewhat misleading about how strong the evidence of the observed value x bears against H_0. In view of the relation between rejection on the basis of critical values or on the basis of a p-value, the one-tailed p-value should be compared to $\alpha/2$. On the other hand, since people are inclined to compare p-values with the significance level α itself, one could also double the one-tailed p-value and compare this with α. This double-tail probability is called a *two-tailed p-value*. It doesn't make much of a difference, as long as one *also* reports whether the reported p-value is one-tailed or two-tailed.

Let us illustrate things by means of the findings by the Polish mathematicians. They performed 250 throws with a Belgian 1 Euro coin and recorded heads 140 times (see also Exercise 24.2). The question is whether this provides strong enough evidence against $H_0 : p = 1/2$. The observed value 140 is to the right of 125, the value we would expect if H_0 is true. Hence the one-tailed p-value is $P(X \geq 140)$, where now X has a $Bin(250, \frac{1}{2})$ distribution. By means of the normal approximation (see page 201), we find

$$P(X \geq 140) = P\left(\frac{X - 125}{\sqrt{\frac{1}{4}}\sqrt{250}} \geq \frac{140 - 125}{\sqrt{\frac{1}{4}}\sqrt{250}} \right)$$

$$\approx P(Z \geq 1.90) = 1 - \Phi(1.90) = 0.0287.$$

Therefore the two-tailed p-value is approximately 0.0574, which does not provide very strong evidence against H_0. In fact, the *exact* two-tailed p-value, computed by means of statistical software, is 0.066, which is even larger.

QUICK EXERCISE 26.4 In a Dutch newspaper (*De Telegraaf*, January 3, 2002) it was reported that the Polish mathematicians recorded heads 150 times. What are the one- and two-tailed probabilities is this case? Do they now have a case?

26.3 Type II error

As we have just seen, by setting a significance level α, we are able to control the probability of committing a type I error; it will at most be α. For instance, let us return to the freeway example and suppose that we adopt the decision rule to fine the driver for speeding if her average observed speed is at least 121.9, i.e.,

reject $H_0 : \mu = 120$ in favor of $H_1 : \mu > 120$ whenever $T = \bar{X}_3 \geq 121.9$.

From Section 26.1 we know that with this decision rule, the probability of a
type I error is 0.05. What is the probability of committing a type II error?
This corresponds to the percentage of drivers whose true speed *is* above 120
but who do not get fined because their recorded average speed is below 121.9.

For instance, suppose that a car passes at true speed $\mu = 125$. A type II error
occurs when $T < 121.9$, and since $T = \bar{X}_3$ has an $N(125, 4/3)$ distribution,
the probability that this happens is

$$P(T < 121.9 \mid \mu = 125) = P\left(\frac{T - 125}{2/\sqrt{3}} < \frac{121.9 - 125}{2/\sqrt{3}} \right)$$
$$= \Phi(-2.68) = 0.0036.$$

This looks promising, but now consider a vehicle passing at true speed $\mu = 123$. The probability of committing a type II error in this case is

$$P(T < 121.9 \mid \mu = 123) = P\left(\frac{T - 123}{2/\sqrt{3}} < \frac{121.9 - 123}{2/\sqrt{3}} \right)$$
$$= \Phi(-0.95) = 0.1711.$$

Hence 17.11% of all drivers that pass at speed $\mu = 123$ will not get fined. In
Figure 26.3 the last situation is illustrated. The curve on the left represents the
probability density of the $N(120, 4/3)$ distribution, which is the distribution
of T under the null hypothesis. The shaded area on the right of 121.9 represents
the probability of committing a type I error

$$P(T \geq 121.9 \mid \mu = 120) = 0.05.$$

The curve on the right is the probability density of the $N(123, 4/3)$ distribu-
tion, which is the distribution of T under the alternative $\mu = 123$. The shaded
area on the left of 121.9 represents the probability of a type II error
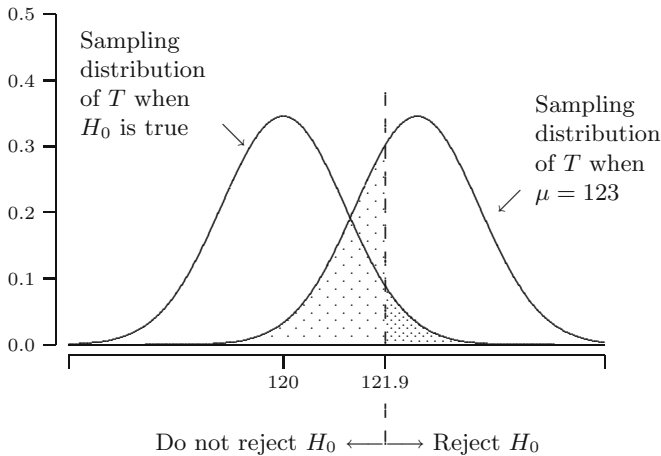
Fig. 26.3. Type I and type II errors in the freeway example.

$$P(T < 121.9 \mid \mu = 123) = 0.1711.$$

Shifting μ further to the right will result in a smaller probability of a type II error. However, shifting μ toward the value 120 leads to a larger probability of a type II error. In fact it can be arbitrarily close to 0.95.

The previous example illustrates that the probability of committing a type II error depends on the actual value of μ in the alternative hypothesis $H_1 : \mu > 120$. The closer μ is to 120, the higher the probability of a type II error will be. In contrast with the probability of a type I error, which is always at most α, the probability of a type II error may be arbitrarily close to $1 - \alpha$. This is illustrated in the next quick exercise.

QUICK EXERCISE 26.5 What is the probability of a type II error in the freeway example if $\mu = 120.1$?

26.4 Relation with confidence intervals

When testing $H_0 : \mu = 120$ against $H_1 : \mu > 120$ at level 0.05 in the freeway example, the critical value was obtained by the formula

$$c_{0.05} = 120 + 1.645 \cdot \frac{2}{\sqrt{3}}.$$

On the other hand, using that $\bar{X}_3$ has an $N(\mu, 4/3)$ distribution, a 95% lower confidence bound for μ in this case can be derived from

$$l_n = \bar{x}_3 - 1.645 \cdot \frac{2}{\sqrt{3}}.$$

Although, at first sight, testing hypotheses and constructing confidence intervals seem to be two separate statistical procedures, they are in fact intimately related. In the freeway example, observe that for a given dataset x_1, x_2, x_3,

we reject $H_0 : \mu = 120$ in favor of $H_1 : \mu > 120$ at level 0.05

$\Leftrightarrow \quad \bar{x}_3 \geq 120 + 1.645 \cdot \dfrac{2}{\sqrt{3}}$

$\Leftrightarrow \quad \bar{x}_3 - 1.645 \cdot \dfrac{2}{\sqrt{3}} \geq 120$

$\Leftrightarrow \quad$ 120 is not in the 95% one-sided confidence interval for μ.

This is not a coincidence. In general, the following applies. Suppose that for some parameter θ we test $H_0 : \theta = \theta_0$. Then

we reject $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta > \theta_0$ at level α

if and only if

θ_0 is not in the $100(1 - \alpha)\%$ *one-sided* confidence interval for θ.

The same relation holds for testing against $H_1 : \theta < \theta_0$, and a similar relation holds between testing against $H_1 : \theta \neq \theta_0$ and two-sided confidence intervals:

> we reject $H_0 : \theta = \theta_0$ in favor of $H_1 : \theta_0 \neq \theta_0$ at level α
>
> if and only if
>
> θ_0 is not in the $100(1 - \alpha)\%$ *two-sided* confidence region for θ.

In fact, one could use these facts to define the $100(1-\alpha)\%$ *confidence region* for a parameter θ as the set of values θ_0 for which the null hypothesis $H_0 : \theta = \theta_0$ is *not rejected* at level α.

It should be emphasized that these relations *only* hold if the random variable that is used to construct the confidence interval relates appropriately to the test statistic. For instance, the preceding relations do not hold if on the one hand, we construct a confidence interval for the parameter μ of an $N(\mu, \sigma^2)$ distribution by means of the studentized mean $(\bar{X}_n - \mu)/(S_n/\sqrt{n})$, and on the other hand, use the sample median Med_n to test a null hypothesis for μ.

26.5 Solutions to the quick exercises

26.1 In the first situation, we reject at significance level $\alpha = 0.05$, which means that the probability of committing a type I error is at most 0.05. This does not necessarily mean that this probability will also be less than or equal to 0.01. Therefore with this information we cannot know whether we also reject at level $\alpha = 0.01$. In the reversed situation, if we reject at level $\alpha = 0.01$, then the probability of committing a type I error is at most 0.01, and is therefore also smaller than 0.05. This means that we also reject at level $\alpha = 0.05$.

26.2 To decide whether we should reject $H_0 : \mu = 120$ at level 0.01, we could compute $P(T \geq 124 \mid H_0)$ and compare this with 0.01. We have already seen that $P(T \geq 124 \mid H_0) = 0.0003$. This is (much) smaller than the significance level $\alpha = 0.01$, so we should reject.

The critical region is $K = [c, \infty)$, where we must solve c from

$$P\left(Z \geq \frac{c - 120}{2/\sqrt{3}}\right) = 0.01.$$

Since $z_{0.01} = 2.326$, this means that $c = 120 + 2.326 \cdot (2/\sqrt{3}) = 122.7$.

26.3 The critical region is of the form $K = \{5, 6, \ldots, c\}$, where the critical value c is the largest value, for which $P(T \leq c \mid H_0)$ is still less than or equal to 0.05. From the table we immediately see that $c = 193$ and that $P(T \in K \mid H_0) = P(T \leq 193 \mid H_0) = 0.0498$, which is not equal to 0.05.

26.4 By means of the normal approximation, for the one-tailed p-value we find

$$\mathrm{P}(X \geq 150) = \mathrm{P}\left(\frac{X - 125}{\sqrt{\frac{1}{4}}\sqrt{250}} \geq \frac{150 - 125}{\sqrt{\frac{1}{4}}\sqrt{250}}\right)$$
$$= \mathrm{P}(Z_n \geq 3.16) \approx 1 - \Phi(3.16) = 0.0008.$$

The two-tailed p-value is 0.0016. This is a lot smaller than the two-tailed p-value 0.0574, corresponding to 140 heads. It seems that with 150 heads the mathematicians would have a case; the Belgian Euro coin would then appear not to be fair.

26.5 The probability of a type II error is

$$\mathrm{P}(T < 121.9 \mid \mu = 120.1) = \mathrm{P}\left(\frac{T - 120.1}{2/\sqrt{3}} < \frac{121.9 - 120.1}{2/\sqrt{3}}\right)$$
$$= \Phi(1.56) = 0.9406.$$

26.6 Exercises

26.1 Polygraphs that are used in criminal investigations are supposed to indicate whether a person is lying or telling the truth. However the procedure is not infallible, as is illustrated by the following example. An experienced polygraph examiner was asked to make an overall judgment for each of a total 280 records, of which 140 were from guilty suspects and 140 from innocent suspects. The results are listed in Table 26.2. We view each judgment as a problem of hypothesis testing, with the null hypothesis corresponding to "suspect is innocent" and the alternative hypothesis to "suspect is guilty." Estimate the probabilities of a type I error and a type II error that apply to this polygraph method on the basis of Table 26.2.

26.2 Consider the testing problem in Exercise 25.11. Compute the probability of committing a type II error if the true value of μ is 1.

26.3 □ One generates a number x from a uniform distribution on the interval $[0, \theta]$. One decides to test $H_0 : \theta = 2$ against $H_1 : \theta \neq 2$ by rejecting H_0 if $x \leq 0.1$ or $x \geq 1.9$.

a. Compute the probability of committing a type I error.

b. Compute the probability of committing a type II error if the true value of θ is 2.5.

26.4 To investigate the hypothesis that a horse's chances of winning an eight-horse race on a circular track are affected by its position in the starting lineup,

Table 26.2. Examiners and suspects.

		Suspect's true status	
		Innocent	Guilty
Examiner's assesment	Acquitted	131	15
	Convicted	9	125

Source: F.S. Horvath and J.E. Reid. The reliability of polygraph examiner diagnosis of truth and deception. *Journal of Criminal Law, Criminology, and Police Science*, 62(2):276–281, 1971.

the starting position of each of 144 winners was recorded ([30]). It turned out that 29 of these winners had starting position one (closest to the rail on the inside track). We model the number of winners with starting position one by a random variable T with a $Bin(144, p)$ distribution. We test the hypothesis $H_0 : p = 1/8$ against $H_1 : p > 1/8$ at level $\alpha = 0.01$ with T as test statistic.

a. Argue whether the test procedure involves a right critical value, a left critical value, or both.

b. Use the normal approximation to compute the critical value(s) corresponding to $\alpha = 0.01$, determine the critical region, and report your conclusion about the null hypothesis.

26.5 ⊞ Recall Exercises 23.5 and 24.8 about the 1500 m speed-skating results in the 2002 Winter Olympic Games. The number of races won by skaters starting in the outer lane is modeled by a random variable X with a $Bin(23, p)$ distribution. The question of whether there is an outer lane advantage was investigated in Exercise 24.8 by means of constructing confidence intervals using the normal approximation. In this exercise we examine this question by testing the null hypothesis $H_0 : p = 1/2$ against $H_1 : p > 1/2$ using X as the test statistic. The distribution of X under H_0 is given in Table 26.3. Out of 23 completed races, 15 were won by skaters starting in the outer lane.

a. Compute the p-value corresponding to $x = 15$ and report your conclusion if we perform the test at level 0.05. Does your conclusion agree with the confidence interval you found for p in Exercise 24.8 **b**?

b. Determine the critical region corresponding to significance level $\alpha = 0.05$.

c. Compute the probability of committing a type I error if we base our decision rule on the critical region determined in **b**.

Table 26.3. Left tail probabilities for the $Bin(23, \frac{1}{2})$ distribution.

k	P$(X \le k)$	k	P$(X \le k)$	k	P$(X \le k)$
0	0.0000	8	0.1050	16	0.9827
1	0.0000	9	0.2024	17	0.9947
2	0.0000	10	0.3388	18	0.9987
3	0.0002	11	0.5000	19	0.9998
4	0.0013	12	0.6612	20	1.0000
5	0.0053	13	0.7976	21	1.0000
6	0.0173	14	0.8950	22	1.0000
7	0.0466	15	0.9534	23	1.0000

d. Use the normal approximation to determine the probability of committing a type II error for the case $p = 0.6$, if we base our decision rule on the critical region determined in **b**.

26.6 ⊡ Consider Exercises 25.2 and 25.7. One decides to test $H_0 : \mu = 1472$ against $H_1 : \mu > 1472$ at level $\alpha = 0.05$ on the basis of the recorded value 1718 of the test statistic T.

a. Argue whether the test procedure involves a right critical value, a left critical value, or both.

b. Use the fact that the distribution of T can be approximated by an $N(\mu, \mu)$ distribution to determine the critical value(s) and the critical region, and report your conclusion about the null hypothesis.

26.7 A random sample X_1, X_2 is drawn from a uniform distribution on the interval $[0, \theta]$. We wish to test $H_0 : \theta = 1$ against $H_1 : \theta < 1$ by rejecting if $X_1 + X_2 \le c$. Find the value of c and the critical region that correspond to a level of significance 0.05.
Hint: use Exercise 11.5.

26.8 ⊞ This exercise is meant to illustrate that the shape of the critical region is not necessarily similar to the type of alternative hypothesis. The type of alternative hypothesis *and* the test statistic used determine the shape of the critical region.

Suppose that $X_1, X_2, \ldots, X_n$ form a random sample from an $Exp(\lambda)$ distribution, and we test $H_0 : \lambda = 1$ with test statistics $T = \bar{X}_n$ and $T' = e^{-\bar{X}_n}$.

a. Suppose we test the null hypothesis against $H_1 : \lambda > 1$. Determine for both test procedures whether they involve a right critical value, a left critical value, or both.

b. Same question as in part **a**, but now test against $H_1 : \lambda \ne 1$.

26.9 ⊞ Similar to Exercise 26.8, but with a random sample $X_1, X_2, \ldots, X_n$ from an $N(\mu, 1)$ distribution. We test $H_0 : \mu = 0$ with test statistics $T = (\bar{X}_n)^2$ and $T' = 1/\bar{X}_n$.

a. Suppose that we test the null hypothesis against $H_1 : \mu \neq 0$. Determine the shape of the critical region for both test procedures.

b. Same question as in part **a**, but now test against $H_1 : \mu > 0$.

The *t*-test

In many applications the quantity of interest can be represented by the expectation of the model distribution. In some of these applications one wants to know whether this expectation deviates from some a priori specified value. This can be investigated by means of a statistical test, known as the *t*-test. We consider this test both under the assumption that the model distribution is normal and without the assumption of normality. Furthermore, we discuss a similar test for the slope and the intercept in a simple linear regression model.

27.1 Monitoring the production of ball bearings

A production line in a large industrial corporation are set to produce a specific type of steel ball bearing with a diameter of 1 millimeter. In order to check the performance of the production lines, a number of ball bearings are picked at the end of the day and their diameters are measured. Suppose we observe 20 diameters of ball bearings from the production lines, which are listed in Table 27.1. The average diameter is $\bar{x}_{20} = 1.03$ millimeter. This clearly deviates from the target value 1, but the question is whether the difference can be attributed to chance or whether it is large enough to conclude that the production line is producing ball bearings with a wrong diameter. To answer this question, we model the dataset as a realization of a random sample $X_1, X_2, \ldots, X_{20}$ from a probability distribution with expected value μ. The parameter μ represents the diameter of ball bearings produced by the produc-

Table 27.1. Diameters of ball bearings.

1.018	1.009	1.042	1.053	0.969	1.002	0.988	1.019	1.062	1.032
1.072	0.977	1.062	1.044	1.069	1.029	0.979	1.096	1.079	0.999

tion lines. In order to investigate whether this diameter deviates from 1, we test the null hypothesis $H_0 : \mu = 1$ against $H_1 : \mu \neq 1$.

This example illustrates a situation that often occurs: the data $x_1, x_2, \ldots, x_n$ are a realization of a random sample $X_1, X_2, \ldots, X_n$ from a distribution with expectation μ, and we want to test whether μ equals an a priori specified value, say μ_0. According to the law of large numbers, $\bar{X}_n$ is close to μ for large n. This suggests a test statistic based on $\bar{X}_n - \mu_0$; realizations of $\bar{X}_n - \mu_0$ close to zero are in favor of the null hypothesis. Does $\bar{X}_n - \mu_0$ suffice as a test statistic?

In our example, $\bar{x}_n - \mu_0 = 1.03 - 1 = 0.03$. Should we interpret this as small? First, note that under the null hypothesis $\mathrm{E}\left[\bar{X}_n - \mu_0\right] = \mu - \mu_0 = 0$. Now, if $\bar{X}_n - \mu_0$ would have standard deviation 1, then the value 0.03 is within one standard deviation of $\mathrm{E}\left[\bar{X}_n - \mu_0\right]$. The "$\mu \pm$ a few σ" rule on page 185 then suggests that the value 0.03 is not exceptional; it must be seen as a small deviation. On the other hand, if $\bar{X}_n - \mu_0$ has standard deviation 0.001, then the value 0.03 is 30 standard deviations away from $\mathrm{E}\left[\bar{X}_n - \mu_0\right]$. According to the "$\mu \pm$ a few σ" rule this is very exceptional; the value 0.03 must be seen as a large deviation. The next quick exercise provides a concrete example.

QUICK EXERCISE 27.1 Suppose that $\bar{X}_n$ is a normal random variable with expectation 1 and variance 1. Determine $\mathrm{P}\left(\bar{X}_n - 1 \geq 0.03\right)$. Find the same probability, but for the case where the variance is $(0.01)^2$.

This discussion illustrates that we must standardize $\bar{X}_n - \mu_0$ to incorporate its variation. Recall that

$$\mathrm{Var}\left(\bar{X}_n - \mu_0\right) = \mathrm{Var}\left(\bar{X}_n\right) = \frac{\sigma^2}{n},$$

where σ^2 is the variance of each X_i. Hence, standardizing $\bar{X}_n - \mu_0$ means that we should divide by $\sigma/\sqrt{n}$. Since σ is unknown, we substitute the sample standard deviation S_n for σ. This leads to the following test statistic for the null hypothesis $H_0 : \mu = \mu_0$:

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}.$$

Values of T close to zero are in favor of $H_0 : \mu = \mu_0$. Large positive values of T suggest that $\mu > \mu_0$ and large negative values suggest that $\mu < \mu_0$; both are evidence against H_0.

For the ball bearing data one finds that $s_n = 0.0372$, so that

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} = \frac{1.03 - 1}{0.0372/\sqrt{20}} = 3.607.$$

This is clearly different from zero, but the question is whether this difference is large enough to reject $H_0 : \mu = 1$. To answer this question, we need to know

the probability distribution of T under the null hypothesis. Note that under the null hypothesis $H_0 : \mu = \mu_0$, the test statistic

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

is the studentized mean (see also Chapter 23)

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}.$$

Hence, *under the null hypothesis*, the probability distribution of T is the *same* as that of the studentized mean.

27.2 The one-sample *t*-test

The classical assumption is that the dataset is a realization of a random sample from an $N(\mu, \sigma^2)$ distribution. In that case our test statistic T turns out to have a *t*-distribution under the null hypothesis, as we will see later. For this reason, the test for the null hypothesis $H_0 : \mu = \mu_0$ is called the *(one-sample) t-test*. Without the assumption of normality, we will use the bootstrap to approximate the distribution of T. For large sample sizes, this distribution can be approximated by means of the central limit theorem. We start with the first case.

Normal data

Suppose that the dataset $x_1, x_2, \ldots, x_n$ is a realization of a random sample $X_1, X_2, \ldots, X_n$ from an $N(\mu, \sigma^2)$ distribution. Then, according to the rule on page 349, the studentized mean has a $t(n-1)$ distribution. An immediate consequence is that, under the null hypothesis $H_0 : \mu = \mu_0$, also our test statistic T has a $t(n-1)$ distribution. Therefore, if we test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at level α, then we must reject the null hypothesis in favor of $H_1 : \mu \neq \mu_0$, if

$$T \leq -t_{n-1,\alpha/2} \quad \text{or} \quad T \geq t_{n-1,\alpha/2}.$$

Similar decision rules apply to alternatives $H_1 : \mu > \mu_0$ and $H_1 : \mu < \mu_0$. Suppose that in the ball bearing example we test $H_0 : \mu = 1$ against $H_1 : \mu \neq 1$ at level $\alpha = 0.05$. From Table B.2 we find $t_{19,0.025} = 2.093$. Hence, we must reject if $T \leq -2.093$ or $T \geq 2.093$. For the ball bearing data we found $t = 3.607$, which means we reject the null hypothesis at level $\alpha = 0.05$.

Alternatively, one might report the one-tailed p-value corresponding to the observed value t and compare this with $\alpha/2$. The one-tailed p-value is either a right or a left tail probability, which must be computed by means

of the $t(n-1)$ distribution. In our ball bearing example the one-tailed p-value is the right tail probability $P(T \geq 3.607)$. From Table B.2 we see that this probability is between 0.0005 and 0.0010, which is smaller than $\alpha/2 = 0.025$ (to be precise, by means of a statistical software package we found $P(T \geq 3.607) = 0.00094$). The data provide strong enough evidence against the null hypothesis, so that it seems sensible to adjust the settings of the production line.

QUICK EXERCISE 27.2 Suppose that the data in Table 27.1 are from two separate production lines. The first ten measurements have average 1.0194 and standard deviation 0.0290, whereas the last ten measurements have average 1.0406 and standard deviation 0.0428. Perform the t-test $H_0 : \mu = 1$ against $H_1 : \mu \neq 1$ at level $\alpha = 0.01$ for both datasets separately, assuming normality.

Nonnormal data

Draw a rectangle with height h and width w (let us agree that $w > h$), and within this rectangle draw a square with sides of length h (see Figure 27.1). This creates another (smaller) rectangle with horizontal and vertical sides of
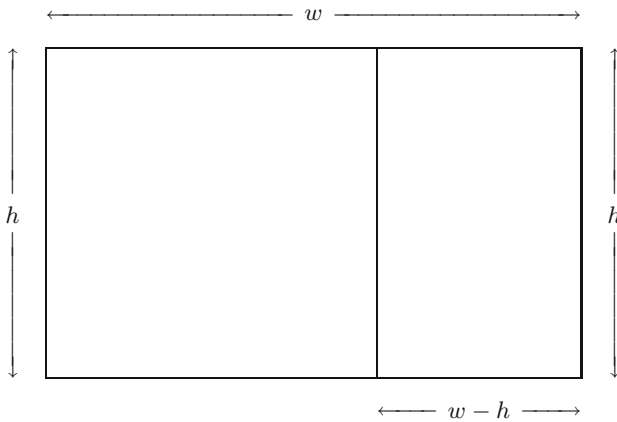
Fig. 27.1. Rectangle with square within.

lengths $w - h$ and h. A large rectangle with a vertical-to-horizontal ratio that is equal to the horizontal-to-vertical ratio for the small rectangle, i.e.,

$$\frac{h}{w} = \frac{w-h}{h},$$

was called a "golden rectangle" by the ancient Greeks, who often used these in their architecture. After solving for h/w, we obtain that the height-to-width

Table 27.2. Ratios for Shoshoni rectangles.

| 0.693 | 0.749 | 0.654 | 0.670 | 0.662 | 0.672 | 0.615 | 0.606 | 0.690 | 0.628 |
| 0.668 | 0.611 | 0.606 | 0.609 | 0.601 | 0.553 | 0.570 | 0.844 | 0.576 | 0.933 |

ratio h/w is equal to the "golden number" $(\sqrt{5}-1)/2 \approx 0.618$. The data in Table 27.2 represent corresponding h/w ratios for rectangles used by Shoshoni Indians to decorate their leather goods. Is it reasonable to assume that they were *also* using golden rectangles? We examine this by means of a t-test.

The observed ratios are modeled as a realization of a random sample from a distribution with expectation μ, where the parameter μ represents the true esthetic preference for height-to-width ratios of the Shoshoni Indians. We want to test

$$H_0 : \mu = 0.618 \quad \text{against} \quad H_1 : \mu \neq 0.618.$$

For the Shoshoni ratios, $\bar{x}_n = 0.6605$ and $s_n = 0.0925$, so that the value of the test statistic is

$$t = \frac{\bar{x}_n - 0.618}{s_n/\sqrt{n}} = \frac{0.6605 - 0.618}{0.0925/\sqrt{20}} = 2.055.$$

Closer examination of the data indicates that the normal distribution is not the right model. For instance, by definition the height-to-width ratios h/w are always between 0 and 1. Because some of the data points are also close to right boundary 1, the normal distribution is inappropriate. If we *cannot* assume a normal model distribution, we can *no longer* conclude that our test statistic has a $t(n-1)$ distribution under the null hypothesis.

Since there is no reason to assume any other particular type of distribution to model the data, we approximate the distribution of T under the null hypothesis. Recall that this distribution is the same as that of the studentized mean (see the end of Section 27.1). To approximate its distribution, we use the empirical bootstrap simulation for the studentized mean, as described on page 351. We generate $10\,000$ bootstrap datasets and for each bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$, we compute

$$t^* = \frac{\bar{x}_n^* - 0.6605}{s_n^*/\sqrt{n}}.$$

In Figure 27.2 the kernel density estimate and empirical distribution function are displayed for $10\,000$ bootstrap values t^*. Suppose we test $H_0 : \mu = 0.618$ against $H_1 : \mu \neq 0.618$ at level $\alpha = 0.05$. In the same way as in Section 23.3, we find the following bootstrap approximations for the critical values:

$$c_l^* = -3.334 \quad \text{and} \quad c_u^* = 1.644.$$
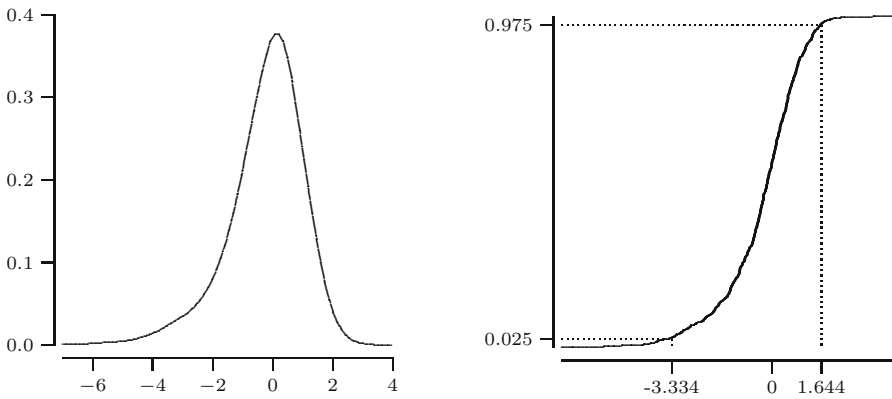
Fig. 27.2. Kernel density estimate and empirical distribution function of 10 000 bootstrap values t^*.

Since for the Shoshoni data the value 2.055 of the test statistic is greater than 1.644, we reject the null hypothesis at level 0.05. Alternatively, we can also compute a bootstrap approximation of the one-tailed p-value corresponding to 2.055, which is the right tail probability $P(T \geq 2.055)$. The bootstrap approximation for this probability is:

$$\frac{\text{number of } t^* \text{values greater than or equal to } 2.055}{10\,000} = 0.0067.$$

Hence $P(T \geq 2.055) \approx 0.0067$, which is smaller than $\alpha/2 = 0.025$. The value 2.055 should be considered as exceptionally large, and we reject the null hypothesis. The esthetic preference for height-to-width ratios of the Shoshoni Indians differs from that of the ancient Greeks.

Large samples

For large sample sizes the distribution of the studentized mean can be approximated by a standard normal distribution (see Section 23.4). This means that for large sample sizes the distribution of the t-test statistic under the null hypothesis can also be approximated by a standard normal distribution. To illustrate this, recall the Old Faithful data. Park rangers in Yellowstone National Park inform the public about the behavior of the geyser, such as the expected time between successive eruptions and the length of the duration of an eruption. Suppose they claim that the expected length of an eruption is 4 minutes (240 seconds). Does this seem likely on the basis of the data from Section 15.1? We investigate this by testing $H_0 : \mu = 240$ against $H_1 : \mu \neq 240$ at level $\alpha = 0.001$, where μ is the expectation of the model distribution. The value of the test statistic is

$$t = \frac{\bar{x}_n - 240}{s_n/\sqrt{n}} = \frac{209.3 - 240}{68.48/\sqrt{272}} = -7.39.$$

The one-tailed p-value $P(T \leq -7.39)$ can be approximated by $P(Z \leq -7.39)$, where Z has an $N(0,1)$ distribution. From Table B.1 we see that this probability is smaller than $P(Z \leq -3.49) = 0.0002$. This is smaller than $\alpha/2 = 0.0005$, so we reject the null hypothesis at level 0.001. In fact the p-value is much smaller: a statistical software package gives $P(Z \leq -7.39) = 7.5 \cdot 10^{-14}$. The data provide overwhelming evidence against $H_0 : \mu = 240$, so that we conclude that the expected length of an eruption is different from 4 minutes.

QUICK EXERCISE 27.3 Compute the critical region K for the test, using the normal approximation, and check that $t = -7.39$ falls in K.

In fact, if we would test $H_0 : \mu = 240$ against $H_1 : \mu < 240$, the p-value corresponding to $t = -7.39$ is the left tail probability $P(T \leq -7.39)$. This probability is very small, so that we also reject the null hypothesis in favor of this alternative and conclude that the expected length of an eruption is smaller than 4 minutes.

27.3 The *t*-test in a regression setting

Is calcium in your drinking water good for your health? In England and Wales, an investigation of environmental causes of disease was conducted. The annual mortality rate (percentage of deaths) and the calcium concentration in the drinking water supply were recorded for 61 large towns. The data in Table 27.3 represent the annual mortality rate averaged over the years 1958–1964, and the calcium concentration in parts per million. In Figure 27.3 the 61 paired measurements are displayed in a scatterplot. The scatterplot shows a slight downward trend, which suggests that higher concentrations of calcium lead to lower mortality rates. The question is whether this is really the case or if the slight downward trend should be attributed to chance.

To investigate this question we model the mortality data by means of a simple linear regression model with normally distributed errors, with the mortality rate as the dependent variable y and the calcium concentration as the independent variable x:

$$Y_i = \alpha + \beta x_i + U_i \quad \text{for } i = 1, 2, \dots, 61,$$

where $U_1, U_2, \dots, U_{61}$ is a random sample from an $N(0, \sigma^2)$ distribution. The parameter β represents the change of the mortality rate if we increase the calcium concentration by one unit. We test the null hypothesis $H_0 : \beta = 0$ (calcium has no effect on the mortality rate) against $H_1 : \beta < 0$ (higher concentration of calcium reduces the mortality rate).

This example illustrates the general situation, where the dataset

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

Table 27.3. Mortality data.

Rate	Calcium	Rate	Calcium	Rate	Calcium	Rate	Calcium
1247	105	1466	5	1299	78	1359	84
1392	73	1307	78	1254	96	1318	122
1260	21	1096	138	1402	37	1309	59
1259	133	1175	107	1486	5	1456	90
1236	101	1369	68	1257	50	1527	60
1627	53	1486	122	1485	81	1519	21
1581	14	1625	13	1668	17	1800	14
1609	18	1558	10	1807	15	1637	10
1755	12	1491	20	1555	39	1428	39
1723	44	1379	94	1742	8	1574	9
1569	91	1591	16	1772	15	1828	8
1704	26	1702	44	1427	27	1724	6
1696	6	1711	13	1444	14	1591	49
1987	8	1495	14	1587	75	1713	71
1557	13	1640	57	1709	71	1625	20
1378	71						
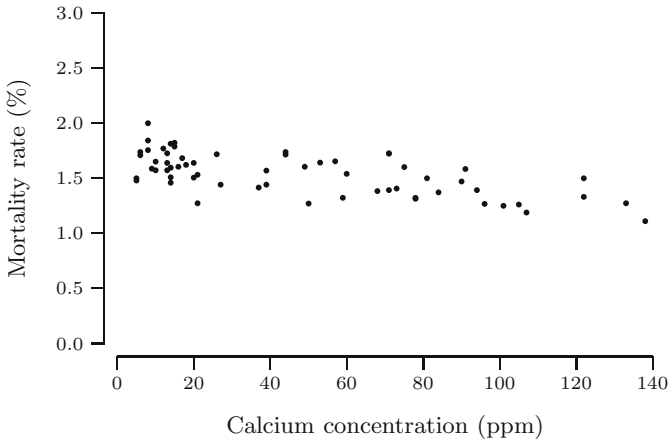
Fig. 27.3. Scatterplot mortality data.

is modeled by a simple linear regression model, and one wants to test a null hypothesis of the form $H_0 : \alpha = \alpha_0$ or $H_0 : \beta = \beta_0$. Similar to the one-sample *t*-test we will construct a test statistic for each of these null hypotheses. With normally distributed errors, these test statistics have a *t*-distribution under the null hypothesis. For this reason, for both null hypotheses the test is called a *t*-test.

The t-test for the slope

For the null hypothesis $H_0 : \beta = \beta_0$, we use as test statistic

$$T_b = \frac{\hat{\beta} - \beta_0}{S_b},$$

where $\hat{\beta}$ is the least squares estimator for β (see Chapter 22) and

$$S_b^2 = \frac{n}{n \sum x_i^2 - (\sum x_i)^2} \hat{\sigma}^2.$$

In this expression,

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

is the estimator for σ^2 as introduced on page 332. It can be shown that

$$\mathrm{Var}\left(\hat{\beta} - \beta_0\right) = \frac{n}{n \sum x_i^2 - (\sum x_i)^2} \sigma^2,$$

so that the random variable S_b^2 is an estimator for the variance of $\hat{\beta} - \beta_0$. Hence, similar to the test statistic for the one-sample t-test, the test statistic T_b compares the estimator $\hat{\beta}$ with the value β_0 and standardizes by dividing by an estimator for the standard deviation of $\hat{\beta} - \beta_0$. Values of T_b close to zero are in favor of the null hypothesis $H_0 : \beta = \beta_0$. Large positive values of T_b suggest that $\beta > \beta_0$, whereas large negative values of T_b suggest that $\beta < \beta_0$.

Recall that in the case of normal random samples the one-sample t-test statistic has a $t(n-1)$ distribution under the null hypothesis. For the same reason, it is also a fact that in the case of normally distributed errors the test statistic T_b has a $t(n-2)$ distribution under the null hypothesis $H_0 : \beta = \beta_0$.

In our mortality example we want to test $H_0 : \beta = 0$ against $H_0 : \beta < 0$. For the data we find $\hat{\beta} = -3.2261$ and $s_b = 0.4847$, so that the value of T_b is

$$t_b = \frac{-3.2261}{0.4847} = -6.656.$$

If we test at level $\alpha = 0.05$, then we must compare this value with the left critical value $-t_{59,0.05}$. This value is not in Table B.2, but we have that

$$-1.676 = -t_{50,0.05} < -t_{59,0.05}.$$

This means that t_b is much smaller than $-t_{59,0.05}$, so that we reject the null hypothesis at level 0.05. How much evidence the value $t_b = -6.656$ bears against the null hypothesis is expressed by the one-tailed p-value $P(T_b \leq -6.656)$. From Table B.2 we can only see that this probability is smaller than 0.0005. By means of a statistical package we find $P(T_b \leq -6.656) = 5.2 \cdot 10^{-9}$. The data provide overwhelming evidence against the null hypothesis. We conclude that higher concentrations of calcium correspond to lower mortality rates.

QUICK EXERCISE 27.4 The data in Table 27.3 can be separated into measurements for towns at least as far north as Derby and towns south of Derby. For the data corresponding to 35 towns at least as far north as Derby, one finds $\hat{\beta} = -1.9313$ and $s_b = 0.8479$. Test $H_0 : \beta = 0$ against $H_0 : \beta < 0$ at level 0.01, i.e., compute the value of the test statistic and report your conclusion about the null hypothesis.

The *t*-test for the intercept

We test the null hypothesis $H_0 : \alpha = \alpha_0$ with test statistic

$$T_a = \frac{\hat{\alpha} - \alpha_0}{S_a}, \qquad (27.1)$$

where $\hat{\alpha}$ is the least squares estimator for α and

$$S_a^2 = \frac{\sum x_i^2}{n \sum x_i^2 - (\sum x_i)^2} \hat{\sigma}^2,$$

with $\hat{\sigma}^2$ defined as before. The random variable S_a^2 is an estimator for the variance

$$\mathrm{Var}(\hat{\alpha} - \alpha_0) = \frac{\sum x_i^2}{n \sum x_i^2 - (\sum x_i)^2} \sigma^2.$$

Again, we compare the estimator $\hat{\alpha}$ with the value α_0 and standardize by dividing by an estimator for the standard deviation of $\hat{\alpha} - \alpha_0$. Values of T_a close to zero are in favor of the null hypothesis $H_0 : \alpha = \alpha_0$. Large positive values of T_a suggest that $\alpha > \alpha_0$, whereas large negative values of T_a suggest that $\alpha < \alpha_0$. Like T_b, in the case of normal errors, the test statistic T_a has a $t(n-2)$ distribution under the null hypothesis $H_0 : \alpha = \alpha_0$.

As an illustration, recall Exercise 17.9 where we modeled the volume y of black cherry trees by means of a linear model without intercept, with independent variable $x = d^2h$, where d and h are the diameter and height of the trees. The scatterplot of the pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_{31}, y_{31})$ is displayed in Figure 27.4. As mentioned in Exercise 17.9, there are physical reasons to leave out the intercept. We want to investigate whether this is confirmed by the data. To this end, we model the data by a simple linear regression model *with* intercept

$$Y_i = \alpha + \beta x_i + U_i \quad \text{for } i = 1, 2, \ldots, 31,$$

where $U_1, U_2, \ldots, U_{31}$ are a random sample from an $N(0, \sigma^2)$ distribution, and we test $H_0 : \alpha = 0$ against $H_1 : \alpha \neq 0$ at level 0.10. The value of the test statistic is

$$t_a = \frac{-0.2977}{0.9636} = -0.3089,$$

and the left critical value is $-t_{29,0.05} = -1.699$. This means we cannot reject the null hypothesis. The data do not provide sufficient evidence against $H_0 : \alpha = 0$, which is confirmed by the one-tailed p-value $P(T_a \leq -0.3089) = 0.3798$ (computed by means of a statistical package). We conclude that the intercept does not contribute significantly to the model.
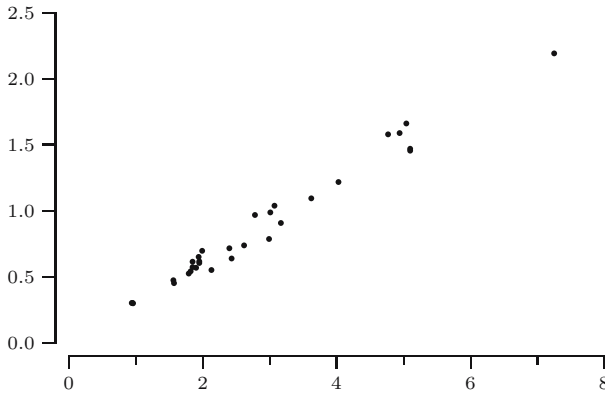
Fig. 27.4. Scatterplot of the black cherry tree data.

27.4 Solutions to the quick exercises

27.1 If Y has an $N(1,1)$ distribution, then $Y - 1$ has an $N(0,1)$ distribution. Therefore, from Table B.1: $P(Y - 1 \geq 0.03) = 0.4880$. If Y has an $N(1,(0.01)^2)$ distribution, then $(Y - 1)/0.01$ has an $N(0,1)$ distribution. In that case,

$$P(Y - 1 \geq 0.03) = P\left(\frac{Y - 1}{0.01} \geq 3\right) = 0.0013.$$

27.2 For the first and last ten measurements the values of the test statistic are

$$t = \frac{1.0194 - 1}{0.0290/\sqrt{10}} = 2.115 \quad \text{and} \quad t = \frac{1.0406 - 1}{0.0428/\sqrt{10}} = 3.000.$$

The critical value $t_{9,0.025} = 2.262$, which means we reject the null hypothesis for the second production line, but not for the first production line.

27.3 The critical region is of the form $K = (-\infty, c_l] \cup [c_u, \infty)$. The right critical value c_u is approximated by $z_{0.0005} = t_{\infty,0.0005} = 3.291$, which can be found in Table B.2. By symmetry of the normal distribution, the left critical value c_l is approximated by $-z_{0.0005} = -3.291$. Clearly, $t = -7.39 < -3.291$, so that it falls in K.

27.4 The value of the test statistic is

$$t_b = \frac{-1.9313}{0.8479} = -2.2778.$$

The left critical value is equal to $-t_{33,0.01}$, which is not in Table B.2, but we see that $-t_{33,0.01} < -t_{40,0.01} = -2.423$. This means that $-t_{33,0.01} < t_b$, so that we cannot reject $H_0 : \beta = 0$ against $H_0 : \beta < 0$ at level 0.01.

27.5 Exercises

27.1 We perform a *t*-test for the null hypothesis $H_0 : \mu = 10$ by means of a dataset consisting of $n = 16$ elements with sample mean 11 and sample variance 4. We use significance level 0.05.

a. Should we reject the null hypothesis in favor of $H_1 : \mu \neq 10$?

b. What if we test against $H_1 : \mu > 10$?

27.2 □ The Cleveland Casting Plant is a large highly automated producer of gray and nodular iron automotive castings for Ford Motor Company. One process variable of interest to Cleveland Casting is the pouring temperature of molten iron. The pouring temperatures (in degrees Fahrenheit) of ten crankshafts are given in Table 27.4. The target setting for the pouring temperature is set at 2550 degrees. One wants to conduct a test at level $\alpha = 0.01$ to determine whether the pouring temperature differs from the target setting.

Table 27.4. Pouring temperatures of ten crankshafts.

2543	2541	2544	2620	2560
2559	2562	2553	2552	2553

© 1995 From A structural model relating process inputs and final product characteristics, *Quality Engineering*, , Vol 7, No. 4, pp. 693-704, by Price, B. and Barth, B. Reproduced by permission of Taylor & Francis, Inc., http//www.taylorandfrancis.com

a. Formulate the appropriate null hypothesis and alternative hypothesis.

b. Compute the value of the test statistic and report your conclusion. You may assume a normal model distribution and use that the sample variance is 517.34.

27.3 Table 27.5 lists the results of tensile adhesion tests on 22 U-700 alloy specimens. The data are loads at failure in MPa. The sample mean is 13.71 and the sample standard deviation is 3.55. You may assume that the data originated from a normal distribution with expectation μ. One is interested in whether the load at failure exceeds 10 MPa. We investigate this by means of a *t*-test for the null hypothesis $H_0 : \mu = 10$.

a. What do you choose as the alternative hypothesis?

b. Compute the value of the test statistic and report your conclusion, when performing the test at level 0.05.

Table 27.5. Loads at failure of U-700 specimens.

19.8	18.5	17.6	16.7	15.8
15.4	14.1	13.6	11.9	11.4
11.4	8.8	7.5	15.4	15.4
19.5	14.9	12.7	11.9	11.4
10.1	7.9			

Source: C.C. Berndt. Instrumented Tensile adhesion tests on plasma sprayed thermal barrier coatings. *Journal of Materials Engineering* II(4): 275-282, Dec 1989. © Springer-Verlag New York Inc.

27.4 Consider the coal data from Table 23.2, where 22 gross calorific value measurements are listed for Daw Mill coal coded 258GB41. We modeled this dataset as a realization of a random sample from an $N(\mu, \sigma^2)$ distribution with μ and σ unknown. We are planning to buy a shipment if the gross calorific value exceeds 31.00 MJ/kg. The sample mean and sample variance of the data are $\bar{x}_n = 31.012$ and $s_n = 0.1294$. Perform a t-test for the null hypothesis $H_0 : \mu = 31.00$ against $H_1 : \mu > 31.00$ using significance level 0.01, i.e., compute the value of the test statistic, the critical value of the test, and report your conclusion.

27.5 ⊞ In the November 1988 issue of *Science* a study was reported on the inbreeding of tropical swarm-founding wasps. Each member of a sample of 197 wasps was captured, frozen, and subjected to a series of genetic tests, from which an inbreeding coefficient was determined. The sample mean and the sample standard deviation of the coefficients are $\bar{x}_{197} = 0.044$ and $s_{197} = 0.884$. If a species does not have the tendency to inbreed, their true inbreeding coefficient is 0. Determine by means of a test whether the inbreeding coefficient for this species of wasp exceeds 0.

a. Formulate the appropriate null hypothesis and alternative hypothesis and compute the value of the test statistic.

b. Compute the p-value corresponding to the value of the test statistic and report your conclusion about the null hypothesis.

27.6 The stopping distance of an automobile is related to its speed. The data in Table 27.6 give the stopping distance in feet and speed in miles per hour of an automobile. The data are modeled by means of simple linear regression model with normally distributed errors, with the square root of the stopping distance as dependent variable y and the speed as independent variable x:

$$Y_i = \alpha + \beta x_i + U_i, \quad \text{for } i = 1, \ldots, 7.$$

For the dataset we find

$$\hat{\alpha} = 5.388, \quad \hat{\beta} = 4.252, \quad s_a = 1.874, \quad s_b = 0.242.$$

Table 27.6. Speed and stopping distance of automobiles.

Speed	20.5	20.5	30.5	30.5	40.5	48.8	57.8
Distance	15.4	13.3	33.9	27.0	73.1	113.0	142.6

Source: K.A. Brownlee. *Statistical theory and methodology in science and engineering.* Wiley, New York, 1960; Table II.9 on page 372.

One would expect that the intercept can be taken equal to 0, since zero speed would yield zero stopping distance. Investigate whether this is confirmed by the data by performing the appropriate test at level 0.10. Formulate the proper null and alternative hypothesis, compute the value of the test statistic, and report your conclusion.

27.7 ⊞ In a study about the effect of wall insulation, the weekly gas consumption (in 1000 cubic feet) and the average outside temperature (in degrees Celsius) was measured of a certain house in southeast England, for 26 weeks before and 30 weeks after cavity-wall insulation had been installed. The house thermostat was set at 20 degrees throughout. The data are listed in Table 27.7. We model the data before insulation by means of a simple linear regression model with normally distributed errors and gas consumption as response variable. A similar model was used for the data after insulation. Given are

Before insulation: $\hat{\alpha} = 6.8538$, $\hat{\beta} = -0.3932$ and $s_a = 0.1184$, $s_b = 0.0196$
After insulation: $\hat{\alpha} = 4.7238$, $\hat{\beta} = -0.2779$ and $s_a = 0.1297$, $s_b = 0.0252$.

a. Use the data before insulation to investigate whether smaller outside temperatures lead to higher gas consumption. Formulate the proper null and alternative hypothesis, compute the value of the test statistic, and report your conclusion, using significance level 0.05.

b. Do the same for the data after insulation.

Table 27.7. Temperature and gas consumption.

Before insulation		After insulation	
Temperature	Gas consumption	Temperature	Gas consumption
−0.8	7.2	−0.7	4.8
−0.7	6.9	0.8	4.6
0.4	6.4	1.0	4.7
2.5	6.0	1.4	4.0
2.9	5.8	1.5	4.2
3.2	5.8	1.6	4.2
3.6	5.6	2.3	4.1
3.9	4.7	2.5	4.0
4.2	5.8	2.5	3.5
4.3	5.2	3.1	3.2
5.4	4.9	3.9	3.9
6.0	4.9	4.0	3.5
6.0	4.3	4.0	3.7
6.0	4.4	4.2	3.5
6.2	4.5	4.3	3.5
6.3	4.6	4.6	3.7
6.9	3.7	4.7	3.5
7.0	3.9	4.9	3.4
7.4	4.2	4.9	3.7
7.5	4.0	4.9	4.0
7.5	3.9	5.0	3.6
7.6	3.5	5.3	3.7
8.0	4.0	6.2	2.8
8.5	3.6	7.1	3.0
9.1	3.1	7.2	2.8
10.2	2.6	7.5	2.6
		8.0	2.7
		8.7	2.8
		8.8	1.3
		9.7	1.5

Comparing two samples

Many applications are concerned with *two* groups of observations of the same kind that originate from two possibly different model distributions, and the question is whether these distributions have different expectations. We describe a test for equality of expectations, where we consider normal and nonnormal model distributions and equal and unequal variances of the model distributions.

28.1 Is dry drilling faster than wet drilling?

Recall the drilling example from Sections 15.5 and 16.4. The question was whether dry drilling is faster than wet drilling. The scatterplots in Figure 15.11 seem to suggest that up to a depth of 250 feet the drill time does not depend on depth. Therefore, for a first investigation of a possible difference between dry and wet drilling we only consider the (mean) drill times up to this depth. A more thorough study can be found in [23].

The boxplots of the drill times for both types of drilling are displayed in Figure 28.1. Clearly, the boxplot for dry drilling is positioned lower than the
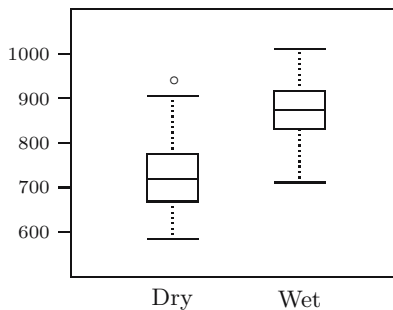
Fig. 28.1. Boxplot of drill times.

one for wet drilling. However, the question is whether this difference can be attributed to chance or if it is large enough to conclude that the dry drill time is shorter than the wet drill time. To answer this question, we model the datasets of dry and wet drill times as realizations of random samples from two distribution functions F and G, one with expected value μ_1 and the other with expected value μ_2. The parameters μ_1 and μ_2 represent the drill times of dry drilling and wet drilling, respectively. We test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$.

This example illustrates a general situation where we compare two datasets

$$x_1, x_2, \ldots, x_n \quad \text{and} \quad y_1, y_2, \ldots, y_m,$$

which are the realization of independent random samples

$$X_1, X_2, \ldots, X_n \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m$$

from two distributions, and we want to test whether the expectations of both distributions are the same. Both the variance σ_X^2 of the X_i and the variance σ_Y^2 of the Y_j are *unknown*.

Note that the null hypothesis is equivalent to the statement $\mu_1 - \mu_2 = 0$. For this reason, similar to Chapter 27, the test statistic for the null hypothesis $H_0 : \mu_1 = \mu_2$ is based on an estimator $\bar{X}_n - \bar{Y}_m$ for the difference $\mu_1 - \mu_2$. As before, we standardize $\bar{X}_n - \bar{Y}_m$ by an estimator for its variance

$$\operatorname{Var}\left(\bar{X}_n - \bar{Y}_m\right) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}.$$

Recall that the sample variances S_X^2 and S_Y^2 of the X_i and Y_j, are unbiased estimators for σ_X^2 and σ_Y^2. We will use a combination of S_X^2 and S_Y^2 to construct an estimator for $\operatorname{Var}\left(\bar{X}_n - \bar{Y}_m\right)$. The actual standardization of $\bar{X}_n - \bar{Y}_m$ depends on whether the variances of the X_i and Y_j are the same. We distinguish between the two cases $\sigma_X^2 = \sigma_Y^2$ and $\sigma_X^2 \neq \sigma_Y^2$. In the next section we consider the case of equal variances.

QUICK EXERCISE 28.1 Looking at the boxplots in Figure 28.1, does the assumption $\sigma_X^2 = \sigma_Y^2$ seem reasonable to you? Can you think of a way to quantify your belief?

28.2 Two samples with equal variances

Suppose that the samples originate from distributions with the *same* (but unknown) variance:

$$\sigma_X^2 = \sigma_Y^2 = \sigma^2.$$

In this case we can *pool* the sample variances S_X^2 and S_Y^2 by constructing a linear combination $aS_X^2 + bS_Y^2$ that is an unbiased estimator for σ^2. One particular choice is the weighted average

$$\frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

It has the property that for normally distributed samples it has the smallest variance among all unbiased linear combinations of S_X^2 and S_Y^2 (see Exercise 28.5). Moreover, the weights depend on the sample sizes. This is appropriate, since if one sample is much larger than the other, the estimate of σ^2 from that sample is more reliable and should receive greater weight.

We find that the *pooled-variance*:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)$$

is an unbiased estimator for

$$\mathrm{Var}\left(\bar{X}_n - \bar{Y}_m\right) = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right).$$

This leads to the following test statistic for the null hypothesis $H_0 : \mu_1 = \mu_2$:

$$T_p = \frac{\bar{X}_n - \bar{Y}_m}{S_p}.$$

As before, we compare the estimator $\bar{X}_n - \bar{Y}_m$ with 0 (the value of $\mu_1 - \mu_2$ under the null hypothesis), and we standardize by dividing by the estimator S_p for the standard deviation of $\bar{X}_n - \bar{Y}_m$. Values of T_p close to zero are in favor of the null hypothesis $H_0 : \mu_1 = \mu_2$. Large positive values of T_p suggest that $\mu_1 > \mu_2$, whereas large negative values suggest that $\mu_1 < \mu_2$.

The next step is to determine the distribution of T_p. Note that under the null hypothesis $H_0 : \mu_1 = \mu_2$, the test statistic T_p is the *pooled studentized mean difference*

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_1 - \mu_2)}{S_p}.$$

Hence, *under the null hypothesis*, the probability distribution of T_p is the *same* as that of the pooled studentized mean difference. To determine its distribution, we distinguish between normal and nonnormal data.

Normal samples

In the same way as the studentized mean of a single normal sample has a $t(n-1)$ distribution (see page 349), it is also a fact that if two independent samples originate from normal distributions, i.e.,

$$X_1, X_2, \ldots, X_n \text{ random sample from } N(\mu_1, \sigma^2)$$
$$Y_1, Y_2, \ldots, Y_m \text{ random sample from } N(\mu_2, \sigma^2),$$

then the pooled studentized mean difference has a $t(n+m-2)$ distribution. Hence, under the null hypothesis, the test statistic T_p has a $t(n+m-2)$

distribution. For this reason, a test for the null hypothesis $H_0 : \mu_1 = \mu_2$ is called a *two-sample t-test*.

Suppose that in our drilling example we model our datasets as realizations of random samples of sizes $n = m = 50$ from two normal distributions with equal variances, and we test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ at level 0.05. For the data we find $\bar{x}_{50} = 727.78$, $\bar{y}_{50} = 873.02$, and $s_p = 13.62$, so that

$$t_p = \frac{727.78 - 873.02}{13.62} = -10.66.$$

We compare this with the left critical value $-t_{98,0.05}$. This value is not in Table B.2, but $-1.676 = -t_{50,0.05} < -t_{98,0.05}$. This means that $t_p < -t_{98,0.05}$, so that we reject $H_0 : \mu_1 = \mu_2$ in favor of $H_1 : \mu_1 < \mu_2$ at level 0.05. The p-value corresponding to $t_p = -10.66$ is the left tail probability $P(T \leq -10.66)$. From Table B.2 we can only see that this is smaller than 0.0005 (a statistical software package gives $P(T \leq -10.66) = 2.25 \cdot 10^{-18}$). The data provide overwhelming evidence against the null hypothesis, so that we conclude that dry drilling is faster than wet drilling.

QUICK EXERCISE 28.2 Suppose that in the ball bearing example of Quick exercise 27.2, we test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, where μ_1 and μ_2 represent the diameters of a ball bearing from the first and second production line. What are the critical values corresponding to level $\alpha = 0.01$?

Nonnormal samples

Similar to the one-sample t-test, if we *cannot* assume normal model distributions, then we can *no longer* conclude that our test statistic has a $t(n + m - 2)$ distribution under the null hypothesis. Recall that under the null hypothesis, the distribution of our test statistic is the same as that of the pooled studentized mean difference (see page 417).

To approximate its distribution, we use the empirical bootstrap simulation for the pooled studentized mean difference

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_1 - \mu_2)}{S_p}.$$

Given datasets $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$, determine their empirical distribution functions F_n and G_m as estimates for F and G. The expectations corresponding to F_n and G_m are $\mu_1^* = \bar{x}_n$ and $\mu_2^* = \bar{y}_m$. Then repeat the following two steps many times:

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n and a bootstrap dataset $y_1^*, y_2^*, \ldots, y_m^*$ from G_m.
2. Compute the pooled studentized mean difference for the bootstrap data:

$$t_p^* = \frac{(\bar{x}_n^* - \bar{y}_m^*) - (\bar{x}_n - \bar{y}_m)}{s_p^*},$$

where $\bar{x}_n^*$ and $\bar{y}_m^*$ are the sample means of the bootstrap datasets, and

$$(s_p^*)^2 = \frac{(n-1)(s_X^*)^2 + (m-1)(s_Y^*)^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)$$

with $(s_X^*)^2$ and $(s_Y^*)^2$ the sample variances of the bootstrap datasets.

The reason that in each iteration we subtract $\bar{x}_n - \bar{y}_m$ is that $\mu_1 - \mu_2$ is the difference of the expectations of the two model distributions. Therefore, according to the bootstrap principle we should replace this by the difference $\bar{x}_n - \bar{y}_m$ of the expectations corresponding to the two empirical distribution functions.

We carried out this bootstrap simulation for the drill times. The result of this simulation can be seen in Figure 28.2, where a histogram and the empirical distribution function are displayed for one thousand bootstrap values of t_p^*. Suppose that we test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ at level 0.05. The bootstrap approximation for the left critical value is $c_l^* = -1.659$. The value of $t_p = -10.66$, computed from the data, is much smaller. Hence, also on the basis of the bootstrap simulation we reject the null hypothesis and conclude that the dry drill time is shorter than the wet drill time.
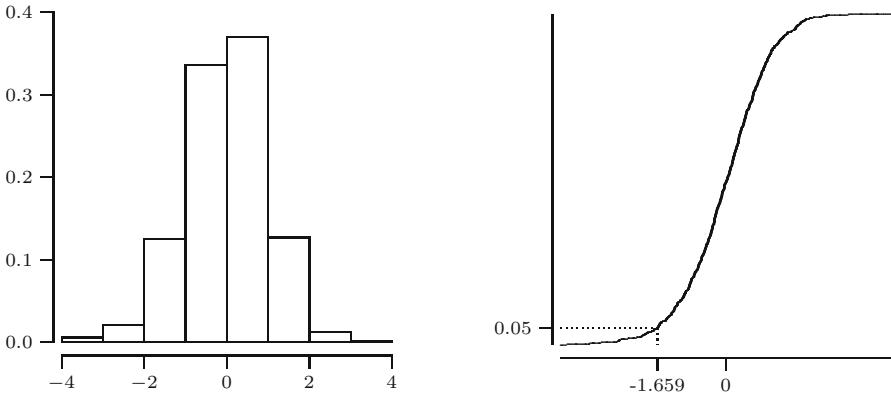
Fig. 28.2. Histogram and empirical distribution function of 1000 bootstrap values for T_p^*.

28.3 Two samples with unequal variances

During an investigation about weather modification, a series of experiments was conducted in southern Florida from 1968 to 1972. These experiments were designed to investigate the use of massive silver-iodide seeding. It was

Table 28.1. Rainfall data.

Unseeded					
1202.6	830.1	372.4	345.5	321.2	244.3
163.0	147.8	95.0	87.0	81.2	68.5
47.3	41.1	36.6	29.0	28.6	26.3
26.1	24.4	21.7	17.3	11.5	4.9
4.9	1.0				

Seeded					
2745.6	1697.8	1656.0	978.0	703.4	489.1
430.0	334.1	302.8	274.7	274.7	255.0
242.5	200.7	198.6	129.6	119.0	118.3
115.3	92.4	40.6	32.7	31.4	17.5
7.7	4.1				

hypothesized that under specified conditions, this leads to invigorated cumulus growth and prolonged lifetimes, thereby causing increased precipitation. In these experiments, 52 isolated cumulus clouds were observed, of which 26 were selected at random and injected with silver-iodide smoke. Rainfall amounts (in acre-feet) were recorded for all clouds. They are listed in Table 28.1. To investigate whether seeding leads to increased rainfall, we test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$, where μ_1 and μ_2 represent the rainfall for unseeded and seeded clouds.

In Figure 28.3 the boxplots of both datasets are displayed. From this we see that the assumption of equal variances may not be realistic. Indeed, this is confirmed by the values $s_X^2 = 77\,521$ and $s_Y^2 = 423\,524$ of the sample variances of the datasets. This means that we need to test $H_0 : \mu_1 = \mu_2$ without the assumption of equal variances. As before, the test statistic will be a standardized version of $\bar{X}_n - \bar{Y}_m$, but S_p^2 is no longer an unbiased estimator for

$$\mathrm{Var}\left(\bar{X}_n - \bar{Y}_m\right) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}.$$

However, if we estimate σ_X^2 and σ_Y^2 by S_X^2 and S_Y^2, then the *nonpooled variance*

$$S_d^2 = \frac{S_X^2}{n} + \frac{S_Y^2}{m}$$

is an unbiased estimator for $\mathrm{Var}\left(\bar{X}_n - \bar{Y}_m\right)$. This leads to test statistic

$$T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d}.$$
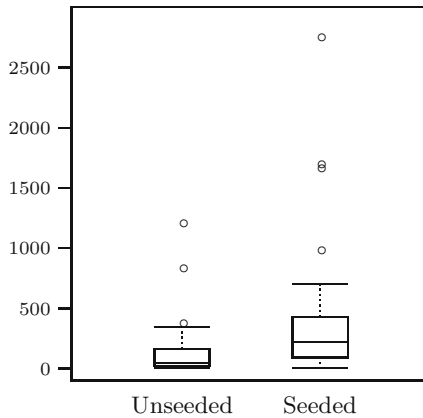
Fig. 28.3. Boxplots of rainfall.

Again, we compare the estimator $\bar{X}_n - \bar{Y}_m$ with zero and standardize by dividing by an estimator for the standard deviation of $\bar{X}_n - \bar{Y}_m$. Values of T_d close to zero are in favor of the null hypothesis $H_0 : \mu_1 = \mu_2$.

QUICK EXERCISE 28.3 Consider the ball bearing example from Quick exercise 27.2. Compute the value of T_d for this example.

Under the null hypothesis $H_0 : \mu_1 = \mu_2$, the test statistic

$$T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d}$$

is equal to the *nonpooled studentized mean difference*

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_1 - \mu_2)}{S_d}.$$

Therefore, the distribution of T_d under the null hypothesis is the same as that of the nonpooled studentized mean difference. Unfortunately, its distribution is not a t-distribution, not even in the case of normal samples. This means that we have to approximate this distribution.

Similar to the previous section, we use the empirical bootstrap simulation for the nonpooled studentized mean difference. The only difference with the procedure outlined in the previous section is that now in each iteration we compute the nonpooled studentized mean difference for the bootstrap datasets:

$$t_d^* = \frac{(\bar{x}_n^* - \bar{y}_m^*) - (\bar{x}_n - \bar{y}_m)}{s_d^*},$$

where $\bar{x}_n^*$ and $\bar{y}_m^*$ are the sample means of the bootstrap datasets, and

$$(s_d^*)^2 = \frac{(s_X^*)^2}{n} + \frac{(s_Y^*)^2}{m}$$
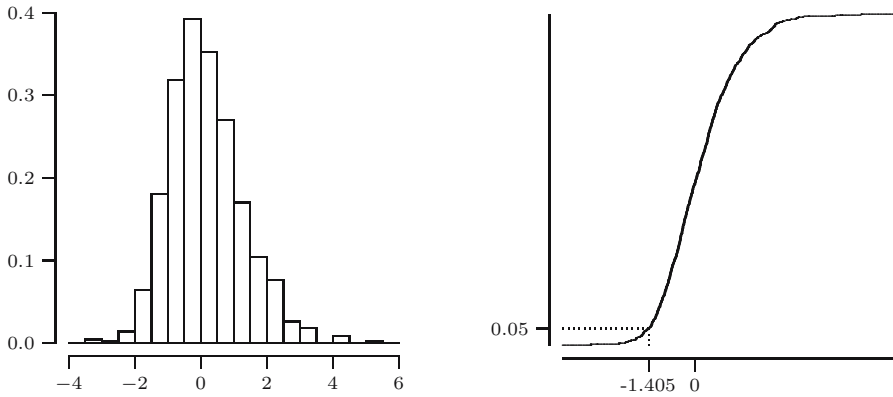
Fig. 28.4. Histogram and empirical distribution function of 1000 bootstrap values of T_d^*.

with $(s_X^*)^2$ and $(s_Y^*)^2$ the sample variances of the bootstrap datasets.

We carried out this bootstrap simulation for the cloud seeding data. The result of this simulation can be seen in Figure 28.4, where a histogram and the empirical distribution function are displayed for one thousand values t_d^*. The bootstrap approximation for the left critical value corresponding to level 0.05 is $c_l^* = -1.405$. For the data we find the value

$$t_d = \frac{164.59 - 441.98}{138.92} = -1.998.$$

This is smaller than c_l^*, so we reject the null hypothesis. Although the evidence against the null hypothesis is not overwhelming, there is some indication that seeding clouds leads to more rainfall.

28.4 Large samples

Variants of the central limit theorem state that as n and m both tend to infinity, the distributions of the pooled studentized mean difference

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_1 - \mu_2)}{S_p}$$

and the nonpooled studentized mean difference

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_1 - \mu_2)}{S_d}$$

both approach the standard normal distribution. This fact can be used to approximate the distribution of the test statistics T_p and T_d under the null hypothesis by a standard normal distribution.

We illustrate this by means of the following example. To investigate whether a restricted diet promotes longevity, two groups of randomly selected rats were put on the different diets. One group of $n = 106$ rats was put on a restricted diet, the other group of $m = 89$ rats on an ad libitum diet (i.e., unrestricted eating). The data in Table 28.2 represent the remaining lifetime in days of two groups of rats after they were put on the different diets. The average lifetimes are $\bar{x}_n = 968.75$ and $\bar{y}_m = 684.01$ days. To investigate whether a restricted diet promotes longevity, we test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$, where μ_1 and μ_2 represent the lifetime of a rat on a restricted diet and on an ad libitum diet, respectively.

If we may assume equal variances, we compute

$$t_p = \frac{968.75 - 684.01}{32.88} = 8.66.$$

This value is larger than the right critical value $z_{0.0005} = 3.291$, which means that we would reject $H_0 : \mu_1 = \mu_2$ in favor of $H_1 : \mu_1 > \mu_2$ at level $\alpha = 0.0005$.

Table 28.2. Rat data.

					Restricted				
105	193	211	236	302	363	389	390	391	403
530	604	605	630	716	718	727	731	749	769
770	789	804	810	811	833	868	871	875	893
897	901	906	907	919	923	931	940	957	958
961	962	974	979	982	1001	1008	1010	1011	1012
1014	1017	1032	1039	1045	1046	1047	1057	1063	1070
1073	1076	1085	1090	1094	1099	1107	1119	1120	1128
1129	1131	1133	1136	1138	1144	1149	1160	1166	1170
1173	1181	1183	1188	1190	1203	1206	1209	1218	1220
1221	1228	1230	1231	1233	1239	1244	1258	1268	1294
1316	1327	1328	1369	1393	1435				

					Ad libitum				
89	104	387	465	479	494	496	514	532	536
545	547	548	582	606	609	619	620	621	630
635	639	648	652	653	654	660	665	667	668
670	675	677	678	678	681	684	688	694	695
697	698	702	704	710	711	712	715	716	717
720	721	730	731	732	733	735	736	738	739
741	743	746	749	751	753	764	765	768	770
773	777	779	780	788	791	794	796	799	801
806	807	815	836	838	850	859	894	963	

Source: B.L. Berger, D.D. Boos, and F.M. Guess. Tests and confidence sets for comparing two mean residual life functions. *Biometrics*, 44:103–115, 1988.

The p-value is the right tail probability $P(T_p \geq 8.66)$, which we approximate by $P(Z \geq 8.66)$, where Z has an $N(0,1)$ distribution. From Table B.1 we see that this probability is smaller than $P(Z \geq 3.49) = 0.0002$. By means of a statistical package we find $P(Z \geq 8.66) = 2.4 \cdot 10^{-16}$.

If we repeat the test without the assumption of equal variances, we compute

$$t_d = \frac{968.75 - 684.01}{31.08} = 9.16,$$

which also leads to rejection of the null hypothesis. In this case, the p-value $P(T_d \geq 9.16) \approx P(Z \geq 9.16)$ is even smaller since $9.16 > 8.66$ (a statistical package gives $P(Z \geq 9.16) = 2.6 \cdot 10^{-18}$). The data provide overwhelming evidence against the null hypothesis, and we conclude that a restricted diet promotes longevity.

28.5 Solutions to the quick exercises

28.1 Just by looking at the boxplots, the authors believe that the assumption $\sigma_X^2 = \sigma_Y^2$ is reasonable. The lengths of the boxplots and their IQRs are almost the same. However, the boxplots do not reveal how the elements of the dataset vary around the center. One way of quantifying our belief would be to compare the sample variances of the datasets. One possibility is to compare the ratio of both sample variances; a ratio close to one would support our belief of equal variances (in case of normal samples, this is a standard test called the F-test).

28.2 In this case we have a right and left critical value. From Quick exercise 27.2 we know that $n = m = 10$, so that the right critical value is $t_{18,0.005} = 2.878$ and the left critical value is $-t_{18,0.005} = -2.878$.

28.3 We first compute $s_d^2 = (0.0290)^2/10 + (0.0428)^2/10 = 0.000267$ and then $t_d = (1.0194 - 1.0406)/\sqrt{0.000267} = -1.297$.

28.6 Exercises

28.1 ☐ The data in Table 28.3 represent salaries (in pounds Sterling) in 72 randomly selected advertisements in the *The Guardian* (April 6, 1992). When a range was given in the advertisement, the midpoint of the range is reproduced in the table. The data are salaries corresponding to two kinds of occupations ($n = m = 72$): (1) creative, media, and marketing and (2) education. The sample mean and sample variance of the two datasets are, respectively:

$$\begin{aligned}
&(1) \quad \bar{x}_{72} = 17\,410 \text{ and } s_x^2 = 41\,258\,741, \\
&(2) \quad \bar{y}_{72} = 19\,818 \text{ and } s_y^2 = 50\,744\,521.
\end{aligned}$$

Table 28.3. Salaries in two kinds of occupations.

Occupation (1)			Occupation (2)		
17703	13796	12000	25899	17378	19236
42000	22958	22900	21676	15594	18780
18780	10750	13440	15053	17375	12459
15723	13552	17574	19461	20111	22700
13179	21000	22149	22485	16799	35750
37500	18245	17547	17378	12587	20539
22955	19358	9500	15053	24102	13115
13000	22000	25000	10998	12755	13605
13500	12000	15723	18360	35000	20539
13000	16820	12300	22533	20500	16629
11000	17709	10750	23008	13000	27500
12500	23065	11000	24260	18066	17378
13000	18693	19000	25899	35403	15053
10500	14472	13500	18021	17378	20594
12285	12000	32000	17970	14855	9866
13000	20000	17783	21074	21074	21074
16000	18900	16600	15053	19401	25598
15000	14481	18000	20739	15053	15053
13944	35000	11406	15053	15083	31530
23960	18000	23000	30800	10294	16799
11389	30000	15379	37000	11389	15053
12587	12548	21458	48000	11389	14359
17000	17048	21262	16000	26544	15344
9000	13349	20000	20147	14274	31000

Suppose that the datasets are modeled as realizations of normal distributions with expectations μ_1 and μ_2, which represent the salaries for occupations (1) and (2).

a. Test the null hypothesis that the salary for both occupations is the same at level $\alpha = 0.05$ under the assumption of equal variances. Formulate the proper null and alternative hypotheses, compute the value of the test statistic, and report your conclusion.

b. Do the same without the assumption of equal variances.

c. As a comparison, one carries out an empirical bootstrap simulation for the nonpooled studentized mean difference. The bootstrap approximations for the critical values are $c_l^* = -2.004$ and $c_u^* = 2.133$. Report your conclusion about the salaries on the basis of the bootstrap results.

28.2 The data in Table 28.4 represent the duration of pregnancy for 1669 women who gave birth in a maternity hospital in Newcastle-upon-Tyne, England, in 1954.

Table 28.4. Durations of pregnancy.

Duration	Medical	Emergency	Social
11		1	
15		1	
17	1		
20		1	
22	1	2	
24	1	3	
25		2	1
26		1	
27	2	2	1
28	1	2	1
29	3	1	
30	3	5	1
31	4	5	2
32	10	9	2
33	6	6	2
34	12	7	10
35	23	11	4
36	26	13	19
37	54	16	30
38	68	35	72
39	159	38	115
40	197	32	155
41	111	27	128
42	55	25	64
43	29	8	16
44	4	5	3
45	3	1	6
46	1	1	1
47	1		
56		1	

Source: D.J. Newell. Statistical aspects of the demand for maternity beds. *Journal of the Royal Statistical Society, Series A*, 127:1–33, 1964.

The durations are measured in complete weeks from the beginning of the last menstrual period until delivery. The pregnancies are divided into those where an admission was booked for medical reasons, those booked for social reasons (such as poor housing), and unbooked emergency admissions. For the three groups the sample means and sample variances are

Medical: 775 observations with $\bar{x} = 39.08$ and $s^2 = 7.77$,
Emergency: 261 observations with $\bar{x} = 37.59$ and $s^2 = 25.33$,
Social: 633 observations with $\bar{x} = 39.60$ and $s^2 = 4.95$.

Suppose we view the datasets as realizations of random samples from normal distributions with expectations μ_1, μ_2, and μ_3 and variances σ_1^2, σ_2^2, and σ_3^2, where μ_i represents the duration of pregnancy for the women from the ith group. We want to investigate whether the duration differs for the different groups. For each combination of two groups test the null hypothesis of equality of μ_i. Compute the values of the test statistic and report your conclusions.

28.3 ⊡ In a seven-day study on the effect of ozone, a group of 23 rats was kept in an ozone-free environment and a group of 22 rats in an ozone-rich environment. From each member in both groups the increase in weight (in grams) was recorded. The results are given in Table 28.5. The interest is in whether ozone affects the increase of weight. We investigate this by testing $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, where μ_1 and μ_2 denote the increases of weight for a rat in the ozone-free and ozone-rich groups. The sample means are

$$\text{Ozone-free:}\quad \bar{x}_{23} = 22.40$$
$$\text{Ozone-rich:}\quad \bar{y}_{22} = 11.01.$$

The pooled standard deviation is $s_p = 4.58$, and the nonpooled standard deviation is $s_d = 4.64$.

Table 28.5. Weight increase of rats.

Ozone-free			Ozone-rich		
41.0	38.4	24.4	10.1	6.1	20.4
25.9	21.9	18.3	7.3	14.3	15.5
13.1	27.3	28.5	−9.9	6.8	28.2
−16.9	17.4	21.8	17.9	−12.9	14.0
15.4	27.4	19.2	6.6	12.1	15.7
22.4	17.7	26.0	39.9	−15.9	54.6
29.4	21.4	22.7	−14.7	44.1	−9.0
26.0	26.6		−9.0		

a. Perform the test at level 0.05 under the assumption of normal data with equal variances, i.e., compute the test statistic and report your conclusion.

b. One also carries out a bootstrap simulation for the test statistic used in **a**, and finds critical values $c_l^* = -1.912$ and $c_u^* = 1.959$. What is your conclusion on the basis of the bootstrap simulation?

c. Also perform the test at level 0.05 without the assumption of equal variances, where you may use the normal approximation for the distribution of the test statistic under the null hypothesis.

d. A bootstrap simulation for the test statistic in **c** yields that the right tail probability corresponding to the observed value of the test statistic in this case is 0.014. What is your conclusion on the basis of the bootstrap simulation?

28.4 Show that in the case when $n = m$, the random variables T_p and T_d are the same.

28.5 ⊞ Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent random samples from normal distributions with variances σ^2. It can be shown that

$$\mathrm{Var}\big(S_X^2\big) = \frac{2\sigma^4}{n-1} \quad \text{and} \quad \mathrm{Var}\big(S_Y^2\big) = \frac{2\sigma^4}{m-1}.$$

Consider linear combinations $aS_X^2 + bS_Y^2$ that are unbiased estimators for σ^2.

a. Show that a and b must satisfy $a + b = 1$.

b. Show that $\mathrm{Var}\big(aS_X^2 + (1-a)S_Y^2\big)$ is minimized for $a = (n-1)/(n+m-2)$ (and hence $b = (m-1)/(n+m-2)$).

28.6 Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be independent random samples from distributions with (possibly unequal) variances σ_X^2 and σ_Y^2.

a. Show that

$$\mathrm{Var}\big(\bar{X}_n - \bar{Y}_m\big) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}.$$

b. Show that the pooled variance S_p^2, as defined on page 417, is a biased estimator for $\mathrm{Var}\big(\bar{X}_n - \bar{Y}_m\big)$.

c. Show that the nonpooled variance S_d^2, as defined on page 420, is the only unbiased estimator for $\mathrm{Var}\big(\bar{X}_n - \bar{Y}_m\big)$ of the form $aS_X^2 + bS_Y^2$.

d. Suppose that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. Show that S_d^2, as defined on page 417, is an unbiased estimator for $\mathrm{Var}\big(\bar{X}_n - \bar{Y}_m\big) = \sigma^2(1/n + 1/m)$.

e. Is S_d^2 also an unbiased estimator for $\mathrm{Var}\big(\bar{X}_n - \bar{Y}_m\big)$ in the case $\sigma_X^2 \neq \sigma_Y^2$? What about when $n = m$?

A

Summary of distributions

Discrete distributions

1. **Bernoulli distribution**: $Ber(p)$, where $0 \leq p \leq 1$.
   $P(X = 1) = p$ and $P(X = 0) = 1 - p$.
   $E[X] = p$ and $Var(X) = p(1 - p)$.

2. **Binomial distribution**: $Bin(n, p)$, where $0 \leq p \leq 1$.
   $$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n.$$
   $E[X] = np$ and $Var(X) = np(1 - p)$.

3. **Geometric distribution**: $Geo(p)$, where $0 < p \leq 1$.
   $P(X = k) = p(1 - p)^{k-1}$ for $k = 1, 2, \ldots$.
   $E[X] = 1/p$ and $Var(X) = (1 - p)/p^2$.

4. **Poisson distribution**: $Pois(\mu)$, where $\mu > 0$.
   $$P(X = k) = \frac{\mu^k}{k!} \, e^{-\mu} \quad \text{for } k = 0, 1, \ldots.$$
   $E[X] = \mu$ and $Var(X) = \mu$.

Continuous distributions

1. **Cauchy distribution**: $Cau(\alpha, \beta)$, where $-\infty < \alpha < \infty$ and $\beta > 0$.
   $$f(x) = \frac{\beta}{\pi \left(\beta^2 + (x - \alpha)^2 \right)} \quad \text{for } -\infty < x < \infty.$$
   $$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x - \alpha}{\beta} \right) \quad \text{for } -\infty < x < \infty.$$
   $E[X]$ and $Var(X)$ do not exist.

2. **Exponential distribution**: $Exp(\lambda)$, where $\lambda > 0$.

   $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

   $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$.

   $\mathrm{E}[X] = 1/\lambda$ and $\mathrm{Var}(X) = 1/\lambda^2$.

3. **Gamma distribution**: $Gam(\alpha, \lambda)$, where $\alpha > 0$ and $\lambda > 0$.

   $$f(x) = \frac{\lambda (\lambda x)^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \quad \text{for } x \geq 0.$$

   $$F(x) = \int_0^x \frac{\lambda (\lambda t)^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)} \, dt \quad \text{for } x \geq 0.$$

   $\mathrm{E}[X] = \alpha/\lambda$ and $\mathrm{Var}(X) = \alpha/\lambda^2$.

4. **Normal distribution**: $N(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma > 0$.

   $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < \infty.$$

   $$F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} \, dt \quad \text{for } -\infty < x < \infty.$$

   $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

5. **Pareto distribution**: $Par(\alpha)$, where $\alpha > 0$.

   $$f(x) = \frac{\alpha}{x^{\alpha+1}} \quad \text{for } x \geq 1.$$

   $F(x) = 1 - x^{-\alpha}$ for $x \geq 1$.

   $\mathrm{E}[X] = \alpha/(\alpha - 1)$ for $\alpha > 1$ and ∞ for $0 < \alpha \leq 1$.

   $\mathrm{Var}(X) = \alpha/((\alpha - 1)^2(\alpha - 2))$ for $\alpha > 2$ and ∞ for $0 < \alpha \leq 1$.

6. **Uniform distribution**: $U(a, b)$, where $a < b$.

   $$f(x) = \frac{1}{b - a} \quad \text{for } a \leq x \leq b.$$

   $$F(x) = \frac{x - a}{b - a} \quad \text{for } a \leq x \leq b.$$

   $\mathrm{E}[X] = (a + b)/2$ and $\mathrm{Var}(X) = (b - a)^2/12$.

B

Tables of the normal and *t*-distributions

Table B.1. Right tail probabilities $1 - \Phi(a) = P(Z \geq a)$ for an $N(0,1)$ distributed random variable Z.

a	0	1	2	3	4	5	6	7	8	9
0.0	5000	4960	4920	4880	4840	4801	4761	4721	4681	4641
0.1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
0.2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
0.3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
0.4	3446	3409	3372	3336	3300	3264	3228	3192	3156	3121
0.5	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
0.6	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
0.7	2420	2389	2358	2327	2296	2266	2236	2206	2177	2148
0.8	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
0.9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
1.0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
1.1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
1.2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
1.3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
1.4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
1.5	0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
1.6	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
1.7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
1.8	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
1.9	0287	0281	0274	0268	0262	0256	0250	0244	0239	0233
2.0	0228	0222	0217	0212	0207	0202	0197	0192	0188	0183
2.1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
2.2	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
2.3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
2.4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064
2.5	0062	0060	0059	0057	0055	0054	0052	0051	0049	0048
2.6	0047	0045	0044	0043	0041	0040	0039	0038	0037	0036
2.7	0035	0034	0033	0032	0031	0030	0029	0028	0027	0026
2.8	0026	0025	0024	0023	0023	0022	0021	0021	0020	0019
2.9	0019	0018	0018	0017	0016	0016	0015	0015	0014	0014
3.0	0013	0013	0013	0012	0012	0011	0011	0011	0010	0010
3.1	0010	0009	0009	0009	0008	0008	0008	0008	0007	0007
3.2	0007	0007	0006	0006	0006	0006	0006	0005	0005	0005
3.3	0005	0005	0005	0004	0004	0004	0004	0004	0004	0003
3.4	0003	0003	0003	0003	0003	0003	0003	0003	0003	0002

Table B.2. Right critical values $t_{m,p}$ of the t-distribution with m degrees of freedom corresponding to right tail probability p: $\mathrm{P}(T_m \geq t_{m,p}) = p$. The last row in the table contains right critical values of the $N(0,1)$ distribution: $t_{\infty,p} = z_p$.

	Right tail probability p							
m	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
∞	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

C

Answers to selected exercises

2.1 $P(A \cup B) = 13/18$.

2.4 Yes.

2.7 0.7.

2.8 $P(D_1 \cup D_2) \leq 2 \cdot 10^{-6}$ and $P(D_1 \cap D_2) \leq 10^{-6}$.

2.11 $p = (-1 + \sqrt{5})/2$.

2.12 a $1/10!$

2.12 b $5! \cdot 5!$

2.12 c $8/63 = 12.7$ percent.

2.14 a

	a	b	c
a	0	1/6	1/6
b	0	0	1/3
c	0	1/3	0

2.14 b $P(\{(a, b), (a, c)\}) = 1/3$.

2.14 c $P(\{(b, c), (c, b)\}) = 2/3$.

2.16 $P(E) = 2/3$.

2.19 a $\Omega = \{2, 3, 4, \dots\}$.

2.19 b $4p^2(1 - p)^3$.

3.1 $7/36$.

3.2 a $P(A \mid B) = 2/11$.

3.2 b No.

3.3 a $P(S_1) = 13/52 = 1/4$, $P(S_2 \mid S_1) = 12/51$, and $P(S_2 \mid S_1^c) = 13/51$.

3.3 b $P(S_2) = 1/4$.

3.4 $P(B \mid T) = 9.1 \cdot 10^{-5}$ and $P(B \mid T^c) = 4.3 \cdot 10^{-6}$.

3.7 a $P(A \cup B) = 1/2$.

3.7 b $P(B) = 1/3$.

3.8 a $P(W) = 0.117$.

3.8 b $P(F \mid W) = 0.846$.

3.9 $P(B \mid A) = 7/15$.

3.14 a $P(W \mid R) = 0$ and $P(W \mid R^c) = 1$.

3.14 b $P(W) = 2/3$.

3.16 a $P(D \mid T) = 0.165$.

3.16 b 0.795.

4.1 a

a	0	1	2
$p_Z(a)$	25/36	10/36	1/36

Z has a $Bin(2, 1/6)$ distribution.

4.1 b $\{M = 2, Z = 0\} = \{(2, 1), (1, 2), (2, 2)\}$, $\{S = 5, Z = 1\} = \emptyset$, and $\{S = 8, Z = 1\} = \{(6, 2), (2, 6)\}$. $P(M = 2, Z = 0) = 1/12$, $P(S = 5, Z = 1) = 0$, and $P(S = 8, Z = 1) = 1/18$.

4.1 c The events are dependent.

4.3

a	0	1/2	3/4
$p(a)$	1/3	1/6	1/2

4.6 a $p_{\bar{X}}(1) = p_{\bar{X}}(3) = 1/27$, $p_{\bar{X}}(4/3) = p_{\bar{X}}(8/3) = 3/27$, $p_{\bar{X}}(5/3) = p_{\bar{X}}(7/3) = 6/27$, and $p_{\bar{X}}(2) = 7/27$.

4.6 b $6/27$.

4.7 a $Bin(1000, 0.001)$.

4.7 b $P(X = 0) = 0.3677$, $P(X = 1) = 0.3681$, and $P(X > 2) = 0.0802$.

4.8 a $Bin(6, 0.8178)$.

4.8 b 0.9999634.

4.10 a Determine $P(R_i = 0)$ first.

4.10 b No!

4.10 c See the birthday problem in Section 3.2.

4.12 No!

4.13 a $Geo(1/N)$.

4.13 b Let D_i be the event that the marked bolt was drawn (for the first time) in the ith draw, and use conditional probabilities in
$P(Y = k) = P(D_1^c \cap \cdots \cap D_{k-1}^c \cap D_k)$.

4.13 c Count the number of ways the event $\{Z = k\}$ can occur, and divide this by the number of ways $\binom{N}{r}$ we can select r objects from N objects.

5.2 $P(1/2 < X \le 3/4) = 5/16$.

5.4 a $P(X < 4\,1/2) = 1/4$.

5.4 b $P(X = 5) = 1/2$.

5.4 c X is neither discrete nor continuous!

5.5 a $c = 1$.

5.5 b $F(x) = 0$ for $x \le -3$;
$F(x) = (x + 3)^2/2$ for $-3 \le x \le -2$;
$F(x) = 1/2$ for $-2 \le x \le 2$;
$F(x) = 1 - (3 - x)^2/2$ for $2 \le x \le 3$;
$F(x) = 1$ for $x \ge 3$.

5.8 a $g(y) = 1/(2\sqrt{ry})$.

5.8 b Yes.

5.8 c Consider $F(r/10)$.

5.9 a $1/2$ and $\{(x, y) : 2 \le x \le 3, 1 \le y \le 3/2\}$.

5.9 b $F(x) = 0$ for $x < 0$;
$F(x) = 2x$ for $0 \le x \le 1/2$;
$F(x) = 1$ for $x > 1/2$.

5.9 c $f(x) = 2$ for $0 \le x \le 1/2$;
$f(x) = 0$ elsewhere.

5.12 2.

5.13 a Change variables from x to $-x$.

5.13 b $P(Z \le -2) = 0.0228$.

6.2 a $1 + 2\sqrt{0.378 \cdots} = 2.2300795$.

6.2 b Smaller.

6.2 c 0.3782739.

6.5 Show, for $a \ge 0$, that $X \le a$ is equivalent with $U \ge e^{-a}$.

6.6 $U = e^{-2X}$.

6.7 $Z = \sqrt{-\ln(1 - U)/5}$, or
$Z = \sqrt{-\ln U/5}$.

6.9 a $6/8$.

6.9 b $Geo(6/8)$.

6.10 a Define $B_i = 1$ if $U_i \le p$ and $B_i = 0$ if $U_i > p$, and N as the position in the sequence of B_i where the first 1 occurs.

6.10 b $P(Z > n) = (1 - p)^n$, for $n = 0, 1, \ldots$; Z has a $Geo(p)$ distribution.

7.1 a Outcomes: $1, 2, 3, 4, 5$, and 6. Each has probability $1/6$.

7.1 b $E[T] = 7/2$, $Var(T) = 35/12$.

7.2 a $E[X] = 1/5$.

7.2 b

y	0	1
$P(Y = y)$	2/5	3/5

and $E[Y] = 3/5$.

7.2 c $E[X^2] = 3/5$.

7.2 d $Var(X) = 14/25$.

7.5 $E[X] = p$ and $Var(X) = p(1 - p)$.

7.6 $195/76$.

7.8 $E[X] = 1/3$.

7.10 a $E[X] = 1/\lambda$ and $E[X^2] = 2/\lambda^2$.

7.10 b $Var(X) = 1/\lambda^2$.

7.11 a 2.

7.11 b The expectation is infinite!

7.11 c $E[X] = \int_1^\infty x \cdot \alpha x^{-\alpha-1} \, dx$.

7.15 a Start with
$Var(rX) = E\left[(rX - E[rX])^2\right]$.

7.15 b Start with $Var(X + s) = E\left[((X + s) - E[X + s])^2\right]$.

7.15 c Apply **b** with rX instead of X.

7.16 $E[X] = 4/9$.

7.17 a If positive terms add to zero, they must all be zero.

7.17 b Note that
$E\left[(V - E[V])^2\right] = \text{Var}(V).$

8.1

y	0	10	20
$P(Y = y)$	0.2	0.4	0.4

8.2 a

y	-1	0	1
$P(Y = y)$	1/6	1/2	1/3

8.2 b

z	-1	0	1
$P(Z = z)$	1/3	1/2	1/6

8.2 c $P(W = 1) = 1$.

8.3 a V has a $U(7, 9)$ distribution.

8.3 b $rU + s$ has a $U(s, s + r)$ distribution if $r > 0$ and a $U(s+r, s)$ distribution if $r < 0$.

8.5 a $x^2(3 - x)/4$ for $0 \le x \le 2$.

8.5 b $F_Y(y) = (3/4)y^4 - (1/4)y^6$ for $0 \le y \le \sqrt{2}$.

8.5 c $3y^3 - (3/2)y^5$ for $0 \le y \le \sqrt{2}$, 0 elsewhere.

8.8 $F_W(w) = 1 - e^{-\gamma w^\alpha}$, with $\gamma = \lambda^\alpha$.

8.10 0.1587.

8.11 Apply Jensen with $-g$.

8.12 a

y	0	1	10	100
$P(Y = y)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

8.12 b $\sqrt{E[X]} \ge E\left[\sqrt{X}\right]$.

8.12 c $\sqrt{E[X]} = 50.25$, but $E\left[\sqrt{X}\right] = 27.75$.

8.18 V has an exponential distribution with parameter $n\lambda$.

8.19 a The upper right quarter of the circle.

8.19 b $F_Z(t) = 1/2 + \arctan(t)/\pi$.

8.19 c $1/[\pi(1 + z^2)]$.

9.2 a $P(X = 0, Y = -1) = 1/6$,
$P(X = 0, Y = 1) = 0$,
$P(X = 1, Y = -1) = 1/6$,
$P(X = 2, Y = -1) = 1/6$,
and $P(X = 2, Y = 1) = 0$.

9.2 b Dependent.

9.5 a $1/16 \le \eta \le 1/4$.

9.5 b No.

9.6 a

		u		
v	0	1	2	
0	1/4	0	1/4	1/2
1	0	1/2	0	1/2
	1/4	1/2	1/4	1

9.6 b Dependent.

9.8 a

z	0	1	2	3
$p_Z(z)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

9.8 b

z	-2	-1	0	1	2	3
$p_{\tilde{X}}(z)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

9.9 a $F_X(x) = 1 - e^{-2x}$ for $x > 0$ and $F_Y(y) = 1 - e^{-y}$ for $y > 0$.

9.9 b $f(x, y) = 2e^{-(2x+y)}$ for $x > 0$ and $y > 0$.

9.9 c $f_X(x) = 2e^{-2x}$ $x > 0$ and $f_Y(y) = e^{-y}$ for $y > 0$.

9.9 d Independent.

9.10 a $41/720$.

9.10 b $F(a, b) = \frac{3}{5}a^2b^2 + \frac{2}{5}a^2b^3$.

9.10 c $F_X(a) = a^2$.

9.10 d $f_X(x) = 2x$ for $0 \le x \le 1$.

9.10 e Independent.

9.11 $27/50$.

9.13 a $1/\pi$.

9.13 b $F_R(r) = r^2$ for $0 \le r \le 1$.

9.13 c $f_X(x) = \frac{2}{\pi}\sqrt{1 - x^2} = f_Y(x)$ for x between -1 and 1.

9.15 a Since $F(a, b) = \frac{\text{area}\,(\Delta \cap \square(a,b))}{\text{area of }\Delta}$, where $\square(a, b)$ is the set of points (x, y), for which $x \le a$ and $y \le b$, one needs to calculate the areas for the various cases.

9.15 b $f(x, y) = 2$ for $(x, y) \in \Delta$, and $f(x, y) = 0$ otherwise.

9.15 c Use the rule on page 122.

9.19 a $a = 5\sqrt{2}$, $b = 4\sqrt{2}$, and $c = 18$.

9.19 b Use that $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$ is the probability density function of an $N(\mu,\sigma^2)$ distributed random variable.

9.19 c $N(0,1/36)$.

10.1 a $\text{Cov}(X,Y)=0.142$. Positively correlated.

10.1 b $\rho(X,Y)=0.0503$.

10.2 a $\text{E}[XY]=0$.

10.2 b $\text{Cov}(X,Y)=0$.

10.2 c $\text{Var}(X+Y)=4/3$.

10.2 d $\text{Var}(X-Y)=4/3$.

10.5 a

	a			
b	0	1	2	
0	8/72	6/72	10/72	1/3
1	12/72	9/72	15/72	1/2
2	4/72	3/72	5/72	1/6
	1/3	1/4	5/12	1

10.5 b $\text{E}[X]=13/12$, $\text{E}[Y]=5/6$, and $\text{Cov}(X,Y)=0$.

10.5 c Yes.

10.6 a $\text{E}[X]=\text{E}[Y]=0$ and $\text{Cov}(X,Y)=0$.

10.6 b $\text{E}[X]=\text{E}[Y]=c$; $\text{E}[XY]=c^2$.

10.6 c No.

10.7 a $\text{Cov}(X,Y)=-1/8$.

10.7 b $\rho(X,Y)=-1/2$.

10.7 c For ε equal to 1/4, 0 or $-1/4$.

10.9 a $\text{P}(X_i=1)=(1-0.001)^{40}=0.96$ and $\text{P}(X_i=41)=0.04$.

10.9 b $\text{E}[X_i]=2.6$ and $\text{E}[X_1+\cdots+X_{25}]=65$.

10.10 a $\text{E}[X]=109/50$, $\text{E}[Y]=157/100$, and $\text{E}[X+Y]=15/4$.

10.10 b $\text{E}[X^2]=1287/250$, $\text{E}[Y^2]=318/125$, and $\text{E}[X+Y]=3633/250$.

10.10 c $\text{Var}(X)=989/2500$, $\text{Var}(Y)=791/10\,000$, and $\text{Var}(X+Y)=4747/10\,000$.

10.14 a Use the alternative expression for the covariance.

10.14 b Use the alternative expression for the covariance.

10.14 c Combine parts **a** and **b**.

10.16 a $\text{Var}(X)+\text{Cov}(X,Y)$.

10.16 b Anything can happen.

10.16 c X and $X+Y$ are positively correlated.

10.18 Solve $0=N(N-1)(N+1)/12+N(N-1)\text{Cov}(X_1,X_2)$.

11.1 a Check that for k between 2 and 6, the summation runs over $\ell=1,\ldots,k-1$, whereas for k between 7 and 12 it runs over $\ell=k-6,\ldots,12$.

11.1 b Check that for $2\le k\le N$, the summation runs over $\ell=1,\ldots,k-1$, whereas for k between $N+1$ and $2N$ it runs over $\ell=k-N,\ldots,2N$.

11.2 a Check that the summation runs over $\ell=0,1,\ldots,k$.

11.2 b Use that $\lambda^{k-\ell}\mu^\ell/(\lambda+\mu)^k$ is equal to $p^\ell\left(1-p\right)^{k-\ell}$, with $p=\mu/(\lambda+\mu)$.

11.4 a $\text{E}[Z]=-3$ and $\text{Var}(Z)=81$.

11.4 b Z has an $N(-3,81)$ distribution.

11.4 c $\text{P}(Z\le 6)=0.8413$.

11.5 Check that for $0\le z<1$, the integral runs over $0\le y\le z$, whereas for $1\le z\le 2$, it runs over $z-1\le y\le 1$.

11.6 Check that the integral runs over $0\le y\le z$.

11.7 Recall that a $Gam(k,\lambda)$ random variable can be represented as the sum of k independent $Exp(\lambda)$ random variables.

11.9 a $f_Z(z)=\dfrac{3}{2}\left(\dfrac{1}{z^2}-\dfrac{1}{z^4}\right)$, for $z\ge 1$.

11.9 b $f_Z(z)=\dfrac{\alpha\beta}{\beta-\alpha}\left(\dfrac{1}{z^{\beta+1}}-\dfrac{1}{z^{\alpha+1}}\right)$, for $z\ge 1$.

12.1 e 1: no, 2: no, 3: okay, 4: okay, 5: okay.

12.5 a 0.00049.

12.5 b 1 (correct to 8281 decimals).

12.6 0.256.

12.7 a $\lambda \approx 0.192$.

12.7 b 0.1583 is close to 0.147.

12.7 c $2.71 \cdot 10^{-5}$.

12.8 a $\mathrm{E}[X(X-1)] = \mu^2$.

12.8 b $\mathrm{Var}(X) = \mu$.

12.11 The probability of the event in the hint equals $(\lambda s)^n e^{-\lambda 2s}/(k!(n-k)!)$.

12.14 a Note: $1-1/n \to 1$ and $1/n \to 0$.

12.14 b $\mathrm{E}[X_n] = (1-1/n)\cdot 0 + (1/n)\cdot 7n = 7$.

13.2 a $\mathrm{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = 1/12$.

13.2 b $1/12$.

13.4 a $n \geq 63$.

13.4 b $n \geq 250$.

13.4 c $n \geq 125$.

13.4 d $n \geq 240$.

13.6 Expected income per game € $1/37$; per year: € 9865.

13.8 a $\mathrm{Var}\big(\bar{Y}_n/2h\big) = 0.171/h\sqrt{n}$.

13.8 b $n \geq 801$.

13.9 a T_n is the average of a sequence of independent and identically distributed random variables.

13.9 b $a = \mathrm{E}\big[X_i^2\big] = 1/3$.

13.10 a $\mathrm{P}(|M_n - 1| > \varepsilon) = (1-\varepsilon)^n$ for $0 \leq \varepsilon \leq 1$.

13.10 b No.

14.2 0.9977.

14.3 17.

14.4 $1/2$.

14.5 Use that X has the same probability distribution as $X_1 + X_2 + \cdots + X_n$, where $X_1, X_2, \ldots, X_n$ are independent $Ber(p)$ distributed random variables.

14.6 a $\mathrm{P}(X \leq 25) \approx 0.5$, $\mathrm{P}(X < 26) \approx 0.6141$.

14.6 b $\mathrm{P}(X \leq 2) \approx 0$.

14.9 a 5.71%.

14.9 b Yes!

14.10 a 91.

14.10 b Use that $(\bar{M}_n - c)/\sigma$ has an $N(0,1)$ distribution.

15.3 a

Bin	Height
(0,250]	0.00297
(250,500]	0.00067
(500,750]	0.00015
(750,1000]	0.00008
(1000,1250]	0.00002
(1250,1500]	0.00004
(1500,1750]	0.00004
(1750,2000]	0
(2250,2500]	0
(2250,2500]	0.00002

15.3 b Skewed.
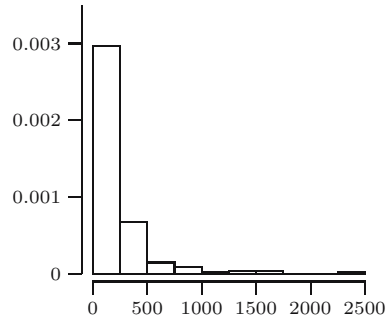
15.4 a

Bin	Height
[0,500]	0.0012741
(500,1000]	0.0003556
(1000,1500]	0.0001778
(1500,2000]	0.0000741
(2000,2500]	0.0000148
(2500,3000]	0.0000148
(3000,3500]	0.0000296
(3500,4000]	0
(4000,4500]	0.0000148
(4500,5000]	0
(5000,5500]	0.0000148
(5500,6000]	0.0000148
(6000,6500]	0.0000148

15.4 b

t	$F_n(t)$	t	$F_n(t)$
0	0	3500	0.9704
500	0.6370	4000	0.9704
1000	0.8148	4500	0.9778
1500	0.9037	5000	0.9778
2000	0.9407	5500	0.9852
2500	0.9481	6000	0.9926
3000	0.9556	6500	1

15.4 c Both are equal to 0.0889.

15.5

Bin	Height
$(0, 1]$	0.2250
$(1, 3]$	0.1100
$(3, 5]$	0.0850
$(5, 8]$	0.0400
$(8, 11]$	0.0230
$(11, 14]$	0.0350
$(14, 18]$	0.0225

15.6 $F_n(7) = 0.9$.

15.11 Use that the number of x_i in $(a, b]$ equals the number of $x_i \leq b$ minus the number of $x_i \leq a$.

15.12 a Bring the integral into the sum, change the integration variable to $u = (t - x_i)/h$, and use the properties of kernel functions.

15.12 b Similar to **a**.

16.1 a Median: 290.

16.1 b Lower quartile: 81; upper quartile: 843; IQR: 762.

16.1 c 144.6.

16.3 a Median: 70; lower quartile: 66.25; upper quartile: 75.

16.3 b
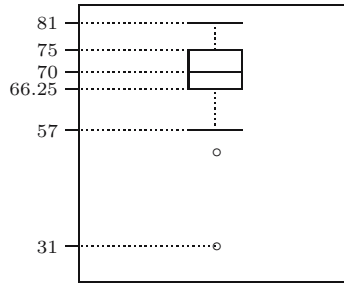
16.3 c Note the position of 31 in the boxplot.

16.4 a Yes, they both equal 7.056.

16.4 b Yes.

16.4 c Yes.

16.6 a Yes.

16.6 b In general this will not be true.

16.6 c Yes.

16.8 MAD is 3.

16.10 a The sample mean goes to infinity, whereas the sample median changes to 4.6.

16.10 b At least three elements need to be replaced.

16.10 c For the sample mean only one; for the sample median at least $\lfloor (n+1)/2 \rfloor$ elements.

16.12 $\bar{x}_n = (N + 1)/2$; $\text{Med}_n = (N + 1)/2$.

16.15 Write $(x_i - \bar{x}_n)^2 = x_i^2 - 2\bar{x}_n x_i + \bar{x}_n^2$.

17.1

$N(3, 1)$	$N(0, 1)$	$N(0, 1)$
$N(3, 1)$	$Exp(1/3)$	$Exp(1)$
$N(0, 1)$	$N(0, 9)$	$Exp(1)$
$N(3, 1)$	$N(0, 9)$	$Exp(1/3)$
$N(0, 9)$	$Exp(1/3)$	$Exp(1)$

17.2

$Exp\,(1/3)$	$N(0,9)$	$Exp\,(1/3)$
$N(0,1)$	$N(3,1)$	$Exp\,(1)$
$N(0,9)$	$N(0,9)$	$N(3,1)$
$Exp\,(1)$	$N(3,1)$	$Exp\,(1)$
$N(0,1)$	$N(0,1)$	$Exp\,(1/3)$

17.3 a $Bin\,(10,p)$.

17.3 b $p = 0.435$.

17.5 a One possibility is $p = 93/331$; another is $p = 29/93$.

17.5 b $p = 474/1285$ or $p = 198/474$.

17.5 c 0.6281 or 0.6741 for smokers and 0.7486 or 0.8026 for nonsmokers.

17.7 a An exponential distribution.

17.7 b One possibility is $\lambda = 0.00469$.

17.9 a Recall the formula for the volume of a cylinder with diameter d (at the base) and height h.

17.9 b $\bar{z}_n = 0.3022$; $\bar{y}/\bar{x} = 0.3028$; least squares: 0.3035.

18.1 $5^6 = 15625$. Not equally likely.

18.3 a 0.0574.

18.3 b 0.0547.

18.3 c 0.000029.

18.4 a 0.3487.

18.4 b $(1 - 1/n)^n$.

18.5 values $0, \pm1, \pm2,$ and ±3 with probabilities 7/27, 6/27, 3/27, and 1/27.

18.7 Determine from which parametric distribution you generate the bootstrap datasets and what the bootstrapped version is of $\bar{X}_n - \mu$.

18.8 a Determine from which $\hat{F}$ you generate the bootstrap datasets and what the bootstrapped version is of $\bar{X}_n - \mu$.

18.8 b Similar to **a**.

18.8 c Similar to **a** and **b**.

18.9 Determine which normal distribution corresponds to $X_1^*, X_2^*, \ldots, X_n^*$ and use this to compute $P\left(|\bar{X}_n^* - \mu^*| > 1\right)$.

19.1 a First show that $E\left[X_1^2\right] = \theta^2/3$, and use linearity of expectations.

19.1 b $\sqrt{T}$ has negative bias.

19.3 $a = 1/n$, $b = 0$.

19.5 $c = n$.

19.6 a Use linearity of expectations and plug in the expressions for $E[M_n]$ and $E\left[\bar{X}_n\right]$.

19.6 b $(nM_n - \bar{X}_n)/(n-1)$.

19.6 c Estimate for δ: 2073.5.

19.8 Check that $E[Y_i] = \beta x_i$ and use linearity of expectations.

20.2 a We prefer T.

20.2 b If $a < 6$ we prefer T; if $a \geq 6$ we prefer S.

20.3 T_1.

20.4 a $E[3L-1]=3E[N+1-M]-1=N$.

20.4 b $(N+1)(N-2)/2$.

20.4 c 4 times.

20.7 $\mathrm{Var}(T_1) = (4 - \theta^2)/n$ and $\mathrm{Var}(T_2) = \theta(4 - \theta)/n$. We prefer T_2.

20.8 a Use linearity of expectations.

20.8 b Differentiate with respect to r.

20.11 $\mathrm{MSE}(T_1) = \sigma^2/(\sum_{i=1}^{n} x_i^2)$, $\mathrm{MSE}(T_2) = (\sigma^2/n^2)\cdot\sum_{i=1}^{n}(1/x_i^2)$, $\mathrm{MSE}(T_3) = \sigma^2\,n/(\sum_{i=1}^{n} x_i)^2$.

21.1 D_2.

21.2 $\hat{p} = 1/4$.

21.4 a Use that $X_1, \ldots, X_n$ are independent $Pois\,(\mu)$ distributed random variables.

21.4 b $\ell(\mu) = \left(\sum_{i=1}^{n} x_i\right)\ln(\mu) - \ln(x_1! \cdot x_2! \cdots x_n!) - n\mu$, $\hat{\mu} = \bar{x}_n$.

21.4 c $e^{-\bar{x}_n}$.

21.5 a $\bar{x}_n$.

21.5 b $\sqrt{\dfrac{1}{n}\sum_{i=1}^{n} x_i^2}$.

21.7 $\sqrt{\frac{1}{2n}\sum_{i=1}^{n} x_i^2}$.

21.8 a $L(\theta) = \frac{C}{43839} \cdot (2 + \theta)^{1997} \cdot \theta^{32} \cdot (1 - \theta)^{1810}$; $\ell(\theta) = \ln(C) - 3839 \ln(4) + 1997 \ln(2 + \theta) + 32 \ln(\theta) + 1810 \ln(1 - \theta)$.

21.8 b 0.0357.

21.8 c $(-b + \sqrt{D})/(2n)$, with $b = -n_1 + n_2 + 2n_3 + 2n_4$, and $D = (n_1 - n_2 - 2n_3 - 2n_4)^2 + 8nn_2$.

21.9 $\hat{\alpha} = x_{(1)}$ and $\hat{\beta} = x_{(n)}$.

21.11 a $1/\bar{x}_n$.

21.11 b $y_{(n)}$.

22.1 a $\hat{\alpha} = 2.35$, $\hat{\beta} = -0.25$.

22.1 b $r_1 = -0.1$, $r_2 = 0.2$, $r_3 = -0.1$.

22.1 c The estimated regression line goes through $(0, 2.35)$ and $(3, 1.6)$.

22.5 Minimize $\sum_{i=1}^{n} (y_i - \beta x_i)^2$.

22.6 2218.45.

22.8 The model with no intercept.

22.10 a $\hat{\alpha} = 7/3$, $\hat{\beta} = -1$, $A(\hat{\alpha}, \hat{\beta}) = 4/3$.

22.10 b $17/9 < \alpha < 7/3$, $\alpha = 2$.

22.10 c $\alpha = 2$, $\beta = -1$.

22.12 a Use that the denominator of $\hat{\beta}$ and that $\sum x_i$ are numbers, *not* random variables.

22.12 b Use that $E[Y_i] = \alpha + \beta x_i$.

22.12 c Simplify the expression in **b**.

22.12 d Combine **a** and **c**.

23.1 $(740.55, 745.45)$.

23.2 $(3.486, 3.594)$.

23.5 a $(0.050, 1.590)$.

23.5 b See Section 23.3.

23.5 c $(0.045, 1.600)$.

23.6 a Rewrite the probability in terms of L_n and U_n.

23.6 b $(3l_n + 7, 3u_n + 7)$.

23.6 c $\tilde{L}_n = 1 - U_n$ and $\tilde{U}_n = 1 - L_n$. The confidence interval: $(-4, 3)$.

23.6 d $(0, 25)$ is a conservative 95% confidence interval for θ.

23.7 $(e^{-3}, e^{-2}) = (0.050, 0.135)$.

23.11 a Yes.

23.11 b Not necessarily.

23.11 c Not necessarily.

24.1 $(0.620, 0.769)$.

24.4 a 609.

24.4 b No.

24.6 a $(1.68, \infty)$.

24.6 b $[0, 2.80)$.

24.8 a $(0.449, 0.812)$.

24.8 b $(0.481, 1]$.

24.9 a See Section 8.4.

24.9 b $c_l = 0.779$, $c_u = 0.996$.

24.9 c $(3.013, 3.851)$.

24.9 d $(m/(1 - \alpha/2)^{1/n}, m/(\alpha/2)^{1/n})$.

25.2 $H_1 : \mu > 1472$.

25.4 a The difference or the ratio of the average numbers of cycles for the two groups.

25.4 b The difference or the ratio of the maximum likelihood estimators $\hat{p}_1$ and $\hat{p}_2$.

25.4 c $H_1 : p_1 < p_2$.

25.5 a Relevant values of T_1 are in $[0, 5]$; those close to 0, or close to 5, are in favor of H_1.

25.5 b Relevant values of T_2 are in $[0, 5]$; only those close to 0 are in favor of H_1.

25.6 a The p-value is 0.23. Do not reject.

25.6 b The p-value is 0.77. Do not reject.

25.6 c The p-value is 0.968. Do not reject.

25.6 d The p-value is 0.019. Reject.

25.6 e The p-value is 0.99. Do not reject.

25.6 f The p-value is smaller than 0.019. Reject.

25.6 g The p-value is smaller than 0.200. We cannot say anything about rejection of H_0.

25.10 a $H_1 : \mu > 23.75$.

25.10 b The p-value is 0.0344.

25.11 0.0456.

26.3 a 0.1.

26.3 b 0.72.

26.5 a The p-value is 0.1050. Do not reject H_0; this agrees with Exercise 24.8 **b**.

26.5 b $K = \{16, 17, \ldots, 23\}$.

26.5 c 0.0466.

26.5 d 0.6950.

26.6 a Right critical value.

26.6 b Right critical value $c = 1535.1$; critical region $[1536, \infty)$.

26.8 a For T we find $K = (0, c_l]$ and for T' we find $K' = [c_u, 1)$.

26.8 b For T we find $K = (0, c_l] \cup [c_u, \infty)$ and for T' we find $K' = (0, c'_l] \cup [c'_u, 1)$.

26.9 a For T we find $K = [c_u, \infty)$ and for T' we find $K' = [c'_l, 0) \cup (0, c'_u]$.

26.9 b For T we find $K = [c_u, \infty)$ and for T' we find $K' = (0, c'_u]$.

27.2 a $H_0 : \mu = 2550$ and $H_1 : \mu \neq 2550$.

27.2 b $t = 1.2096$. Do not reject H_0.

27.5 a $H_0 : \mu = 0$; $H_1 : \mu > 0$; $t = 0.70$.

27.5 b p-value: 0.2420. Do not reject H_0.

27.7 a $H_0 : \beta = 0$ and $H_1 : \beta < 0$; $t_b = -20.06$. Reject H_0.

27.7 b Same testing problem; $t_b = -11.03$. Reject H_0.

28.1 a $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$; $t_p = -2.130$. Reject H_0.

28.1 b $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$; $t_d = -2.130$. Reject H_0.

28.1 c Reject H_0. The salaries differ significantly.

28.3 a $t_p = 2.492$. Reject H_0.

28.3 b Reject H_0.

28.3 c $t_d = 2.463$. Reject H_0.

28.3 d Reject H_0.

28.5 a Determine $\mathrm{E}\left[aS_X^2 + bS_Y^2\right]$, using that S_X^2 and S_Y^2 are both unbiased for σ^2.

28.5 b Determine $\mathrm{E}\left[aS_X^2 + (1-a)S_Y^2\right]$, using that S_X^2 and S_Y^2 are independent, and minimize over a.

Full solutions to selected exercises

2.8 From the rule for the probability of a union we obtain $P(D_1 \cup D_2) \leq P(D_1) + P(D_2) = 2 \cdot 10^{-6}$. Since $D_1 \cap D_2$ is contained in both D_1 and D_2, we obtain $P(D_1 \cap D_2) \leq \min\{P(D_1), P(D_2)\} = 10^{-6}$. Equality may hold in both cases: for the union, take D_1 and D_2 disjoint, for the intersection, take D_1 and D_2 equal to each other.

2.12 a This is the same situation as with the three envelopes on the doormat, but now with ten possibilities. Hence an outcome has probability $1/10!$ to occur.

2.12 b For the five envelopes labeled $1, 2, 3, 4, 5$ there are $5!$ possible orders, and for each of these there are $5!$ possible orders for the envelopes labeled $6, 7, 8, 9, 10$. Hence in total there are $5! \cdot 5!$ outcomes.

2.12 c There are $32 \cdot 5! \cdot 5!$ outcomes in the event "dream draw." Hence the probability is $32 \cdot 5! 5!/10! = 32 \cdot 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5/(6 \cdot 7 \cdot 8 \cdot 9 \cdot 10) = 8/63 = 12.7$ percent.

2.14 a Since door a is never opened, $P((a, a)) = P((b, a)) = P((c, a)) = 0$. If the candidate chooses a (which happens with probability $1/3$), then the quizmaster chooses without preference from doors b and c. This yields that $P((a, b)) = P((a, c)) = 1/6$. If the candidate chooses b (which happens with probability $1/3$), then the quizmaster can only open door c. Hence $P((b, c)) = 1/3$. Similarly, $P((c, b)) = 1/3$. Clearly, $P((b, b)) = P((c, c)) = 0$.

2.14 b If the candidate chooses a then she or he wins; hence the corresponding event is $\{(a, a), (a, b), (a, c)\}$, and its probability is $1/3$.

2.14 c To end with a the candidate should have chosen b or c. So the event is $\{(b, c), (c, b)\}$ and $P(\{(b, c), (c, b)\}) = 2/3$.

2.16 Since $E \cap F \cap G = \emptyset$, the three sets $E \cap F$, $F \cap G$, and $E \cap G$ are disjoint. Since each has probability $1/3$, they have probability 1 together. From these two facts one deduces $P(E) = P(E \cap F) + P(E \cap G) = 2/3$ (make a diagram or use that $E = E \cap (E \cap F) \cup E \cap (F \cap G) \cup E \cap (E \cap G)$).

3.1 Define the following events: B is the event "point B is reached on the second step," C is the event "the path to C is chosen on the first step," and similarly we define D and E. Note that the events C, D, and E are mutually exclusive and that one of them must occur. Furthermore, that we can only reach B by first going to C

or D. For the computation we use the law of total probability, by conditioning on the result of the first step:

$$\begin{aligned}
P(B) &= P(B \cap C) + P(B \cap D) + P(B \cap E) \\
&= P(B \,|\, C) \, P(C) + P(B \,|\, D) \, P(D) + P(B \,|\, E) \, P(E) \\
&= \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{1}{3} + \frac{1}{3} \cdot 0 = \frac{7}{36}.
\end{aligned}$$

3.2 a Event A has three outcomes, event B has 11 outcomes, and $A \cap B = \{(1,3), (3,1)\}$. Hence we find $P(B) = 11/36$ and $P(A \cap B) = 2/36$ so that

$$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)} = \frac{2/36}{11/36} = \frac{2}{11}.$$

3.2 b Because $P(A) = 3/36 = 1/12$ and this is not equal to $2/11 = P(A \,|\, B)$ the events A and B are *dependent*.

3.3 a There are 13 spades in the deck and each has probability $1/52$ of being chosen, hence $P(S_1) = 13/52 = 1/4$. Given that the first card is a spade there are $13 - 1 = 12$ spades left in the deck with $52 - 1 = 51$ remaining cards, so $P(S_2 \,|\, S_1) = 12/51$. If the first card is not a spade there are 13 spades left in the deck of 51, so $P(S_2 \,|\, S_1^c) = 13/51$.

3.3 b We use the law of total probability (based on $\Omega = S_1 \cup S_1^c$):

$$\begin{aligned}
P(S_2) &= P(S_2 \cap S_1) + P(S_2 \cap S_1^c) = P(S_2 \,|\, S_1) \, P(S_1) + P(S_2 \,|\, S_1^c) \, P(S_1^c) \\
&= \frac{12}{51} \cdot \frac{1}{4} + \frac{13}{51} \cdot \frac{3}{4} = \frac{12 + 39}{51 \cdot 4} = \frac{1}{4}.
\end{aligned}$$

3.7 a The best approach to a problem like this one is to write out the conditional probability and then see if we can somehow combine this with $P(A) = 1/3$ to solve the puzzle. Note that $P(B \cap A^c) = P(B \,|\, A^c) \, P(A^c)$ and that $P(A \cup B) = P(A) + P(B \cap A^c)$. So

$$P(A \cup B) = \frac{1}{3} + \frac{1}{4} \cdot \left(1 - \frac{1}{3}\right) = \frac{1}{3} + \frac{1}{6} = \frac{1}{2}.$$

3.7 b From the conditional probability we find $P(A^c \cap B^c) = P(A^c \,|\, B^c) \, P(B^c) = \frac{1}{2}(1 - P(B))$. Recalling DeMorgan's law we know $P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 1/3$. Combined this yields an equation for $P(B)$: $\frac{1}{2}(1 - P(B)) = 1/3$ from which we find $P(B) = 1/3$.

3.8 a This asks for $P(W)$. We use the law of total probability, decomposing $\Omega = F \cup F^c$. Note that $P(W \,|\, F) = 0.99$.

$$\begin{aligned}
P(W) &= P(W \cap F) + P(W \cap F^c) = P(W \,|\, F) \, P(F) + P(W \,|\, F^c) \, P(F^c) \\
&= 0.99 \cdot 0.1 + 0.02 \cdot 0.9 = 0.099 + 0.018 = 0.117.
\end{aligned}$$

3.8 b We need to determine $P(F \,|\, W)$, and this can be done using Bayes' rule. Some of the necessary computations have already been done in **a**, we can copy $P(W \cap F)$ and $P(W)$ and get:

$$P(F \,|\, W) = \frac{P(F \cap W)}{P(W)} = \frac{0.099}{0.117} = 0.846.$$

4.1 a In two independent throws of a die there are 36 possible outcomes, each occurring with probability 1/36. Since there are 25 ways to have no 6's, 10 ways to have one 6, and one way to have two 6's, we find that $p_Z(0) = 25/36$, $p_Z(1) = 10/36$, and $p_Z(2) = 1/36$. So the probability mass function p_Z of Z is given by the following table:

z	0	1	2
$p_Z(z)$	$\frac{25}{36}$	$\frac{10}{36}$	$\frac{1}{36}$

The distribution function F_Z is given by

$$F_Z(a) = \begin{cases} 0 & \text{for } a < 0 \\ \frac{25}{36} & \text{for } 0 \le a < 1 \\ \frac{25}{36} + \frac{10}{36} = \frac{35}{36} & \text{for } 1 \le a < 2 \\ \frac{25}{36} + \frac{10}{36} + \frac{1}{36} = 1 & \text{for } a \ge 2. \end{cases}$$

Z is the sum of two independent $Ber(1/6)$ distributed random variables, so Z has a $Bin(2, 1/6)$ distribution.

4.1 b If we denote the outcome of the two throws by (i, j), where i is the outcome of the first throw and j the outcome of the second, then $\{M = 2, Z = 0\} = \{(2, 1), (1, 2), (2, 2)\}$, $\{S = 5, Z = 1\} = \emptyset$, $\{S = 8, Z = 1\} = \{(6, 2), (2, 6)\}$. Furthermore, $P(M = 2, Z = 0) = 3/36$, $P(S = 5, Z = 1) = 0$, and $P(S = 8, Z = 1) = 2/36$.

4.1 c The events are dependent, because, e.g., $P(M = 2, Z = 0) = \frac{3}{36}$ differs from $P(M = 2) \cdot P(Z = 0) = \frac{3}{36} \cdot \frac{253}{36}$.

4.10 a Each R_i has a Bernoulli distribution, because it can only attain the values 0 and 1. The parameter is $p = P(R_i = 1)$. It is not easy to determine $P(R_i = 1)$, but it is fairly easy to determine $P(R_i = 0)$. The event $\{R_i = 0\}$ occurs when none of the m people has chosen the ith floor. Since they make their choices independently of each other, and each floor is selected by each of these m people with probability 1/21, it follows that

$$P(R_i = 0) = \left(\frac{20}{21}\right)^m.$$

Now use that $p = P(R_i = 1) = 1 - P(R_i = 0)$ to find the desired answer.

4.10 b If $\{R_1 = 0\}, \ldots, \{R_{20} = 0\}$, we must have that $\{R_{21} = 1\}$, so we cannot conclude that the events $\{R_1 = a_1\}, \ldots, \{R_{21} = a_{21}\}$, where a_i is 0 or 1, are independent. Consequently, we cannot use the argument from Section 4.3 to conclude that S_m is $Bin(21, p)$. In fact, S_m is not $Bin(21, p)$ distributed, as the following shows. The elevator will stop at least once, so $P(S_m = 0) = 0$. However, if S_m would have a $Bin(21, p)$ distribution, then $P(S_m = 0) = (1 - p)^{21} > 0$, which is a contradiction.

4.10 c This exercise is a variation on finding the probability of no coincident birthdays from Section 3.2. For $m = 2$, $S_2 = 1$ occurs precisely if the two persons entering the elevator select the same floor. The first person selects any of the 21 floors, the second selects the same floor with probability 1/21, so $P(S_2 = 1) = 1/21$. For $m = 3$, $S_3 = 1$ occurs if the second and third persons entering the elevator both select the same floor as was selected by the first person, so $P(S_3 = 1) = (1/21)^2 = 1/441$. Furthermore, $S_3 = 3$ occurs precisely when all three persons choose a different floor. Since there are $21 \cdot 20 \cdot 19$ ways to do this out of a total of 21^3 possible ways, we

find that $P(S_3 = 3) = 380/441$. Since S_3 can only attain the values $1, 2, 3$, it follows that $P(S_3 = 2) = 1 - P(S_3 = 1) - P(S_3 = 3) = 60/441$.

4.13 a Since we wait for the first time we draw the marked bolt in independent draws, each with a $Ber(p)$ distribution, where p is the probability to draw the bolt (so $p = 1/N$), we find, using a reasoning as in Section 4.4, that X has a $Geo(1/N)$ distribution.

4.13 b Clearly, $P(Y = 1) = 1/N$. Let D_i be the event that the marked bolt was drawn (for the first time) in the ith draw. For $k = 2, \ldots, N$ we have that

$$P(Y = k) = P(D_1^c \cap \cdots \cap D_{k-1}^c \cap D_k)$$
$$= P(D_k \mid D_1^c \cap \cdots \cap D_{k-1}^c) \cdot P(D_1^c \cap \cdots \cap D_{k-1}^c).$$

Now $P(D_k \mid D_1^c \cap \cdots \cap D_{k-1}^c) = \frac{1}{N-k+1}$,

$$P(D_1^c \cap \cdots \cap D_{k-1}^c) = P(D_{k-1}^c \mid D_1^c \cap \cdots \cap D_{k-2}^c) \cdot P(D_1^c \cap \cdots \cap D_{k-2}^c),$$

and

$$P(D_{k-1}^c \mid D_1^c \cap \cdots \cap D_{k-1}^c) = 1 - P(D_{k-1} \mid D_1^c \cap \cdots \cap D_{k-1}^c) = 1 - \frac{1}{N-k+2}.$$

Continuing in this way, we find after k steps that

$$P(Y = k) = \frac{1}{N-k+1} \cdot \frac{N-k+1}{N-k+2} \cdot \frac{N-k+2}{N-k+3} \cdots \frac{N-2}{N-1} \cdot \frac{N-1}{N} = \frac{1}{N}.$$

See also Section 9.3, where the distribution of Y is derived in a different way.

4.13 c For $k = 0, 1, \ldots, r$, the probability $P(Z = k)$ is equal to the number of ways the event $\{Z = k\}$ can occur, divided by the number of ways $\binom{N}{r}$ we can select r objects from N objects, see also Section 4.3. Since one can select k marked bolts from m marked ones in $\binom{m}{k}$ ways, and $r - k$ nonmarked bolts from $N - m$ nonmarked ones in $\binom{N-m}{r-k}$ ways, it follows that

$$P(Z = k) = \frac{\binom{m}{k}\binom{N-m}{r-k}}{\binom{N}{r}}, \quad \text{for } k = 0, 1, 2, \ldots, r.$$

5.4 a Let T be the time until the next arrival of a bus. Then T has $U(4, 6)$ distribution. Hence $P(X \leq 4.5) = P(T \leq 4.5) = \int_4^{4.5} 1/2 \, dx = 1/4$.

5.4 b Since Jensen leaves when the next bus arrives after more than 5 minutes, $P(X = 5) = P(T > 5) = \int_5^6 \frac{1}{2} \, dx = 1/2$.

5.4 c Since $P(X = 5) = 0.5 > 0$, X cannot be continuous. Since X can take any of the uncountable values in $[4, 5]$, it can also not be discrete.

5.8 a The probability density $g(y) = 1/(2\sqrt{ry})$ has an asymptote in 0 and decreases to $1/2r$ in the point r. Outside $[0, r]$ the function is 0.

5.8 b The second darter is better: for each $0 < b < r$ one has $(b/r)^2 < \sqrt{b/r}$ so the second darter always has a larger probability to get closer to the center.

5.8 c Any function F that is 0 left from 0, increasing on $[0, r]$, takes the value 0.9 in $r/10$, and takes the value 1 in r and to the right of r is a correct answer to this question.

5.13 a This follows with a change of variable transformation $x \mapsto -x$ in the integral:
$\Phi(-a) = \int_{-\infty}^{-a} \phi(x)\, dx = \int_a^{\infty} \phi(-x)\, dx = \int_a^{\infty} \phi(x)\, dx = 1 - \Phi(a)$.

5.13 b This is straightforward: $P(Z \le -2) = \Phi(-2) = 1 - \Phi(2) = 0.0228$.

6.5 We see that

$$X \le a \quad \Leftrightarrow \quad -\ln U \le a \quad \Leftrightarrow \quad \ln U \ge -a \quad \Leftrightarrow \quad U \ge e^{-a},$$

and so $P(X \le a) = P(U \ge e^{-a}) = 1 - P(U \le e^{-a}) = 1 - e^{-a}$, where we use
$P(U \le p) = p$ for $0 \le p \le 1$ applied to $p = e^{-a}$ (remember that $a \ge 0$).

6.7 We need to obtain F^{inv}, and do this by solving $F(x) = u$, for $0 \le u \le 1$:

$$1 - e^{-5x^2} = u \quad \Leftrightarrow \quad e^{-5x^2} = 1 - u \quad \Leftrightarrow \quad -5x^2 = \ln(1-u)$$
$$\Leftrightarrow \quad x^2 = -0.2\ln(1-u) \quad \Leftrightarrow \quad x = \sqrt{-0.2\ln(1-u)}.$$

The solution is $Z = \sqrt{-0.2\ln U}$ (replacing $1 - U$ by U, see Exercise 6.3). Note that
Z^2 has an $Exp(5)$ distribution.

6.10 a Define random variables $B_i = 1$ if $U_i \le p$ and $B_i = 0$ if $U_i > p$. Then
$P(B_i = 1) = p$ and $P(B_i = 0) = 1 - p$: each B_i has a $Ber(p)$ distribution. If $B_1 = B_2 = \cdots = B_{k-1} = 0$ and $B_k = 1$, then $N = k$, i.e., N is the position in the
sequence of Bernoulli random variables, where the first 1 occurs. This is a $Geo(p)$
distribution. This can be verified by computing the probability mass function: for
$k \ge 1$,

$$P(N = k) = P(B_1 = B_2 = \cdots = B_{k-1} = 0, B_k = 1)$$
$$= P(B_1 = 0)\, P(B_2 = 0) \cdots P(B_{k-1} = 0)\, P(B_k = 1)$$
$$= (1 - p)^{k-1} p.$$

6.10 b If Y is (a real number!) greater than n, then rounding *upwards* means we
obtain $n + 1$ or higher, so $\{Y > n\} = \{Z \ge n + 1\} = \{Z > n\}$. Therefore,
$P(Z > n) = P(Y > n) = e^{-\lambda n} = (e^{-\lambda})^n$. From $\lambda = -\ln(1 - p)$ we see: $e^{-\lambda} = 1 - p$,
so the last probability is $(1 - p)^n$. From $P(Z > n - 1) = P(Z = n) + P(Z > n)$ we
find: $P(Z = n) = P(Z > n - 1) - P(Z > n) = (1 - p)^{n-1} - (1 - p)^n = (1 - p)^{n-1} p$.
Z has a $Geo(p)$ distribution.

6.12 We need to generate stock prices for the next five years, or 60 months. So we
need sixty $U(0,1)$ random variables $U_1, \ldots, U_{60}$. Let S_i denote the stock price in
month i, and set $S_0 = 100$, the initial stock price. From the U_i we obtain the stock
movement, as follows, for $i = 1, 2, \ldots$:

$$S_i = \begin{cases} 0.95\, S_{i-1} & \text{if } U_i < 0.25, \\ S_{i-1} & \text{if } 0.25 \le U_i \le 0.75, \\ 1.05\, S_{i-1} & \text{if } U_i > 0.75. \end{cases}$$

We have carried this out, using the realizations below:

1–10:	0.72	0.03	0.01	0.81	0.97	0.31	0.76	0.70	0.71 0.25
11–20:	0.88	0.25	0.89	0.95	0.82	0.52	0.37	0.40	0.82 0.04
21–30:	0.38	0.88	0.81	0.09	0.36	0.93	0.00	0.14	0.74 0.48
31–40:	0.34	0.34	0.37	0.30	0.74	0.03	0.16	0.92	0.25 0.20
41–50:	0.37	0.24	0.09	0.69	0.91	0.04	0.81	0.95	0.29 0.47
51–60:	0.19	0.76	0.98	0.31	0.70	0.36	0.56	0.22	0.78 0.41

We do not list all the stock prices, just the ones that matter for our investment strategy (you can verify this). We first wait until the price drops below € 95, which happens at $S_4 = 94.76$. Our money has been in the bank for four months, so we own € $1000 \cdot 1.005^4 = $ € 1020.15, for which we can buy $1020.15/94.76 = 10.77$ shares. Next we wait until the price hits € 110, this happens at $S_{15} = 114.61$. We sell the our shares for € $10.77 \cdot 114.61 = $ € 1233.85, and put the money in the bank. At $S_{42} = 92.19$ we buy stock again, for the € $1233.85 \cdot 1.005^{27} = $ € 1411.71 that has accrued in the bank. We can buy 15.31 shares. For the rest of the five year period nothing happens, the final price is $S_{60} = 100.63$, which puts the value of our portfolio at € 1540.65.

For a real simulation the above should be repeated, say, one thousand times. The one thousand net results then give us an impression of the probability distribution that corresponds to this model and strategy.

7.6 Since f is increasing on the interval $[2, 3]$ we know from the interpretation of expectation as center of gravity that the expectation should lie closer to 3 than to 2. The computation: $\mathrm{E}[Z] = \int_2^3 \frac{3}{19} z^3 \, \mathrm{d}z = \left[\frac{3}{76} z^4\right]_2^3 = 2\frac{43}{76}$.

7.15 a We use the change-of-units rule for the expectation twice:

$$\mathrm{Var}(rX) = \mathrm{E}\left[(rX - \mathrm{E}[rX]^2)\right] = \mathrm{E}\left[(rX - r\mathrm{E}[X])^2\right]$$
$$= \mathrm{E}\left[r^2(X - \mathrm{E}[X])^2\right] = r^2\mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = r^2\mathrm{Var}(X).$$

7.15 b Now we use the change-of-units rule for the expectation once:

$$\mathrm{Var}(X + s) = \mathrm{E}\left[((X + s) - \mathrm{E}[X + s])^2\right]$$
$$= \mathrm{E}\left[((X + s) - \mathrm{E}[X] + s)^2\right] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{Var}(X).$$

7.15 c With first **b**, and then **a**: $\mathrm{Var}(rX + s) = \mathrm{Var}(rX) = r^2\mathrm{Var}(X).$

7.17 a Since $a_i \geq 0$ and $p_i \geq 0$ it must follow that $a_1 p_1 + \cdots + a_r p_r \geq 0$. So $0 = \mathrm{E}[U] = a_1 p_1 + \cdots + a_r p_r \geq 0$. As we may assume that all $p_i > 0$, it follows that $a_1 = a_2 = \cdots = a_r = 0$.

7.17 b Let $m = \mathrm{E}[V] = p_1 b_1 + \cdots + p_r b_r$. Then the random variable $U = (V - \mathrm{E}[V])^2$ takes the values $a_1 = (b_1 - m)^2, \ldots, a_r = (b_r - m)^2$. Since $\mathrm{E}[U] = \mathrm{Var}(V) = 0$, part **a** tells us that $0 = a_1 = (b_1 - m)^2, \ldots, 0 = a_r = (b_r - m)^2$. But this is only possible if $b_1 = m, \ldots, b_r = m$. Since $m = \mathrm{E}[V]$, this is the same as saying that $P(V = \mathrm{E}[V]) = 1$.

8.2 a First we determine the possible values that Y can take. Here these are $-1, 0,$ and 1. Then we investigate which x-values lead to these y-values and sum the probabilities of the x-values to obtain the probability of the y-value. For instance,

$$P(Y = 0) = P(X = 2) + P(X = 4) + P(X = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Similarly, we obtain for the two other values

$$P(Y = -1) = P(X = 3) = \frac{1}{6}, \qquad P(Y = 1) = P(X = 1) + P(X = 5) = \frac{1}{3}.$$

8.2 b The values taken by Z are -1, 0, and 1. Furthermore

$$P(Z = 0) = P(X = 1) + P(X = 3) + P(X = 5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2},$$

and similarly $P(Z = -1) = 1/3$ and $P(Z = 1) = 1/6$.

8.2 c Since for any α one has $\sin^2(\alpha) + \cos^2(\alpha) = 1$, W can only take the value 1, so $P(W = 1) = 1$.

8.10 Because of symmetry: $P(X \geq 3) = 0.500$. Furthermore: $\sigma^2 = 4$, so $\sigma = 2$. Then $Z = (X - 3)/2$ is an $N(0, 1)$ distributed random variable, so that $P(X \leq 1) = P((X - 3)/2) \leq (1 - 3)/2 = P(Z \leq -1) = P(Z \geq 1) = 0.1587$.

8.11 Since $-g$ is a convex function, Jensen's inequality yields that $-g(\mathrm{E}[X]) \leq \mathrm{E}[-g(X)]$. Since $\mathrm{E}[-g(X)] = -\mathrm{E}[g(X)]$, the inequality follows by multiplying both sides by -1.

8.12 a The possible values Y can take are $\sqrt{0} = 0$, $\sqrt{1} = 1$, $\sqrt{100} = 10$, and $\sqrt{10\,000} = 100$. Hence the probability mass function is given by

y	0	1	10	100
$P(Y = y)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

8.12 b Compute the second derivative: $\frac{d^2}{dx^2}\sqrt{x} = -\frac{1}{4}x^{-3/2} < 0$. Hence $g(x) = -\sqrt{x}$ is a convex function. Jensen's inequality yields that $\sqrt{\mathrm{E}[X]} \geq \mathrm{E}\left[\sqrt{X}\right]$.

8.12 c We obtain $\sqrt{\mathrm{E}[X]} = \sqrt{(0 + 1 + 100 + 10\,000)/4} = 50.25$, but

$$\mathrm{E}\left[\sqrt{X}\right] = \mathrm{E}[Y] = (0 + 1 + 10 + 100)/4 = 27.75.$$

8.19 a This happens for all φ in the interval $[\pi/4, \pi/2]$, which corresponds to the upper right quarter of the circle.

8.19 b Since $\{Z \leq t\} = \{X \leq \arctan(t)\}$, we obtain

$$F_Z(t) = P(Z \leq t) = P(X \leq \arctan(t)) = \frac{1}{2} + \frac{1}{\pi}\arctan(t).$$

8.19 c Differentiating F_Z we obtain that the probability density function of Z is

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \frac{d}{dz}\left(\frac{1}{2} + \frac{1}{\pi}\arctan(z)\right) = \frac{1}{\pi(1 + z^2)} \quad \text{for } -\infty < z < \infty.$$

9.2 a From $P(X = 1, Y = 1) = 1/2$, $P(X = 1) = 2/3$, and the fact that $P(X = 1) = P(X = 1, Y = 1) + P(X = 1, Y = -1)$, it follows that $P(X = 1, Y = -1) = 1/6$. Since $P(Y = 1) = 1/2$ and $P(X = 1, Y = 1) = 1/2$, we must have: $P(X = 0, Y = 1)$ and $P(X = 2, Y = 1)$ are both zero. From this and the fact that $P(X = 0) = 1/6 = P(X = 2)$ one finds that $P(X = 0, Y = -1) = 1/6 = P(X = 2, Y = -1)$.

9.2 b Since, e.g., $P(X = 2, Y = 1) = 0$ is different from $P(X = 2)P(Y = 1) = \frac{1}{6} \cdot \frac{1}{2}$, one finds that X and Y are dependent.

9.8 a Since X can attain the values 0 and 1 and Y the values 0 and 2, Z can attain the values 0, 1, 2, and 3 with probabilities: $P(Z = 0) = P(X = 0, Y = 0) = 1/4$, $P(Z = 1) = P(X = 1, Y = 0) = 1/4$, $P(Z = 2) = P(X = 0, Y = 2) = 1/4$, and $P(Z = 3) = P(X = 1, Y = 2) = 1/4$.

9.8 b Since $\tilde{X} = \tilde{Z} - \tilde{Y}$, $\tilde{X}$ can attain the values -2, -1, 0, 1, 2, and 3 with probabilities

$$P\left(\tilde{X} = -2\right) = P\left(\tilde{Z} = 0, \tilde{Y} = 2\right) = 1/8,$$

$$P\left(\tilde{X} = -1\right) = P\left(\tilde{Z} = 1, \tilde{Y} = 2\right) = 1/8,$$

$$P\left(\tilde{X} = 0\right) = P\left(\tilde{Z} = 0, \tilde{Y} = 0\right) + P\left(\tilde{Z} = 2, \tilde{Y} = 2\right) = 1/4,$$

$$P\left(\tilde{X} = 1\right) = P\left(\tilde{Z} = 1, \tilde{Y} = 0\right) + P\left(\tilde{Z} = 3, \tilde{Y} = 2\right) = 1/4,$$

$$P\left(\tilde{X} = 2\right) = P\left(\tilde{Z} = 2, \tilde{Y} = 0\right) = 1/8,$$

$$P\left(\tilde{X} = 3\right) = P\left(\tilde{Z} = 3, \tilde{Y} = 0\right) = 1/8.$$

We have the following table:

z	-2	-1	0	1	2	3
$p_{\tilde{X}}(z)$	1/8	1/8	1/4	1/4	1/8	1/8

9.9 a One has that $F_X(x) = \lim_{y \to \infty} F(x, y)$. So for $x \le 0$: $F_X(x) = 0$, and for $x > 0$: $F_X(x) = F(x, \infty) = 1 - e^{-2x}$. Similarly, $F_Y(y) = 0$ for $y \le 0$, and for $y > 0$: $F_Y(y) = F(\infty, y) = 1 - e^{-y}$.

9.9 b For $x > 0$ and $y > 0$: $f(x, y) = \frac{\partial^2}{\partial x \, \partial y} F(x, y) = \frac{\partial}{\partial x}\left(e^{-y} - e^{-(2x+y)}\right) = 2e^{-(2x+y)}$.

9.9 c There are two ways to determine $f_X(x)$:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^{\infty} e^{-(2x+y)}\, dy = 2e^{-2x} \quad \text{for } x > 0$$

and

$$f_X(x) = \frac{d}{dx} F_X(x) = 2e^{-2x} \quad \text{for } x > 0.$$

Using either way one finds that $f_Y(y) = e^{-y}$ for $y > 0$.

9.9 d Since $F(x, y) = F_X(x) F_Y(y)$ for all x, y, we find that X and Y are independent.

9.11 To determine $P(X < Y)$ we must integrate $f(x, y)$ over the region G of points (x, y) in $\mathbb{R}^2$ for which x is smaller than y:

$$P(X < Y) = \iint_{\{(x,y)\in\mathbb{R}^2;\, x<y\}} f(x, y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{y} f(x, y)\, dx\right) dy = \int_0^1 \left(\int_0^y \frac{12}{5} xy(1 + y)\, dx\right) dy$$

$$= \frac{12}{5} \int_0^1 y(1 + y) \left(\int_0^y x\, dx\right) dy = \frac{12}{10} \int_0^1 y^3(1 + y)\, dy = \frac{27}{50}.$$

Here we used that $f(x, y) = 0$ for (x, y) outside the unit square.

9.15 a Setting $\square(a, b)$ as the set of points (x, y), for which $x \le a$ and $y \le b$, we have that

$$F(a, b) = \frac{\text{area}\left(\Delta \cap \square(a, b)\right)}{\text{area of } \Delta}.$$

- If $a < 0$ or if $b < 0$ (or both), then area $\left(\Delta \cap \square(a, b)\right) = \emptyset$, so $F(a, b) = 0$,

- If $(a,b) \in \Delta$, then area $(\Delta \cap \square(a,b)) = a(b - \frac{1}{2}a)$, so $F(a,b) = a(2b - a)$,
- If $0 \leq b \leq 1$, and $a > b$, then area $(\Delta \cap \square(a,b)) = \frac{1}{2}b^2$, so $F(a,b) = b^2$,
- If $0 \leq a \leq 1$, and $b > 1$, then area $(\Delta \cap \square(a,b)) = a - \frac{1}{2}a^2$, so $F(a,b) = 2a - a^2$,
- If both $a > 1$ and $b > 1$, then area $(\Delta \cap \square(a,b)) = \frac{1}{2}$, so $F(a,b) = 1$.

9.15 b Since $f(x,y) = \frac{\partial^2}{\partial x\,\partial y}F(x,y)$, we find for $(x,y) \in \Delta$ that $f(x,y) = 2$. Furthermore, $f(x,y) = 0$ for (x,y) outside the triangle Δ.

9.15 c For x between 0 and 1,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{x}^{1} 2\,dy = 2(1-x).$$

For y between 0 and 1,

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{0}^{y} 2\,dx = 2y.$$

10.6 a When $c = 0$, the joint distribution becomes

b	a			$P(Y=b)$
	-1	0	1	
-1	2/45	9/45	4/45	1/3
0	7/45	5/45	3/45	1/3
1	6/45	1/45	8/45	1/3
$P(X=a)$	1/3	1/3	1/3	1

We find $E[X] = (-1) \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0$, and similarly $E[Y] = 0$. By leaving out terms where either $X = 0$ or $Y = 0$, we find

$$E[XY] = (-1) \cdot (-1) \cdot \frac{2}{45} + (-1) \cdot 1 \cdot \frac{4}{45} + 1 \cdot (-1) \cdot \frac{6}{45} + 1 \cdot 1 \cdot \frac{8}{45} = 0,$$

which implies that $\text{Cov}(X,Y) = E[XY] - E[X]E[Y] = 0$.

10.6 b Note that the variables X and Y in part **b** are equal to the ones from part **a**, shifted by c. If we write U and V for the variables from **a**, then $X = U + c$ and $Y = V + c$. According to the rule on the covariance under change of units, we then immediately find $\text{Cov}(X,Y) = \text{Cov}(U+c, V+c) = \text{Cov}(U,V) = 0$.
Alternatively, one could also compute the covariance from $\text{Cov}(X,Y) = E[XY] - E[X]E[Y]$. We find $E[X] = (c-1) \cdot \frac{1}{3} + c \cdot \frac{1}{3} + (c+1) \cdot \frac{1}{3} = c$, and similarly $E[Y] = c$. Since

$$E[XY] = (c-1) \cdot (c-1) \cdot \frac{2}{45} + (c-1) \cdot c \cdot \frac{9}{45} + (c+1) \cdot (c+1) \cdot \frac{4}{45}$$
$$+ c \cdot (c-1) \cdot \frac{7}{45} + c \cdot c \cdot \frac{5}{45} + c \cdot (c+1) \cdot \frac{3}{45}$$
$$+ (c+1) \cdot (c-1) \cdot \frac{6}{45} + (c+1) \cdot c \cdot \frac{1}{45} + (c+1) \cdot (c+1) \cdot \frac{8}{45} = c^2,$$

we find $\text{Cov}(X,Y) = E[XY] - E[X]E[Y] = c^2 - c \cdot c = 0$.

10.6 c No, X and Y are not independent. For instance, $P(X = c, Y = c + 1) = 1/45$, which differs from $P(X = c) \, P(Y = c + 1) = 1/9$.

10.9 a If the aggregated blood sample tests negative, we do not have to perform additional tests, so that X_i takes on the value 1. If the aggregated blood sample tests positive, we have to perform 40 additional tests for the blood sample of each person in the group, so that X_i takes on the value 41. We first find that $P(X_i = 1) = P(\text{no infections in group of } 40) = (1 - 0.001)^{40} = 0.96$, and therefore $P(X_i = 41) = 1 - P(X_i = 1) = 0.04$.

10.9 b First compute $E[X_i] = 1 \cdot 0.96 + 41 \cdot 0.04 = 2.6$. The expected total number of tests is $E[X_1 + X_2 + \cdots + X_{25}] = E[X_1] + E[X_2] + \cdots + E[X_{25}] = 25 \cdot 2.6 = 65$. With the original procedure of blood testing, the total number of tests is $25 \cdot 40 = 1000$. On average the alternative procedure would only require 65 tests. Only with very small probability one would end up with doing more than 1000 tests, so the alternative procedure is better.

10.10 a We find

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_0^3 \frac{2}{225} (9x^3 + 7x^2) \, dx = \frac{2}{225} \left[\frac{9}{4} x^4 + \frac{7}{3} x^3 \right]_0^3 = \frac{109}{50},$$

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y) \, dy = \int_1^2 \frac{1}{25} (3y^3 + 12y^2) \, dy = \frac{1}{25} \left[\frac{3}{4} y^4 + 4y^3 \right]_1^2 = \frac{157}{100},$$

so that $E[X + Y] = E[X] + E[Y] = 15/4$.

10.10 b We find

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) \, dx = \int_0^3 \frac{2}{225} (9x^4 + 7x^3) \, dx = \frac{2}{225} \left[\frac{9}{5} x^5 + \frac{7}{4} x^4 \right]_0^3 = \frac{1287}{250},$$

$$E[Y^2] = \int_{-\infty}^{\infty} y^2 f_Y(y) \, dy = \int_1^2 \frac{1}{25} (3y^4 + 12y^3) \, dy = \frac{1}{25} \left[\frac{3}{5} y^5 + 3y^4 \right]_1^2 = \frac{318}{125},$$

$$E[XY] = \int_0^3 \int_1^2 xy f(x, y) \, dy \, dx = \int_0^3 \int_1^2 \frac{2}{75} (2x^3 y^2 + x^2 y^3) \, dy \, dx$$

$$= \frac{4}{75} \int_0^3 x^3 \left(\int_1^2 y^2 \, dy \right) dx + \frac{2}{75} \int_0^3 x^2 \left(\int_1^2 y^3 \, dy \right) dx$$

$$= \frac{4}{75} \frac{7}{3} \int_0^3 x^3 \, dx + \frac{2}{75} \frac{15}{4} \int_0^3 x^2 \, dx = \frac{171}{50},$$

so that $E[(X + Y)^2] = E[X^2] + E[Y^2] + 2E[XY] = 3633/250$.

10.10 c We find

$$\operatorname{Var}(X) = E[X^2] - (E[X])^2 = \frac{1287}{250} - \left(\frac{109}{50} \right)^2 = \frac{989}{2500},$$

$$\operatorname{Var}(Y) = E[Y^2] - (E[Y])^2 = \frac{318}{125} - \left(\frac{157}{100} \right)^2 = \frac{791}{10\,000},$$

$$\operatorname{Var}(X + Y) = E[(X + Y)^2] - (E[X + Y])^2 = \frac{3633}{250} - \left(\frac{15}{4} \right)^2 = \frac{939}{2000}.$$

Hence, $\operatorname{Var}(X) + \operatorname{Var}(Y) = 0.4747$, which differs from $\operatorname{Var}(X + Y) = 0.4695$.

10.14 a By using the alternative expression for the covariance and linearity of expectations, we find

$$
\begin{aligned}
\mathrm{Cov}&(X+s, Y+u) \\
&= \mathrm{E}\left[(X+s)(Y+u)\right] - \mathrm{E}\left[X+s\right]\mathrm{E}\left[Y+u\right] \\
&= \mathrm{E}\left[XY + sY + uX + su\right] - (\mathrm{E}\left[X\right]+s)(\mathrm{E}\left[Y\right]+u) \\
&= (\mathrm{E}\left[XY\right] + s\mathrm{E}\left[Y\right] + u\mathrm{E}\left[X\right] + su) - (\mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] + s\mathrm{E}\left[Y\right] + u\mathrm{E}\left[X\right] + su) \\
&= \mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] \\
&= \mathrm{Cov}(X, Y).
\end{aligned}
$$

10.14 b By using the alternative expression for the covariance and the rule on expectations under change of units, we find

$$
\begin{aligned}
\mathrm{Cov}(rX, tY) &= \mathrm{E}\left[(rX)(tY)\right] - \mathrm{E}\left[rX\right]\mathrm{E}\left[tY\right] \\
&= \mathrm{E}\left[rtXY\right] - (r\mathrm{E}\left[X\right])(t\mathrm{E}\left[Y\right]) \\
&= rt\mathrm{E}\left[XY\right] - rt\mathrm{E}\left[X\right]\mathrm{E}\left[Y\right] \\
&= rt\left(\mathrm{E}\left[XY\right] - \mathrm{E}\left[X\right]\mathrm{E}\left[Y\right]\right) \\
&= rt\mathrm{Cov}(X, Y).
\end{aligned}
$$

10.14 c First applying part **a** and then part **b** yields

$$
\mathrm{Cov}(rX + s, tY + u) = \mathrm{Cov}(rX, tY) = rt\mathrm{Cov}(X, Y).
$$

10.18 First note that $X_1 + X_2 + \cdots + X_N$ is the sum of all numbers, which is a nonrandom constant. Therefore, $\mathrm{Var}(X_1 + X_2 + \cdots + X_N) = 0$. In Section 9.3 we argued that, although we draw without replacement, each X_i has the same distribution. By the same reasoning, we find that each pair (X_i, X_j), with $i \neq j$, has the same joint distribution, so that $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_1, X_2)$ for all pairs with $i \neq j$. Direct application of Exercise 10.17 with $\sigma^2 = (N-1)(N+1)$ and $\gamma = \mathrm{Cov}(X_1, X_2)$ gives

$$
0 = \mathrm{Var}(X_1 + X_2 + \cdots + X_N) = N \cdot \frac{(N-1)(N+1)}{12} + N(N-1)\mathrm{Cov}(X_1, X_2).
$$

Solving this identity gives $\mathrm{Cov}(X_1, X_2) = -(N+1)/12$.

11.2 a By using the rule on addition of two independent discrete random variables, we have

$$
\mathrm{P}(X + Y = k) = p_Z(k) = \sum_{\ell=0}^{\infty} p_X(k - \ell)p_Y(\ell).
$$

Because $p_X(a) = 0$ for $a \leq -1$, all terms with $\ell \geq k+1$ vanish, so that

$$
\mathrm{P}(X + Y = k) = \sum_{\ell=0}^{k} \frac{1^{k-\ell}}{(k-\ell)!}\mathrm{e}^{-1} \cdot \frac{1^{\ell}}{\ell!}\mathrm{e}^{-1} = \frac{\mathrm{e}^{-2}}{k!}\sum_{\ell=0}^{k}\binom{k}{\ell} = \frac{2^k}{k!}\mathrm{e}^{-2},
$$

also using $\sum_{\ell=0}^{k}\binom{k}{\ell} = 2^k$ in the last equality.

11.2 b Similar to part **a**, by using the rule on addition of two independent discrete random variables and leaving out terms for which $p_X(a) = 0$, we have

$$P(X + Y = k) = \sum_{\ell=0}^{k} \frac{\lambda^{k-\ell}}{(k-\ell)!} e^{-\lambda} \cdot \frac{\mu^{\ell}}{\ell!} e^{-\mu} = \frac{(\lambda+\mu)^k}{k!} e^{-(\lambda+\mu)} \sum_{\ell=0}^{k} \binom{k}{\ell} \frac{\lambda^{k-\ell}\mu^{\ell}}{(\lambda+\mu)^k}.$$

Next, write

$$\frac{\lambda^{k-\ell}\mu^{\ell}}{(\lambda+\mu)^k} = \left(\frac{\mu}{\lambda+\mu}\right)^{\ell}\left(\frac{\lambda}{\lambda+\mu}\right)^{k-\ell} = \left(\frac{\mu}{\lambda+\mu}\right)^{\ell}\left(1 - \frac{\mu}{\lambda+\mu}\right)^{k-\ell} = p^{\ell}(1-p)^{k-\ell}$$

with $p = \mu/(\lambda+\mu)$. This means that

$$P(X + Y = k) = \frac{(\lambda+\mu)^k}{k!} e^{-(\lambda+\mu)} \sum_{\ell=0}^{k} \binom{k}{\ell} p^{\ell}(1-p)^{k-\ell} = \frac{(\lambda+\mu)^k}{k!} e^{-(\lambda+\mu)},$$

using that $\sum_{\ell=0}^{k} \binom{k}{\ell} p^{\ell}(1-p)^{k-\ell} = 1$.

11.4 a From the fact that X has an $N(2,5)$ distribution, it follows that $E[X] = 2$ and $\text{Var}(X) = 5$. Similarly, $E[Y] = 5$ and $\text{Var}(Y) = 9$. Hence by linearity of expectations,

$$E[Z] = E[3X - 2Y + 1] = 3E[X] - 2E[Y] + 1 = 3 \cdot 2 - 2 \cdot 5 + 1 = -3.$$

By the rules for the variance and covariance,

$$\text{Var}(Z) = 9\text{Var}(X) + 4\text{Var}(Y) - 12\text{Cov}(X,Y) = 9 \cdot 5 + 4 \cdot 9 - 12 \cdot 0 = 81,$$

using that $\text{Cov}(X,Y) = 0$, due to independence of X and Y.

11.4 b The random variables $3X$ and $-2Y + 1$ are independent and, according to the rule for the normal distribution under a change of units (page 106), it follows that they both have a normal distribution. Next, the sum rule for independent normal random variables then yields that $Z = (3X) + (-2Y + 1)$ also has a normal distribution. Its parameters are the expectation and variance of Z. From **a** it follows that Z has an $N(-3, 81)$ distribution.

11.4 c From **b** we know that Z has an $N(-3, 81)$ distribution, so that $(Z + 3)/9$ has a standard normal distribution. Therefore

$$P(Z \leq 6) = P\left(\frac{Z+3}{9} \leq \frac{6+3}{9}\right) = \Phi(1),$$

where Φ is the standard normal distribution function. From Table B.1 we find that $\Phi(1) = 1 - 0.1587 = 0.8413$.

11.9 a According to the product rule on page 160,

$$f_Z(z) = \int_1^z f_Y\left(\frac{z}{x}\right) f_X(x) \frac{1}{x} \, dx = \int_1^z \frac{1}{\left(\frac{z}{x}\right)^2} \frac{3}{x^4} \frac{1}{x} \, dx$$

$$= \frac{3}{z^2} \int_1^z \frac{1}{x^3} \, dx = \frac{3}{z^2}\left[-\frac{1}{2}x^{-2}\right]_1^z = \frac{3}{2}\frac{1}{z^2}\left(1 - \frac{1}{z^2}\right)$$

$$= \frac{3}{2}\left(\frac{1}{z^2} - \frac{1}{z^4}\right).$$

11.9 b According to the product rule,

$$f_Z(z) = \int_1^z f_Y\left(\frac{z}{x}\right) f_X(x)\frac{1}{x}\,\mathrm{d}x = \int_1^z \frac{\beta}{\left(\frac{z}{x}\right)^{\beta+1}} \frac{\alpha}{x^{\alpha+1}}\frac{1}{x}\,\mathrm{d}x$$

$$= \frac{\alpha\beta}{z^{\beta+1}} \int_1^z x^{\beta-\alpha-1}\,\mathrm{d}x = \frac{\alpha\beta}{z^{\beta+1}}\left[\frac{x^{\beta-\alpha}}{\beta-\alpha}\right]_1^z = \frac{\alpha\beta}{\alpha-\beta}\frac{1}{z^{\beta+1}}\left(1 - z^{\beta-\alpha}\right)$$

$$= \frac{\alpha\beta}{\beta-\alpha}\left(\frac{1}{z^{\beta+1}} - \frac{1}{z^{\alpha+1}}\right).$$

12.1 e This is certainly open to discussion. Bankruptcies: no (they come in clusters, don't they?). Eggs: no (I suppose after one egg it takes the chicken some time to produce another). Examples 3 and 4 are the best candidates. Example 5 could be modeled by the Poisson process if the crossing is not a dangerous one; otherwise authorities might take measures and destroy the homogeneity.

12.6 The expected numbers of flaws in 1 meter is $100/40 = 2.5$, and hence the number of flaws X has a $Pois(2.5)$ distribution. The answer is $\mathrm{P}(X = 2) = \frac{1}{2!}(2.5)^2\mathrm{e}^{-2.5} = 0.256$.

12.7 a It is reasonable to estimate λ with (nr. of cars)/(total time in sec.) $= 0.192$.

12.7 b $19/120 = 0.1583$, and if $\lambda = 0.192$ then $\mathrm{P}(N(10) = 0) = \mathrm{e}^{-0.192\cdot 10} = 0.147$.

12.7 c $\mathrm{P}(N(10) = 10)$ with λ from **a** seems a reasonable approximation of this probability. It equals $\mathrm{e}^{-1.92} \cdot (0.192 \cdot 10)^{10}/10! = 2.71 \cdot 10^{-5}$.

12.11 Following the hint, we obtain:

$$\mathrm{P}(N([0, s] = k, N([0, 2s]) = n) = \mathrm{P}(N([0, s]) = k, N((s, 2s]) = n - k)$$
$$= \mathrm{P}(N([0, s]) = k) \cdot \mathrm{P}(N((s, 2s]) = n - k)$$
$$= (\lambda s)^k\mathrm{e}^{-\lambda s}/(k!) \cdot (\lambda s)^{n-k}\mathrm{e}^{-\lambda s}/((n-k)!)$$
$$= (\lambda s)^n\mathrm{e}^{-\lambda 2s}/(k!(n-k)!).$$

So

$$\mathrm{P}(N([0, s]) = k \mid N([0, 2s]) = n) = \frac{\mathrm{P}(N([0, s]) = k, N([0, 2s]) = n)}{\mathrm{P}(N([0, 2s]) = n)}$$
$$= n!/(k!(n-k)!) \cdot (\lambda s)^n/(2\lambda s)^n$$
$$= n!/(k!(n-k)!) \cdot (1/2)^n.$$

This holds for $k = 0, \ldots, n$, so we find the $Bin(n, \frac{1}{2})$ distribution.

13.2 a From the formulas for the $U(a, b)$ distribution, substituting $a = -1/2$ and $b = 1/2$, we derive that $\mathrm{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = 1/12$.

13.2 b We write $S = X_1 + X_2 + \cdots + X_{100}$, for which we find $\mathrm{E}[S] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_{100}] = 0$ and, by independence, $\mathrm{Var}(S) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_{100}) = 100\cdot\frac{1}{12} = 100/12$. We find from Chebyshev's inequality:

$$\mathrm{P}(|S| > 10) = \mathrm{P}(|S - 0| > 10) \le \frac{\mathrm{Var}(S)}{10^2} = \frac{1}{12}.$$

13.4 a Because X_i has a $Ber(p)$ distribution, $\mathrm{E}\,[X_i] = p$ and $\mathrm{Var}(X_i) = p\,(1-p)$, and so $\mathrm{E}\,[\bar{X}_n] = p$ and $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_i)\,/n = p\,(1-p)/n$. By Chebyshev's inequality:

$$\mathrm{P}\big(|\bar{X}_n - p| \ge 0.2\big) \le \frac{p\,(1-p)/n}{(0.2)^2} = \frac{25p(1-p)}{n}.$$

The right-hand side should be at most 0.1 (note that we switched to the complement). If $p = 1/2$ we therefore require $25/(4n) \le 0.1$, or $n \ge 25/(4 \cdot 0.1) = 62.5$, i.e., $n \ge 63$. Now, suppose $p \ne 1/2$, using $n = 63$ and $p(1-p) \le 1/4$ we conclude that $25p(1-p)/n \le 25 \cdot (1/4)/63 = 0.0992 < 0.1$, so (because of the inequality) the computed value satisfies for other values of p as well.

13.4 b For arbitrary $a > 0$ we conclude from Chebyshev's inequality:

$$\mathrm{P}\big(|\bar{X}_n - p| \ge a\big) \le \frac{p\,(1-p)/n}{a^2} = \frac{p(1-p)}{na^2} \le \frac{1}{4na^2},$$

where we used $p\,(1-p) \le 1/4$ again. The question now becomes: when $a = 0.1$, for what n is $1/(4na^2) \le 0.1$? We find: $n \ge 1/(4 \cdot 0.1 \cdot (0.1)^2) = 250$, so $n = 250$ is large enough.

13.4 c From part **a** we know that an error of size 0.2 or occur with a probability of at most $25/4n$, regardless of the values of p. So, we need $25/(4n) \le 0.05$, i.e., $n \ge 25/(4 \cdot 0.05) = 125$.

13.4 d We compute $\mathrm{P}\big(\bar{X}_n \le 0.5\big)$ for the case that $p = 0.6$. Then $\mathrm{E}\,[\bar{X}_n] = 0.6$ and $\mathrm{Var}(\bar{X}_n) = 0.6 \cdot 0.4/n$. Chebyshev's inequality cannot be used directly, we need an intermediate step: the probability that $\bar{X}_n \le 0.5$ is contained in the event "the prediction is off by at least 0.1, in either direction." So

$$\mathrm{P}\big(\bar{X}_n \le 0.5\big) \le \mathrm{P}\big(|\bar{X}_n - 0.6| \ge 0.1\big) \le \frac{0.6 \cdot 0.4/n}{(0.1)^2} = \frac{24}{n}$$

For $n \ge 240$ this probability is 0.1 or smaller.

13.9 a The statement looks like the law of large numbers, and indeed, if we look more closely, we see that T_n is the average of an i.i.d. sequence: define $Y_i = X_i^2$, then $T_n = \bar{Y}_n$. The law of large numbers now states: if $\bar{Y}_n$ is the average of n independent random variables with expectation μ and variance σ^2, then for any $\varepsilon > 0$: $\lim_{n\to\infty} \mathrm{P}\big(|\bar{Y}_n - \mu| > \varepsilon\big) = 0$. So, if $a = \mu$ and the variance σ^2 is finite, then it is true.

13.9 b We compute expectation and variance of Y_i: $\mathrm{E}\,[Y_i] = \mathrm{E}\,[X_i^2] = \int_{-1}^{1} \frac{1}{2}x^2\,\mathrm{d}x = 1/3$. And: $\mathrm{E}\,[Y_i^2] = \mathrm{E}\,[X_i^4] = \int_{-1}^{1} \frac{1}{2}x^4\,\mathrm{d}x = 1/5$, so $\mathrm{Var}(Y_i) = 1/5 - (1/3)^2 = 4/45$. The variance is finite, so indeed, the law of large numbers applies, and the statement is true if $a = \mathrm{E}\,[X_i^2] = 1/3$.

14.3 First note that $\mathrm{P}\big(|\bar{X}_n - p| < 0.2\big) = 1 - \mathrm{P}\big(\bar{X}_n - p \ge 0.2\big) - \mathrm{P}\big(\bar{X}_n - p \le -0.2\big)$. Because $\mu = p$ and $\sigma^2 = p(1-p)$, we find, using the central limit theorem:

$$\mathrm{P}\big(\bar{X}_n - p \ge 0.2\big) = \mathrm{P}\left(\sqrt{n}\frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \ge \sqrt{n}\frac{0.2}{\sqrt{p(1-p)}}\right)$$

$$= \mathrm{P}\left(Z_n \ge \sqrt{n}\frac{0.2}{\sqrt{p(1-p)}}\right) \approx \mathrm{P}\left(Z \ge \sqrt{n}\frac{0.2}{\sqrt{p(1-p)}}\right),$$

where Z has an $N(0,1)$ distribution. Similarly,

$$P(\bar{X}_n - p \le -0.2) \approx P\left(Z \ge \sqrt{n}\frac{0.2}{\sqrt{p(1-p)}}\right),$$

so we are looking for the smallest positive integer n such that

$$1 - 2P\left(Z \ge \sqrt{n}\frac{0.2}{\sqrt{p(1-p)}}\right) \ge 0.9,$$

i.e., the smallest positive integer n such that

$$P\left(Z \ge \sqrt{n}\frac{0.2}{\sqrt{p(1-p)}}\right) \le 0.05.$$

From Table B.1 it follows that

$$\sqrt{n}\frac{0.2}{\sqrt{p(1-p)}} \ge 1.645.$$

Since $p(1-p) \le 1/4$ for all p between 0 and 1, we see that n should be at least 17.

14.5 In Section 4.3 we have seen that X has the same probability distribution as $X_1 + X_2 + \cdots + X_n$, where $X_1, X_2, \ldots, X_n$ are independent $Ber(p)$ distributed random variables. Recall that $E[X_i] = p$, and $\text{Var}(X_i) = p(1-p)$. But then we have for any real number a that

$$P\left(\frac{X - np}{\sqrt{np(1-p)}} \le a\right) = P\left(\frac{X_1 + X_2 + \cdots + X_n - np}{\sqrt{np(1-p)}} \le a\right) = P(Z_n \le a);$$

see also (14.1). It follows from the central limit theorem that

$$P\left(\frac{X - np}{\sqrt{np(1-p)}} \le a\right) \approx \Phi(a),$$

i.e., the random variable $\frac{X-np}{\sqrt{np(1-p)}}$ has a distribution that is approximately standard normal.

14.9 a The probability that for a chain of at least 50 meters more than 1002 links are needed is the same as the probability that a chain of 1002 chains is shorter than 50 meters. Assuming that the random variables $X_1, X_2, \ldots, X_{1002}$ are independent, and using the central limit theorem, we have that

$$P(X_1 + X_2 + \cdots + X_{1002} < 5000) \approx P\left(Z < \sqrt{1002} \cdot \frac{\frac{5000}{1002} - 5}{\sqrt{0.04}}\right) = 0.0571,$$

where Z has an $N(0,1)$ distribution. So about 6% of the customers will receive a free chain.

14.9 b We now have that

$$P(X_1 + X_2 + \cdots + X_{1002} < 5000) \approx P(Z < 0.0032),$$

which is slightly larger than $1/2$. So about half of the customers will receive a free chain. Clearly something has to be done: a seemingly minor change of expected value has major consequences!

15.6 Because $(2-0) \cdot 0.245 + (4-2) \cdot 0.130 + (7-4) \cdot 0.050 + (11-7) \cdot 0.020 + (15-11) \cdot 0.005 = 1$, there are no data points outside the listed bins. Hence

$$
\begin{aligned}
F_n(7) &= \frac{\text{number of } x_i \leq 7}{n} \\
&= \frac{\text{number of } x_i \text{ in bins } (0, 2], (2, 4] \text{ and } (4, 7]}{n} \\
&= \frac{n \cdot (2-0) \cdot 0.245 + n \cdot (4-2) \cdot 0.130 + n \cdot (7-4) \cdot 0.050}{n} \\
&= 0.490 + 0.260 + 0.150 = 0.9.
\end{aligned}
$$

15.11 The height of the histogram on a bin $(a, b]$ is

$$
\begin{aligned}
\frac{\text{number of } x_i \text{ in } (a, b]}{n(b-a)} &= \frac{(\text{number of } x_i \leq b) - (\text{number of } x_i \leq a)}{n(b-a)} \\
&= \frac{F_n(b) - F_n(a)}{b-a}.
\end{aligned}
$$

15.12 a By inserting the expression for $f_{n,h}(t)$, we get

$$
\begin{aligned}
\int_{-\infty}^{\infty} t \cdot f_{n,h}(t) \, \mathrm{d}t &= \int_{-\infty}^{\infty} t \cdot \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t-x_i}{h}\right) \mathrm{d}t \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{t}{h} K\left(\frac{t-x_i}{h}\right) \mathrm{d}t.
\end{aligned}
$$

For each i fixed we find with change of integration variables $u = (t - x_i)/h$,

$$
\begin{aligned}
\int_{-\infty}^{\infty} \frac{t}{h} K\left(\frac{t-x_i}{h}\right) \mathrm{d}t &= \int_{-\infty}^{\infty} (x_i + hu) K(u) \, \mathrm{d}u \\
&= x_i \int_{-\infty}^{\infty} K(u) \, \mathrm{d}u + h \int_{-\infty}^{\infty} uK(u) \, \mathrm{d}u = x_i,
\end{aligned}
$$

using that K integrates to one and that $\int_{-\infty}^{\infty} uK(u) \, \mathrm{d}u = 0$, because K is symmetric. Hence

$$
\int_{-\infty}^{\infty} t \cdot f_{n,h}(t) \, \mathrm{d}t = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{t}{h} K\left(\frac{t-x_i}{h}\right) \mathrm{d}t = \frac{1}{n} \sum_{i=1}^{n} x_i.
$$

15.12 b By means of similar reasoning

$$
\begin{aligned}
\int_{-\infty}^{\infty} t^2 \cdot f_{n,h}(t) \, \mathrm{d}t &= \int_{-\infty}^{\infty} t^2 \cdot \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t-x_i}{h}\right) \mathrm{d}t \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{t^2}{h} K\left(\frac{t-x_i}{h}\right) \mathrm{d}t.
\end{aligned}
$$

For each i:

$$\int_{-\infty}^{\infty} \frac{t^2}{h} K\left(\frac{t - x_i}{h}\right) \, dt$$

$$= \int_{-\infty}^{\infty} (x_i + hu)^2 K(u) \, du = \int_{-\infty}^{\infty} (x_i^2 + 2x_i hu + h^2 u^2) K(u) \, du$$

$$= x_i^2 \int_{-\infty}^{\infty} K(u) \, du + 2x_i h \int_{-\infty}^{\infty} u K(u) \, du + h^2 \int_{-\infty}^{\infty} u^2 K(u) \, du$$

$$= x_i^2 + h^2 \int_{-\infty}^{\infty} u^2 K(u) \, du,$$

again using that K integrates to one and that K is symmetric.

16.3 a Because $n = 24$, the sample median is the average of the 12th and 13th elements. Since these are both equal to 70, the sample median is also 70. The lower quartile is the pth empirical quantile for $p = 1/4$. We get $k = \lfloor p(n + 1) \rfloor = 6$, so that

$$q_n(0.25) = x_{(6)} + 0.25 \cdot (x_{(7)} - x_{(6)}) = 66 + 0.25 \cdot (67 - 66) = 66.25.$$

Similarly, the upper quartile is the pth empirical quantile for $p = 3/4$:

$$q_n(0.75) = x_{(18)} + 0.75 \cdot (x_{(19)} - x_{(18)}) = 75 + 0.75 \cdot (75 - 75) = 75.$$

16.3 b In part **a** we found the sample median and the two quartiles. From this we compute the IQR: $q_n(0.75) - q_n(0.25) = 75 - 66.25 = 8.75$. This means that

$$q_n(0.25) - 1.5 \cdot \text{IQR} = 66.25 - 1.5 \cdot 8.75 = 53.125,$$
$$q_n(0.75) + 1.5 \cdot \text{IQR} = 75 + 1.5 \cdot 8.75 = 88.125.$$

Hence, the last element below 88.125 is 88, and the first element above 53.125 is 57. Therefore, the upper whisker runs until 88 and the lower whisker until 57, with two elements 53 and 31 below. This leads to the following boxplot:
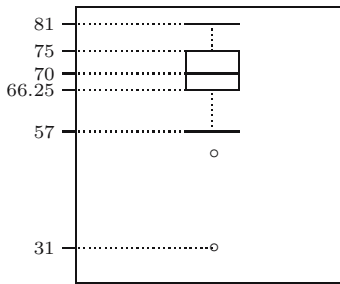
16.3 c The values 53 and 31 are outliers. Value 31 is far away from the bulk of the data and appears to be an *extreme* outlier.

16.6 a Yes, we find $\bar{x} = (1 + 5 + 9)/3 = 15/3 = 5$, $\bar{y} = (2 + 4 + 6 + 8)/4 = 20/4 = 5$, so that $(\bar{x} + \bar{y})/2 = 5$. The average for the combined dataset is also equal to 5: $(15 + 20)/7 = 5$.

16.6 b The mean of $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m$ equals

$$\frac{x_1 + \cdots + x_n + y_1 + \cdots + y_m}{n + m} = \frac{n\bar{x}_n + m\bar{y}_m}{n + m} = \frac{n}{n + m} \bar{x}_n + \frac{m}{n + m} \bar{y}_m.$$

In general, this is not equal to $(\bar{x}_n + \bar{y}_m)/2$. For instance, replace 1 in the first dataset by 4. Then $\bar{x}_n = 6$ and $\bar{y}_m = 5$, so that $(\bar{x}_n + \bar{y}_m)/2 = 5\frac{1}{2}$. However, the average of the combined dataset is $38/7 = 5\frac{2}{7}$.

16.6 c Yes, $m = n$ implies $n/(n + m) = m/(n + m) = 1/2$. From the expressions found in part **b** we see that the sample mean of the combined dataset equals $(\bar{x}_n + \bar{y}_m)/2$.

16.8 The ordered combined dataset is 1, 2, 4, 5, 6, 8, 9, so that the sample median equals 5. The absolute deviations from 5 are: 4, 3, 1, 0, 1, 3, 4, and if we put them in order: 0, 1, 1, 3, 3, 4, 4. The MAD is the sample median of the absolute deviations, which is 3.

16.15 First write

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2\bar{x}_n x_i + \bar{x}_n^2\right) = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}_n\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}_n^2.$$

Next, by inserting

$$\frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}_n \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\bar{x}_n^2 = \frac{1}{n}\cdot n \cdot \bar{x}_n^2 = \bar{x}_n^2,$$

we find

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}_n^2 + \bar{x}_n^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}_n^2.$$

17.3 a The model distribution corresponds to the number of women in a queue. A queue has 10 positions. The occurrence of a woman in any position is independent of the occurrence of a woman in other positions. At each position a woman occurs with probability p. Counting the occurrence of a woman as a "success," the number of women in a queue corresponds to the number of successes in 10 independent experiments with probability p of success and is therefore modeled by a $Bin(10, p)$ distribution.

17.3 b We have 100 queues and the number of women x_i in the ith queue is a realization of a $Bin(10, p)$ random variable. Hence, according to Table 17.2, the average number of women $\bar{x}_{100}$ resembles the expectation $10p$ of the $Bin(10, p)$ distribution. We find $\bar{x}_{100} = 435/100 = 4.35$, so an estimate for p is $4.35/10 = 0.435$.

17.7 a If we model the series of disasters by a Poisson process, then as a property of the Poisson process, the interdisaster times should follow an exponential distribution (see Section 12.3). This is indeed confirmed by the histogram and empirical distribution of the observed interdisaster times; they resemble the probability density and distribution function of an exponential distribution.

17.7 b The average length of a time interval is $40\,549/190 = 213.4$ days. Following Table 17.2 this should resemble the expectation of the $Exp(\lambda)$ distribution, which is $1/\lambda$. Hence, as an estimate for λ we could take $190/40\,549 = 0.00469$.

17.9 a A (perfect) cylindrical cone with diameter d (at the base) and height h has volume $\pi d^2 h/12$, or about $0.26d^2 h$. The effective wood of a tree is the trunk without the branches. Since the trunk is similar to a cylindrical cone, one can expect a linear relation between the effective wood and $d^2 h$.

17.9 b We find

$$\bar{z}_n = \frac{\sum y_i/x_i}{n} = \frac{9.369}{31} = 0.3022$$

$$\bar{y}/\bar{x} = \frac{(\sum y_i)/n}{(\sum x_i)/n} = \frac{26.486/31}{87.456/31} = 0.3028$$

$$\text{least squares} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{95.498}{314.644} = 0.3035.$$

18.3 a Note that generating from the empirical distribution function is the same as choosing one of the elements of the original dataset with equal probability. Hence, an element in the bootstrap dataset equals 0.35 with probability 0.1. The number of ways to have exactly three out of ten elements equal to 0.35 is $\binom{10}{3}$, and each has probability $(0.1)^3(0.9)^7$. Therefore, the probability that the bootstrap dataset has exactly three elements equal to 0.35 is equal to $\binom{10}{3}(0.1)^3(0.9)^7 = 0.0574$.

18.3 b Having at most two elements less than or equal to 0.38 means that 0, 1, or 2 elements are less than or equal to 0.38. Five elements of the original dataset are smaller than or equal to 0.38, so that an element in the bootstrap dataset is less than or equal to 0.38 with probability 0.5. Hence, the probability that the bootstrap dataset has at most two elements less than or equal to 0.38 is equal to $(0.5)^{10} + \binom{10}{1}(0.5)^{10} + \binom{10}{2}(0.5)^{10} = 0.0547$.

18.3 c Five elements of the dataset are smaller than or equal to 0.38 and two are greater than 0.42. Therefore, obtaining a bootstrap dataset with two elements less than or equal to 0.38, and the other elements greater than 0.42 has probability $(0.5)^2 (0.2)^8$. The number of such bootstrap datasets is $\binom{10}{2}$. So the answer is $\binom{10}{2}(0.5)^2 (0.2)^8 = 0.000029$.

18.7 For the parametric bootstrap, we must estimate the parameter θ by $\hat{\theta} = (n + 1)m_n/n$, and generate bootstrap samples from the $U(0,\hat{\theta})$ distribution. This distribution has expectation $\mu_{\hat{\theta}} = \hat{\theta}/2 = (n+1)m_n/(2n)$. Hence, for each bootstrap sample $x_1^*, x_2^*, \ldots, x_n^*$ compute $\bar{x}_n^* - \mu_{\hat{\theta}} = \bar{x}_n^* - (n+1)m_n/(2n)$.
Note that this is different from the *empirical* bootstrap simulation, where one would estimate μ by $\bar{x}_n$ and compute $\bar{x}_n^* - \bar{x}_n$.

18.8 a Since we know nothing about the distribution of the interfailure times, we estimate F by the empirical distribution function F_n of the software data and we estimate the expectation μ of F by the expectation $\mu^* = \bar{x}_n = 656.8815$ of F_n. The bootstrapped centered sample mean is the random variable $\bar{X}_n^* - 656.8815$. The corresponding empirical bootstrap simulation is described as follows:

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n, i.e., draw with replacement 135 numbers from the software data.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - 656.8815$$

where $\bar{x}_n$ is the sample mean of $x_1^*, x_2^*, \ldots, x_n^*$.

Repeat steps 1 and 2 one thousand times.

18.8 b Because the interfailure times are now assumed to have an $Exp(\lambda)$ distribution, we must estimate λ by $\hat{\lambda} = 1/\bar{x}_n = 0.0015$ and estimate F by the distribution

function of the $Exp(0.0015)$ distribution. Estimate the expectation $\mu = 1/\lambda$ of the $Exp(\lambda)$ distribution by $\mu^* = 1/\hat{\lambda} = \bar{x}_n = 656.8815$. Also now, the bootstrapped centered sample mean is the random variable $\bar{X}_n^* - 656.8815$. The corresponding parametric bootstrap simulation is described as follows:

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from the $Exp(0.0015)$ distribution.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - 656.8815,$$

where $\bar{x}_n$ is the sample mean of $x_1^*, x_2^*, \ldots, x_n^*$.

Repeat steps 1 and 2 one thousand times. We see that in this simulation the bootstrapped centered sample mean is the *same* in both cases: $\bar{X}_n^* - \bar{x}_n$, but the corresponding simulation procedures differ in step 1.

18.8 c Estimate λ by $\hat{\lambda} = \ln 2/m_n = 0.0024$ and estimate F by the distribution function of the $Exp(0.0024)$ distribution. Estimate the expectation $\mu = 1/\lambda$ of the $Exp(\lambda)$ distribution by $\mu^* = 1/\hat{\lambda} = 418.3816$. The corresponding parametric bootstrap simulation is described as follows:

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from the $Exp(0.0024)$ distribution.
2. Compute the centered sample mean for the bootstrap dataset:

$$\bar{x}_n^* - 418.3816,$$

where $\bar{x}_n$ is the sample mean of $x_1^*, x_2^*, \ldots, x_n^*$.

Repeat steps 1 and 2 one thousand times. We see that in this parametric bootstrap simulation the bootstrapped centered sample mean is *different* from the one in the empirical bootstrap simulation: $\bar{X}_n^* - (\ln 2)/m_n$ instead of $\bar{X}_n^* - \bar{x}_n$.

19.1 a From the formulas for the expectation and variance of uniform random variables we know that $\mathrm{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = (2\theta)^2/12 = \theta^2/3$. Hence $\mathrm{E}[X_i^2] = \mathrm{Var}(X_i) + (\mathrm{E}[X_i])^2 = \theta^2/3$. Therefore, by linearity of expectations

$$\mathrm{E}[T] = \frac{3}{n}\left(\frac{\theta^2}{3} + \cdots + \frac{\theta^2}{3}\right) = \frac{3}{n} \cdot n \cdot \frac{\theta^2}{3} = \theta^2.$$

Since $\mathrm{E}[T] = \theta^2$, the random variable T is an unbiased estimator for θ^2.

19.1 b The function $g(x) = -\sqrt{x}$ is a strictly convex function, because $g''(x) = (x^{-3/4})/4 > 0$. Therefore, by Jensen's inequality, $-\sqrt{\mathrm{E}[T]} < -\mathrm{E}\left[\sqrt{T}\right]$. Since, from part **a** we know that $\mathrm{E}[T] = \theta^2$, this means that $\mathrm{E}\left[\sqrt{T}\right] < \theta$. In other words, $\sqrt{T}$ is a biased estimator for θ, with negative bias.

19.8 From the model assumptions it follows that $\mathrm{E}[Y_i] = \beta x_i$ for each i. Using linearity of expectations, this implies that

$$\mathrm{E}[B_1] = \frac{1}{n}\left(\frac{\mathrm{E}[Y_1]}{x_1} + \cdots + \frac{\mathrm{E}[Y_n]}{x_n}\right) = \frac{1}{n}\left(\frac{\beta x_1}{x_1} + \cdots + \frac{\beta x_n}{x_n}\right) = \beta,$$

$$\mathrm{E}[B_2] = \frac{\mathrm{E}[Y_1] + \cdots + \mathrm{E}[Y_n]}{x_1 + \cdots + x_n} = \frac{\beta x_1 + \cdots + \beta x_n}{x_1 + \cdots + x_n} = \beta,$$

$$\mathrm{E}[B_3] = \frac{x_1\mathrm{E}[Y_1] + \cdots + x_n\mathrm{E}[Y_n]}{x_1^2 + \cdots + x_n^2} = \frac{\beta x_1^2 + \cdots + \beta x_n^2}{x_1^2 + \cdots + x_n^2} = \beta.$$

20.2 a Compute the mean squared errors of S and T: $\text{MSE}(S) = \text{Var}(S) + [\text{bias}(S)]^2 = 40 + 0 = 40$; $\text{MSE}(T) = \text{Var}(T) + [\text{bias}(T)]^2 = 4 + 9 = 13$. We prefer T, because it has a smaller MSE.

20.2 b Compute the mean squared errors of S and T: $\text{MSE}(S) = 40$, as in **a**; $\text{MSE}(T) = \text{Var}(T) + [\text{bias}(T)]^2 = 4 + a^2$. So, if $a < 6$: prefer T. If $a \geq 6$: prefer S. The preferences are based on the MSE criterion.

20.3 $\text{Var}(T_1) = 1/(n\lambda^2)$, $\text{Var}(T_2) = 1/\lambda^2$; hence we prefer T_1, because of its smaller variance.

20.8 a This follows directly from linearity of expectations:

$$\text{E}[T] = \text{E}\left[r\bar{X}_n + (1-r)\bar{Y}_m\right] = r\text{E}\left[\bar{X}_n\right] + (1-r)\text{E}\left[\bar{Y}_m\right] = r\mu + (1-r)\mu = \mu.$$

20.8 b Using that $\bar{X}_n$ and $\bar{Y}_m$ are independent, we find $\text{MSE}(T) = \text{Var}(T) = r^2\text{Var}\left(\bar{X}_n\right) + (1-r)^2\text{Var}\left(\bar{Y}_m\right) = r^2 \cdot \sigma^2/n + (1-r)^2 \cdot \sigma^2/m$.
To find the minimum of this parabola we differentiate with respect to r and equate the result to 0: $2r/n - 2(1-r)/m = 0$. This gives the minimum value: $2rm - 2n(1-r) = 0$ or $r = n/(n+m)$.

21.1 Setting $X_i = j$ if *red* appears in the ith experiment for the first time on the jth throw, we have that X_1, X_2, and X_3 are independent $Geo\,(p)$ distributed random variables, where p is the probability that *red* appears when throwing the selected die. The likelihood function is

$$L(p) = \text{P}(X_1 = 3, X_2 = 5, X_3 = 4) = (1-p)^2p \cdot (1-p)^4p \cdot (1-p)^3p$$
$$= p^3(1-p)^9,$$

so for D_1 one has that $L(p) = L(\frac{5}{6}) = \left(\frac{5}{6}\right)^3\left(1 - \frac{5}{6}\right)^9$, whereas for D_2 one has that $L(p) = L(\frac{1}{6}) = \left(\frac{1}{6}\right)^3\left(1 - \frac{1}{6}\right)^9 = 5^6 \cdot L(\frac{5}{6})$. It is very likely that we picked D_2.

21.4 a The likelihood $L(\mu)$ is given by

$$L(\mu) = \text{P}(X_1 = x_1, \dots, X_n = x_n) = \text{P}(X_1 = x_1) \cdots \text{P}(X_n = x_n)$$
$$= \frac{\mu^{x_1}}{x_1!} \cdot \text{e}^{-\mu} \cdots \frac{\mu^{x_n}}{x_n!} \cdot \text{e}^{-\mu} = \frac{\text{e}^{-n\mu}}{x_1! \cdots x_n!}\mu^{x_1 + x_2 + \cdots + x_n}.$$

21.4 b We find that the loglikelihood $\ell(\mu)$ is given by

$$\ell(\mu) = \left(\sum_{i=1}^{n} x_i\right)\ln(\mu) - \ln(x_1! \cdots x_n!) - n\mu.$$

Hence

$$\frac{\text{d}\ell}{\text{d}\mu} = \frac{\sum x_i}{\mu} - n,$$

and we find—after checking that we indeed have a maximum!—that $\bar{x}_n$ is the maximum likelihood estimate for μ.

21.4 c In **b** we have seen that $\bar{x}_n$ is the maximum likelihood estimate for μ. Due to the invariance principle from Section 21.4 we thus find that $\text{e}^{-\bar{x}_n}$ is the maximum likelihood estimate for $\text{e}^{-\mu}$.

21.8 a The likelihood $L(\theta)$ is given by

$$L(\theta) = C \cdot \left(\frac{1}{4}(2+\theta)\right)^{1997} \cdot \left(\frac{1}{4}\theta\right)^{32} \cdot \left(\frac{1}{4}(1-\theta)\right)^{906} \cdot \left(\frac{1}{4}(1-\theta)\right)^{904}$$

$$= \frac{C}{4^{3839}} \cdot (2+\theta)^{1997} \cdot \theta^{32} \cdot (1-\theta)^{1810},$$

where C is the number of ways we can assign 1997 starchy-greens, 32 sugary-whites, 906 starchy-whites, and 904 sugary-greens to 3839 plants. Hence the loglikelihood $\ell(\theta)$ is given by

$$\ell(\theta) = \ln(C) - 3839\ln(4) + 1997\ln(2+\theta) + 32\ln(\theta) + 1810\ln(1-\theta).$$

21.8 b A short calculation shows that

$$\frac{d\ell(\theta)}{d\theta} = 0 \qquad \Leftrightarrow \qquad 38100\theta^2 - 1655\theta - 64 = 0,$$

so the maximum likelihood estimate of θ is (after checking that $L(\theta)$ indeed attains a maximum for this value of θ):

$$\frac{-1655 + \sqrt{3714385}}{7620} = 0.0357.$$

21.8 c In this general case the likelihood $L(\theta)$ is given by

$$L(\theta) = C \cdot \left(\frac{1}{4}(2+\theta)\right)^{n_1} \cdot \left(\frac{1}{4}\theta\right)^{n_2} \cdot \left(\frac{1}{4}(1-\theta)\right)^{n_3} \cdot \left(\frac{1}{4}(1-\theta)\right)^{n_4} \cdot$$

$$= \frac{C}{4^n} \cdot (2+\theta)^{n_1} \cdot \theta^{n_2} \cdot (1-\theta)^{n_3+n_4},$$

where C is the number of ways we can assign n_1 starchy-greens, n_2 sugary-whites, n_3 starchy-whites, and n_4 sugary-greens to n plants. Hence the loglikelihood $\ell(\theta)$ is given by

$$\ell(\theta) = \ln(C) - n\ln(4) + n_1\ln(2+\theta) + n_2\ln(\theta) + (n_3+n_4)\ln(1-\theta).$$

A short calculation shows that

$$\frac{d\ell(\theta)}{d\theta} = 0 \qquad \Leftrightarrow \qquad n\theta^2 - (n_1 - n_2 - 2n_3 - 2n_4)\theta - 2n_2 = 0,$$

so the maximum likelihood estimate of θ is (after checking that $L(\theta)$ indeed attains a maximum for this value of θ):

$$\frac{n_1 - n_2 - 2n_3 - 2n_4 + \sqrt{(n_1 - n_2 - 2n_3 - 2n_4)^2 + 8nn_2}}{2n}.$$

21.11 a Since the dataset is a realization of a random sample from a $Geo(1/N)$ distribution, the likelihood is $L(N) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$, where each X_i has a $Geo(1/N)$ distribution. So

$$L(N) = \left(1 - \frac{1}{N}\right)^{x_1-1}\frac{1}{N}\left(1 - \frac{1}{N}\right)^{x_2-1}\frac{1}{N}\cdots\left(1 - \frac{1}{N}\right)^{x_n-1}\frac{1}{N}$$

$$= \left(1 - \frac{1}{N}\right)^{\left(-n+\sum_{i=1}^{n} x_i\right)}\left(\frac{1}{N}\right)^n.$$

But then the loglikelihood is equal to

$$\ell(N) = -n \ln N + \left(-n + \sum_{i=1}^{n} x_i \right) \ln \left(1 - \frac{1}{N} \right).$$

Differentiating to N yields

$$\frac{\mathrm{d}}{\mathrm{d}N} \left(\ell(N) \right) = \frac{-n}{N} + \left(-n + \sum_{i=1}^{n} x_i \right) \frac{1}{N(N-1)},$$

Now $\frac{\mathrm{d}}{\mathrm{d}N} \left(\ell(N) \right) = 0$ if and only if $N = \bar{x}_n$. Because $\ell(N)$ attains its maximum at $\bar{x}_n$, we find that the maximum likelihood estimate of N is $\hat{N} = \bar{x}_n$.

21.11 b Since $P(Y = k) = 1/N$ for $k = 1, 2, \ldots, N$, the likelihood is given by

$$L(N) = \left(\frac{1}{N} \right)^n \quad \text{for } N \geq y_{(n)},$$

and $L(N) = 0$ for $N < y_{(n)}$. So $L(N)$ attains its maximum at $y_{(n)}$; the maximum likelihood estimate of N is $\hat{N} = y_{(n)}$.

22.1 a Since $\sum x_i y_i = 12.4$, $\sum x_i = 9$, $\sum y_i = 4.8$, $\sum x_i^2 = 35$, and $n = 3$, we find (c.f. (22.1) and (22.2)), that

$$\hat{\beta} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} = \frac{3 \cdot 12.4 - 9 \cdot 4.8}{3 \cdot 35 - 9^2} = -\frac{1}{4},$$

and $\hat{\alpha} = \bar{y}_n - \hat{\beta} \bar{x}_n = 2.35$.

22.1 b Since $r_i = y_i - \hat{\alpha} - \hat{\beta} x_i$, for $i = 1, \ldots, n$, we find $r_1 = 2 - 2.35 + 0.25 = -0.1$, $r_2 = 1.8 - 2.35 + 0.75 = 0.2$, $r_3 = 1 - 2.35 + 1.25 = -0.1$, and $r_1 + r_2 + r_3 = -0.1 + 0.2 - 0.1 = 0$.

22.1 c See Figure D.1.
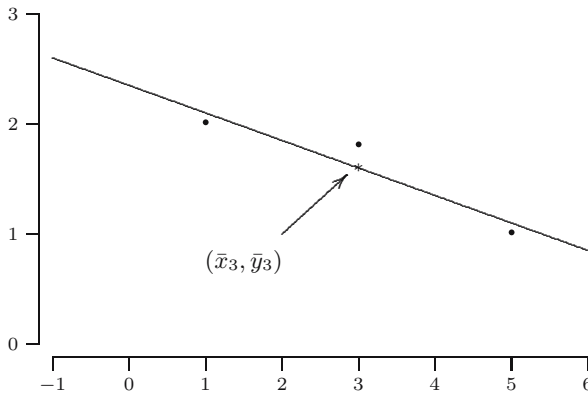
Fig. D.1. Solution of Exercise 22.1 **c**.

22.5 With the assumption that $\alpha = 0$, the method of least squares tells us now to minimize

$$S(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2 \, .$$

Now

$$\frac{\mathrm{d}S(\beta)}{\mathrm{d}\beta} = -2 \sum_{i=1}^{n} (y_i - \beta x_i) x_i = -2 \left(\sum_{i=1}^{n} x_i y_i - \beta \sum_{i=1}^{n} x_i^2 \right),$$

so

$$\frac{\mathrm{d}S(\beta)}{\mathrm{d}\beta} = 0 \quad \Leftrightarrow \quad \beta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Because $S(\beta)$ has a minimum for this last value of β, we see that the least squares estimator $\hat{\beta}$ of β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}.$$

22.12 a Since the denominator of $\hat{\beta}$ is a number, *not* a random variable, one has that

$$\mathrm{E}\left[\hat{\beta}\right] = \frac{\mathrm{E}\left[n(\sum x_i Y_i) - (\sum x_i)(\sum Y_i)\right]}{x \sum x_i^2 - (\sum x_i)^2}.$$

Furthermore, the numerator of this last fraction can be written as

$$\mathrm{E}\left[n \sum x_i Y_i\right] - \mathrm{E}\left[(\sum x_i)(\sum Y_i)\right],$$

which is equal to

$$n \sum (x_i \mathrm{E}\left[Y_i\right]) - (\sum x_i) \sum \mathrm{E}\left[Y_i\right].$$

22.12 b Substituting $\mathrm{E}\left[Y_i\right] = \alpha + \beta x_i$ in the last expression, we find that

$$\mathrm{E}\left[\hat{\beta}\right] = \frac{n \sum (x_i(\alpha + \beta x_i)) - (\sum x_i)\left[\sum(\alpha + \beta x_i)\right]}{x \sum x_i^2 - (\sum x_i)^2}.$$

22.12 c The numerator of the previous expression for $\mathrm{E}\left[\hat{\beta}\right]$ can be simplified to

$$\frac{n\alpha \sum x_i + n\beta \sum x_i^2 - n\alpha \sum x_i - \beta(\sum x_i)(\sum x_i)}{n \sum x_i^2 - (\sum x_i)^2},$$

which is equal to

$$\frac{\beta(n \sum x_i^2 - (\sum x_i)^2)}{n \sum x_i^2 - (\sum x_i)^2}.$$

22.12 d From **c** it now follows that $\mathrm{E}\left[\hat{\beta}\right] = \beta$, i.e., $\hat{\beta}$ is an unbiased estimator for β.

23.5 a The standard confidence interval for the mean of a normal sample with unknown variance applies, with $n = 23$, $\bar{x} = 0.82$ and $s = 1.78$, so:

$$\left(\bar{x} - t_{22,0.025} \cdot \frac{s}{\sqrt{23}}, \ \bar{x} + t_{22,0.025} \cdot \frac{s}{\sqrt{23}} \right).$$

The critical values come from the $t(22)$ distribution: $t_{22,0.025} = 2.074$. The actual interval becomes:

$$\left(0.82 - 2.074 \cdot \frac{1.78}{\sqrt{23}}, \ 0.82 + 2.074 \cdot \frac{1.78}{\sqrt{23}} \right) = (0.050, \ 1.590).$$

23.5 b Generate one thousand samples of size 23, by drawing with replacement from the 23 numbers

$$1.06, \quad 1.04, \quad 2.62, \quad \ldots, \quad 2.01.$$

For each sample $x_1^*, x_2^*, \ldots, x_{23}^*$ compute: $t^* = \bar{x}_{23}^* - 0.82/(s_{23}^*/\sqrt{23})$, where $s_{23}^* = \sqrt{\frac{1}{22} \sum (x_i^* - \bar{x}_{23}^*)^2}$.

23.5 c We need to estimate the critical value c_l^* such that $\mathrm{P}(T^* \le c_l^*) \approx 0.025$. We take $c_l^* = -2.101$, the 25th of the ordered values, an estimate for the $25/1000 = 0.025$ quantile. Similarly, c_l^* is estimated by the 976th, which is 2.088.
The bootstrap confidence interval uses the c^* values instead of the t-distribution values $\pm t_{n-1,\alpha/2}$, but beware: c_l^* is from the *left tail* and appears on the *right-hand side* of the interval and c_u^* on the left-hand side:

$$\left(\bar{x}_n - c_u^* \frac{s_n}{\sqrt{n}}, \; \bar{x}_n - c_l^* \frac{s_n}{\sqrt{n}} \right).$$

Substituting $c_l^* = -2.101$ and $c_u^* = 2.088$, the confidence interval becomes:

$$\left(0.82 - 2.088 \cdot \frac{1.78}{\sqrt{23}}, \; 0.82 + 2.101 \cdot \frac{1.78}{\sqrt{23}} \right) = (0.045, \, 1.600).$$

23.6 a Because events described by inequalities do not change when we multiply the inqualities by a positive constant or add or subtract a constant, the following equalities hold: $\mathrm{P}\left(\tilde{L}_n < \theta < \tilde{U}_n \right) = \mathrm{P}(3L_n + 7 < 3\mu + 7 < 3U_n + 7) = \mathrm{P}(3L_n < 3\mu < 3U_n) = \mathrm{P}(L_n < \mu < U_n)$, and this equals 0.95, as is given.

23.6 b The confidence interval for θ is obtained as the realization of $(\tilde{L}_n, \tilde{U}_n)$, that is: $(\tilde{l}_n, \tilde{u}_n) = (3l_n + 7, 3u_n + 7)$. This is obtained by transforming the confidence interval for μ (using the transformation that is applied to μ to get θ).

23.6 c We start with $\mathrm{P}(L_n < \mu < U_n) = 0.95$ and try to get $1 - \mu$ in the middle: $\mathrm{P}(L_n < \mu < U_n) = \mathrm{P}(-L_n > -\mu > -U_n) = \mathrm{P}(1 - L_n > 1 - \mu > 1 - U_n) = \mathrm{P}(1 - U_n < 1 - \mu < 1 - L_n)$, where we see that the minus sign causes an interchange: $\tilde{L}_n = 1 - U_n$ and $\tilde{U}_n = 1 - L_n$. The confidence interval: $(1 - 5, 1 - (-2)) = (-4, 3)$.

23.6 d If we knew that L_n and U_n were always positive, then we could conclude: $\mathrm{P}(L_n < \mu < U_n) = \mathrm{P}\left(L_n^2 < \mu^2 < U_n^2 \right)$ and we could just square the numbers in the confidence interval for μ to get the one for θ. Without the positivity assumption, the sharpest conclusion you can draw from $L_n < \mu < U_n$ is that μ^2 is smaller than the maximum of L_n^2 and U_n^2. So, $0.95 = \mathrm{P}(L_n < \mu < U_n) \le \mathrm{P}\left(0 \le \mu^2 < \max\{L_n^2, U_n^2\} \right)$ and the confidence interval $[0, \max\{l_n^2, u_n^2\}) = [0, 25)$ has a confidence of *at least* 95%. This kind of problem may occur when the transformation is not one-to-one (both -1 and 1 are mapped to 1 by squaring).

23.11 a For the 98% confidence interval the same formula is used as for the 95% interval, replacing the critical values by larger ones. This is the case, no matter whether the critical values are from the normal or t-distribution, or from a bootstrap experiment. Therefore, the 98% interval *contains* the 95%, and so must also contain the number 0.

23.11 b From a new bootstrap experiment we would obtain new and, most probably, different values c_u^* and c_l^*. It therefore could be, if the number 0 is close to the edge of the first bootstrap confidence interval, that it is just outside the new interval.

23.11 c The new dataset will resemble the old one in many ways, but things like the sample mean would most likely differ from the old one, and so there is no guarantee that the number 0 will again be in the confidence interval.

24.6 a The environmentalists are interested in a *lower* confidence bound, because they would like to make a statement like "We are 97.5% confidence that the concentration exceeds 1.68 ppm [and that is much too high.]" We have normal data, with σ unknown so we use $s_{16} = \sqrt{1.12} = 1.058$ as an estimate and use the critical value corresponding to 2.5% from the $t(15)$ distribution: $t_{15,0.025} = 2.131$. The lower confidence bound is $2.24 - 2.131 \cdot 1.058/\sqrt{16} = 2.24 - 0.56 = 1.68$, the interval: $(1.68, \infty)$.

24.6 b For similar reasons, the plant management constructs an *upper* confidence bound ("We are 97.5% confident pollution does not exceed 2.80 [and this is acceptable.]"). The computation is the same except for a minus sign: $2.24 + 2.131 \cdot 1.058/\sqrt{16} = 2.24 + 0.56 = 2.80$, so the interval is $[0, 2.80)$. Note that the computed upper and lower bounds are in fact the endpoints of the 95% two-sided confidence interval.

24.9 a From Section 8.4 we know: $P(M \le a) = [F_X(a)]^{12}$, so $P(M/\theta \le t) = P(M \le \theta t) = [F_X(\theta t)]^{12}$. Since X_i has a $U(0,\theta)$ distribution, $F_X(\theta t) = t$, for $0 \le t \le 1$. Substituting this shows the result.

24.9 b For c_l we need to solve $(c_l)^{12} = \alpha/2$, or $c_l = (\alpha/2)^{1/12} = (0.05)^{1/12} = 0.7791$. For c_u we need to solve $(c_u)^{12} = 1 - \alpha/2$, or $c_u = (1 - \alpha/2)^{1/12} = (0.95)^{1/12} = 0.9958$.

24.9 c From **b** we know that $P(c_l < M/\theta < c_u) = P(0.7790 < M/\theta < 0.9958) = 0.90$. Rewriting this equation, we get: $P(0.7790\,\theta < M < 0.9958\,\theta) = 0.90$ and $P(M/0.9958 < \theta < M/0.7790) = 0.90$. This means that $(m/0.9958, m/0.7790) = (3.013, 3.851)$ is a 90% confidence interval for θ.

24.9 d From **b** we derive the general formula:

$$P\left((\alpha/2)^{1/n} < \frac{M}{\theta} < (1 - \alpha/2)^{1/n}\right) = 1 - \alpha.$$

The left hand inequality can be rewritten as $\theta < M/(\alpha/2)^{1/n}$ and the right hand one as $M/(1 - \alpha/2)^{1/n} < \theta$. So, the statement above can be rewritten as:

$$P\left(\frac{M}{(1 - \alpha/2)^{1/n}} < \theta < \frac{M}{(\alpha/2)^{1/n}}\right) = 1 - \alpha,$$

so that the general formula for the confidence interval becomes:

$$\left(\frac{m}{(1 - \alpha/2)^{1/n}}, \frac{m}{(\alpha/2)^{1/n}}\right).$$

25.4 a Denote the observed numbers of cycles for the smokers by $X_1, X_2, \ldots, X_{n_1}$ and similarly $Y_1, Y_2, \ldots, Y_{n_2}$ for the nonsmokers. A test statistic should compare estimators for p_1 and p_2. Since the geometric distributions have expectations $1/p_1$

and $1/p_2$, we could compare the estimator $1/\bar{X}_{n_1}$ for p_1 with the estimator $1/\bar{Y}_{n_2}$ for p_2, or simply compare $\bar{X}_{n_1}$ with $\bar{Y}_{n_2}$. For instance, take test statistic $T = \bar{X}_{n_1} - \bar{Y}_{n_2}$. Values of T close to zero are in favor of H_0, and values far away from zero are in favor of H_1. Another possibility is $T = \bar{X}_{n_1}/\bar{Y}_{n_2}$.

25.4 b In this case, the maximum likelihood estimators $\hat{p}_1$ and $\hat{p}_2$ give better indications about p_1 and p_2. They can be compared in the same way as the estimators in **a**.

25.4 c The probability of getting pregnant during a cycle is p_1 for the smoking women and p_2 for the nonsmokers. The alternative hypothesis should express the belief that smoking women are *less likely* to get pregnant than nonsmoking women. Therefore take $H_1 : p_1 < p_2$.

25.10 a The alternative hypothesis should express the belief that the gross calorific *exceeds* 23.75 MJ/kg. Therefore take $H_1 : \mu > 23.75$.

25.10 b The p-value is the probability $P(\bar{X}_n \geq 23.788)$ under the null hypothesis. We can compute this probability by using that under the null hypothesis $\bar{X}_n$ has an $N(23.75, (0.1)^2/23)$ distribution:

$$P(\bar{X}_n \geq 23.788) = P\left(\frac{\bar{X}_n - 23.75}{0.1/\sqrt{23}} \geq \frac{23.788 - 23.75}{0.1/\sqrt{23}} \right) = P(Z \geq 1.82),$$

where Z has an $N(0,1)$ distribution. From Table B.1 we find $P(Z \geq 1.82) = 0.0344$.

25.11 A type I error occurs when $\mu = 0$ and $|t| \geq 2$. When $\mu = 0$, then T has an $N(0,1)$ distribution. Hence, by symmetry of the $N(0,1)$ distribution and Table B.1, we find that the probability of committing a type I error is

$$P(|T| \geq 2) = P(T \leq -2) + P(T \geq 2) = 2 \cdot P(T \geq 2) = 2 \cdot 0.0228 = 0.0456.$$

26.5 a The p-value is $P(X \geq 15)$ under the null hypothesis $H_0 : p = 1/2$. Using Table 26.3 we find $P(X \geq 15) = 1 - P(X \leq 14) = 1 - 0.8950 = 0.1050$.

26.5 b Only values close to 23 are in favor of $H_1 : p > 1/2$, so the critical region is of the form $K = \{c, c+1, \ldots, 23\}$. The critical value c is the smallest value, such that $P(X \geq c) \leq 0.05$ under $H_0 : p = 1/2$, or equivalently, $1 - P(X \leq c-1) \leq 0.05$, which means $P(X \leq c-1) \geq 0.95$. From Table 26.3 we conclude that $c - 1 = 15$, so that $K = \{16, 17, \ldots, 23\}$.

26.5 c A type I error occurs if $p = 1/2$ and $X \geq 16$. The probability that this happens is $P(X \geq 16 \mid p = 1/2) = 1 - P(X \leq 15 \mid p = 1/2) = 1 - 0.9534 = 0.0466$, where we have used Table 26.3 once more.

26.5 d In this case, a type II error occurs if $p = 0.6$ and $X \leq 15$. To approximate $P(X \leq 15 \mid p = 0.6)$, we use the same reasoning as in Section 14.2, but now with $n = 23$ and $p = 0.6$. Write X as the sum of independent Bernoulli random variables: $X = R_1 + \cdots + R_n$, and apply the central limit theorem with $\mu = p = 0.6$ and $\sigma^2 = p(1-p) = 0.24$. Then

$$P(X \leq 15) = P(R_1 + \cdots + R_n \leq 15)$$
$$= P\left(\frac{R_1 + \cdots + R_n - n\mu}{\sigma\sqrt{n}} \leq \frac{15 - n\mu}{\sigma\sqrt{n}} \right)$$
$$= P\left(Z_{23} \geq \frac{15 - 13.8}{\sqrt{0.24}\sqrt{23}} \right) \approx \Phi(0.51) = 0.6950.$$

26.8 a Test statistic $T = \bar{X}_n$ takes values in $(0, \infty)$. Recall that the $Exp(\lambda)$ distribution has expectation $1/\lambda$, and that according to the law of large numbers $\bar{X}_n$ will be close to $1/\lambda$. Hence, values of $\bar{X}_n$ close to 1 are in favor of $H_0 : \lambda = 1$, and *only* values of $\bar{X}_n$ close to zero are in favor $H_1 : \lambda > 1$. Large values of $\bar{X}_n$ also provide evidence against $H_0 : \lambda = 1$, but even stronger evidence against $H_1 : \lambda > 1$. We conclude that $T = \bar{X}_n$ has critical region $K = (0, c_l]$. This is an example in which the alternative hypothesis and the test statistic deviate from the null hypothesis in *opposite* directions.
Test statistic $T' = e^{-\bar{X}_n}$ takes values in $(0, 1)$. Values of $\bar{X}_n$ close to zero correspond to values of T' close to 1, and large values of $\bar{X}_n$ correspond to values of T' close to 0. Hence, *only* values of T' close to 1 are in favor $H_1 : \lambda > 1$. We conclude that T' has critical region $K' = [c_u, 1)$. Here the alternative hypothesis and the test statistic deviate from the null hypothesis in the *same* direction.

26.8 b Again, values of $\bar{X}_n$ close to 1 are in favor of $H_0 : \lambda = 1$. Values of $\bar{X}_n$ close to zero suggest $\lambda > 1$, whereas large values of $\bar{X}_n$ suggest $\lambda < 1$. Hence, both small and large values of $\bar{X}_n$ are in favor of $H_1 : \lambda \neq 1$. We conclude that $T = \bar{X}_n$ has critical region $K = (0, c_l] \cup [c_u, \infty)$.
Small and large values of $\bar{X}_n$ correspond to values of T' close to 1 and 0. Hence, values of T' both close to 0 and close 1 are in favor of $H_1 : \lambda \neq 1$. We conclude that T' has critical region $K' = (0, c_l'] \cup [c_u', 1)$. Both test statistics deviate from the null hypothesis in the same directions as the alternative hypothesis.

26.9 a Test statistic $T = (\bar{X}_n)^2$ takes values in $[0, \infty)$. Since μ is the expectation of the $N(\mu, 1)$ distribution, according to the law of large numbers, $\bar{X}_n$ is close to μ. Hence, values of $\bar{X}_n$ close to zero are in favor of $H_0 : \mu = 0$. Large negative values of $\bar{X}_n$ suggest $\mu < 0$, and large positive values of $\bar{X}_n$ suggest $\mu > 0$. Therefore, both large negative and large positive values of $\bar{X}_n$ are in favor of $H_1 : \mu \neq 0$. These values correspond to large positive values of T, so T has critical region $K = [c_u, \infty)$. This is an example in which the test statistic deviates from the null hypothesis in *one* direction, whereas the alternative hypothesis deviates in *two* directions.
Test statistic T' takes values in $(-\infty, 0) \cup (0, \infty)$. Large negative values and large positive values of $\bar{X}_n$ correspond to values of T' close to zero. Therefore, T' has critical region $K' = [c_l', 0) \cup (0, c_u']$. This is an example in which the test statistic deviates from the null hypothesis for small values, whereas the alternative hypothesis deviates for large values.

26.9 b Only large positive values of $\bar{X}_n$ are in favor of $\mu > 0$, which correspond to large values of T. Hence, T has critical region $K = [c_u, \infty)$. This is an example where the test statistic has the *same type* of critical region with a one-sided or two-sided alternative. Of course, the critical value c_u in part **b** is different from the one in part **a**.
Large positive values of $\bar{X}_n$ correspond to small positive values of T'. Hence, T' has critical region $K' = (0, c_u']$. This is another example where the test statistic deviates from the null hypothesis for small values, whereas the alternative hypothesis deviates for large values.

27.5 a The interest is whether the inbreeding coefficient exceeds 0. Let μ represent this coefficient for the species of wasps. The value 0 is the a priori specified value of the parameter, so test null hypothesis $H_0 : \mu = 0$. The alternative hypothesis should express the belief that the inbreeding coefficient *exceeds* 0. Hence, we take alternative hypothesis $H_1 : \mu > 0$. The value of the test statistic is

$$t = \frac{0.044}{0.884/\sqrt{197}} = 0.70.$$

27.5 b Because $n = 197$ is large, we approximate the distribution of T under the null hypothesis by an $N(0,1)$ distribution. The value $t = 0.70$ lies to the right of zero, so the p-value is the right tail probability $P(T \geq 0.70)$. By means of the normal approximation we find from Table B.1 that the right tail probability

$$P(T \geq 0.70) \approx 1 - \Phi(0.70) = 0.2420.$$

This means that the value of the test statistic is not very far in the (right) tail of the distribution and is therefore not to be considered exceptionally large. We do not reject the null hypothesis.

27.7 a The data are modeled by a simple linear regression model: $Y_i = \alpha + \beta x_i$, where Y_i is the gas consumption and x_i is the average outside temperature in the ith week. Higher gas consumption as a consequence of smaller temperatures corresponds to $\beta < 0$. It is natural to consider the value 0 as the a priori specified value of the parameter (it corresponds to no change of gas consumption). Therefore, we take null hypothesis $H_0 : \beta = 0$. The alternative hypothesis should express the belief that the gas consumption increases as a consequence of smaller temperatures. Hence, we take alternative hypothesis $H_1 : \beta < 0$. The value of the test statistic is

$$t_b = \frac{\hat{\beta}}{s_b} = \frac{-0.3932}{0.0196} = -20.06.$$

The test statistic T_b has a t-distribution with $n - 2 = 24$ degrees of freedom. The value -20.06 is smaller than the left critical value $t_{24,0.05} = -1.711$, so we reject.

27.7 b For the data after insulation, the value of the test statistic is

$$t_b = \frac{-0.2779}{0.0252} = -11.03,$$

and T_b has a $t(28)$ distribution. The value -11.03 is smaller than the left critical value $t_{28,0.05} = -1.701$, so we reject.

28.5 a When $aS_X^2 + bS_Y^2$ is unbiased for σ^2, we should have $E\left[aS_X^2 + bS_Y^2\right] = \sigma^2$. Using that S_X^2 and S_Y^2 are both unbiased for σ^2, i.e., $E\left[S_X^2\right] = \sigma^2$ and $E\left[S_Y^2\right] = \sigma^2$, we get

$$E\left[aS_X^2 + bS_Y^2\right] = aE\left[S_X^2\right] + bE\left[S_Y^2\right] = (a + b)\sigma^2.$$

Hence, $E\left[aS_X^2 + bS_Y^2\right] = \sigma^2$ for all $\sigma > 0$ if and only if $a + b = 1$.

28.5 b By independence of S_X^2 and S_Y^2 write

$$\text{Var}\left(aS_X^2 + (1 - a)S_Y^2\right) = a^2\text{Var}\left(S_X^2\right) + (1 - a)^2\text{Var}\left(S_Y^2\right)$$

$$= \left(\frac{a^2}{n - 1} + \frac{(1 - a)^2}{m - 1}\right)2\sigma^4.$$

To find the value of a that minimizes this, differentiate with respect to a and put the derivative equal to zero. This leads to

$$\frac{2a}{n - 1} - \frac{2(1 - a)}{m - 1} = 0.$$

Solving for a yields $a = (n - 1)/(n + m - 2)$. Note that the second derivative of $\text{Var}\left(aS_X^2 + (1 - a)S_Y^2\right)$ is positive so that this is indeed a minimum.

References

1. J. Bernoulli. *Ars Conjectandi.* Basel, 1713.
2. J. Bernoulli. The most probable choice between several discrepant observations and the formation therefrom of the most likely induction. ( ):3–33, 1778. With a comment by Euler.
3. P. Billingsley. *Probability and measure.* John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication.
4. L.D. Brown, T.T. Cai, and A. DasGupta. Interval estimation for a binomial proportion. *Stat. Science*, 16(2):101–133, 2001.
5. S.R. Dalal, E.B. Fowlkes, and B. Hoadley. Risk analysis of the space shuttle: pre-*Challenger* prediction of failure. *J. Am. Stat. Assoc.*, 84:945–957, 1989.
6. J. Daugman. Wavelet demodulation codes, statistical independence, and pattern recognition. In *Institute of Mathematics and its Applications, Proc. 2nd IMA-IP: Mathematical Methods, Algorithms, and Applications (Blackledge and Turner, Eds)*, pages 244–260. Horwood, London, 2000.
7. B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.
8. W. Feller. *An introduction to probability theory and its applications, Vol. II.* John Wiley & Sons Inc., New York, 1971.
9. R.A. Fisher. On an absolute criterion for fitting frequency curves. *Mess. Math.*, 41:155–160, 1912.
10. R.A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1(4):3–32, 1921.
11. H.S. Fogler. *Elements of chemical reaction engineering.* Prentice-Hall, Upper Saddle River, 1999.
12. D. Freedman and P. Diaconis. On the histogram as a density estimator: L_2 theory. *Z. Wahrsch. Verw. Gebiete*, 57(4):453–476, 1981.
13. C.F. Gauss. *Theoria motus corporum coelestium in sectionis conicis solem ambientum. In: Werke. Band VII.* Georg Olms Verlag, Hildesheim, 1973. Reprint of the 1906 original.
14. P. Hall. *The bootstrap and Edgeworth expansion.* Springer-Verlag, New York, 1992.
15. R. Herz, H.G. Schlichter, and W. Siegener. *Angewandte Statistik für Verkehrs- und Regionalplaner.* Werner-Ingenieur-Texte 42, Werner-Verlag, Düsseldorf, 1992.

16. J.L. Lagrange. *Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observations.* Paris, 1770–73. Œvres 2, 1886.

17. J.H. Lambert. *Photometria.* Augustae Vindelicorum, 1760.

18. R.J. MacKay and R.W. Oldford. Scientific method, statistical method and the speed of light. *Stat. Science*, 15(3):254–278, 2000.

19. J. Moynagh, H. Schimmel, and G.N. Kramer. The evaluation of tests for the diagnosis of transmissible spongiform encephalopathy in bovines. Technical report, European Commission, Directorate General XXIV, Brussels, 1999.

20. V. Pareto. *Cours d'economie politique.* Rouge, Lausanne et Paris, 1897.

21. E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33:1065–1076, 1962.

22. K. Pearson. *Philos. Trans.*, 186:343–414, 1895.

23. R. Penner and D.G. Watts. Mining information. *The Amer. Stat.*, 45:4–9, 1991.

24. Commission Rogers. Report on the space shuttle *Challenger* accident. Technical report, Presidential commission on the Space Shuttle Challenger Accident, Washington, DC, 1986.

25. M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27:832–837, 1956.

26. S.M. Ross. *A first course in probability.* Prentice-Hall, Inc., New Jersey, sixth edition, 1984.

27. R. Ruggles and H. Brodie. An empirical approach to economic intelligence in World War II. *Journal of the American Statistical Association*, 42:72–91, 1947.

28. E. Rutherford and H. Geiger (with a note by H. Bateman). The probability variations in the distribution of α particles. *Phil.Mag.*, 6:698–704, 1910.

29. D.W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

30. S. Siegel and N.J. Castellan. *Nonparametric statistics for the behavioral sciences.* McGraw-Hill, New York, second edition, 1988.

31. B.W. Silverman. *Density estimation for statistics and data analysis.* Chapman & Hall, London, 1986.

32. K. Singh. On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics*, 9:1187–1195, 1981.

33. S.M. Stigler. *The history of statistics — the measurement of uncertainty before 1900.* Cambridge, Massachusetts, 1986.

34. H.A. Sturges. *J. Amer. Statist. Ass.*, 21, 1926.

35. J.W. Tukey. *Exploratory data analysis.* Addison-Wesley, Reading, 1977.

36. S.A. van de Geer. *Applications of empirical process theory.* Cambridge University Press, Cambridge, 2000.

37. J.G. Wardrop. Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers*, 1, 1952.

38. C.R. Weinberg and B.C. Gladen. The beta-geometric distribution applied to comparative fecundability studies. *Biometrics*, 42(3):547–560, 1986.

39. H. Westergaard. *Contributions to the history of statistics.* Agathon, New York, 1968.

40. E.B. Wilson. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.*, 22:209–212, 1927.

41. D.R. Witte et al. Cardiovascular mortality in Dutch men during 1996 European foolball championship: longitudinal population study. *British Medical Journal*, 321:1552–1554, 2000.

List of symbols

$\emptyset$	empty set, page 14
α	significance level, page 384
A^c	complement of the event A, page 14
$A \cap B$	intersection of A and B, page 14
$A \subset B$	A subset of B, page 15
$A \cup B$	union of A and B, page 14
$Ber(p)$	Bernoulli distribution with parameter p, page 45
$Bin(n, p)$	binomial distribution with parameters n and p, page 48
c_l, c_u	left and right critical values, page 388
$Cau(\alpha, \beta)$	Cauchy distribution with parameters α en β, page 161
$\mathrm{Cov}(X, Y)$	covariance between X and Y, page 139
$\mathrm{E}[X]$	expectation of the random variable X, page 90, 91
$Exp(\lambda)$	exponential distribution with parameter λ, page 62
Φ	distribution function of the standard normal distribution, page 65
ϕ	probability density of the standard normal distribution, page 65
f	probability density function, page 57
f	joint probability density function, page 119
F	distribution function, page 44
F	joint distribution function, page 118
F^{inv}	inverse function of distribution function F, page 73
F_n	empirical distribution function, page 219
$f_{n,h}$	kernel density estimate, page 213
$Gam(\alpha, \lambda)$	gamma distribution with parameters α en λ, page 157
$Geo(p)$	geometric distribution with parameter p, page 49
H_0, H_1	null hypothesis and alternative hypothesis, page 374

$L(\theta)$	likelihood function, page 317
$\ell(\theta)$	loglikelihood function, page 319
Med_n	sample median of a dataset, page 231
$n!$	n factorial, page 14
$N(\mu, \sigma^2)$	normal distribution with parameters μ and σ^2, page 64
Ω	sample space, page 13
$Par(\alpha)$	Pareto distribution with parameter α, page 63
$Pois(\mu)$	Poisson distribution with parameter μ, page 170
$\mathrm{P}(A\,|\,C)$	conditional probability of A given C, page 26
$\mathrm{P}(A)$	probability of the event A, page 16
$q_n(p)$	pth empirical quantile, page 234
q_p	pth quantile or $100p$th percentile, page 66
$\rho(X, Y)$	correlation coefficient between X and Y, page 142
s_n^2	sample variance of a dataset, page 233
S_n^2	sample variance of random sample, page 292
$t(m)$	t-distribution with m degrees of freedom, page 348
$t_{m,p}$	critical value of the $t(m)$ distribution, page 348
$U(\alpha, \beta)$	uniform distribution with parameters α and β, page 60
$\text{Var}(X)$	variance of the random variable X, page 96
$\bar{x}_n$	sample mean of a dataset, page 231
$\bar{X}_n$	average of the random variables $X_1, \ldots, X_n$, page 182
z_p	critical value of the $N(0, 1)$ distribution, page 345

Index