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

British Library Cataloguing in Publication Data
A modern introduction to probability and statistics. —
(Springer texts in statistics)
1. Probabilities 2. Mathematical statistics
I. Dekking, F. M.
519.2
ISBN 1852338962

Library of Congress Cataloging-in-Publication Data
A modern introduction to probability and statistics : understanding why and how / F.M. Dekking ... [et al.].
p. cm. — (Springer texts in statistics)
Includes bibliographical references and index.
ISBN 1-85233-896-2
1. Probabilities—Textbooks. 2. Mathematical statistics—Textbooks. I. Dekking, F.M. II. Series.
QA273.M645 2005
519.2—dc22 2004057700

ISBN-10: 1-85233-896-2
ISBN-13: 978-1-85233-896-1

Springer Science+Business Media
springeronline.com

Printed in the United States of America
12/3830/543210 Printed on acid-free paper SPIN 10943403

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 $\boxplus$ and those with only a short answer are marked with $\boxdot$ (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
January 2005

F. M. Dekking
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.

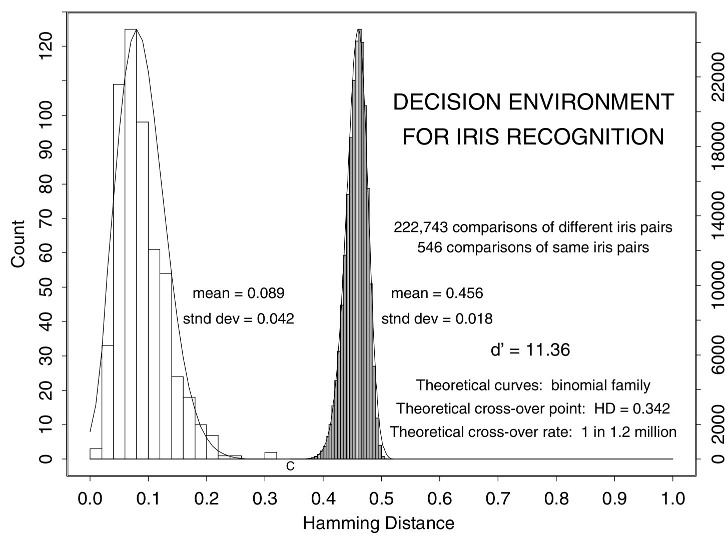


Fig. 1.1. Comparison of same and different iris pairs.

Source: J.Daugman. *Second IMA Conference on Image Processing: Mathematical Methods, Algorithms and Applications*, 2000. © Ellis Horwood Publishing Limited.

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.”

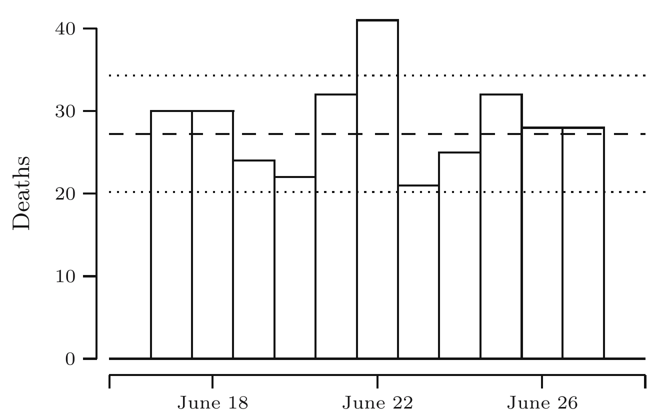


Fig. 1.2. Number of deaths from infarction or stroke in (part of) June 1996.

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.

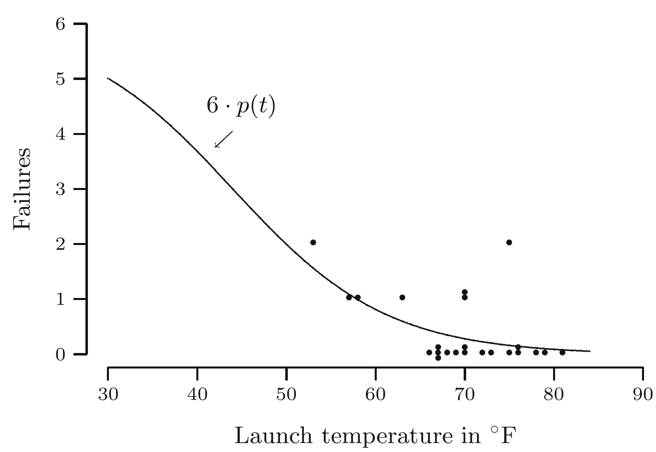
1.4 The space shuttle *Challenger*

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.



Source: based on data from Volume VI of the Report of the Presidential Commission on the space shuttle Challenger accident, Washington, DC, 1986.

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:

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

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:

$$\text{time} = \frac{\alpha/2\pi}{\text{rps}}.$$

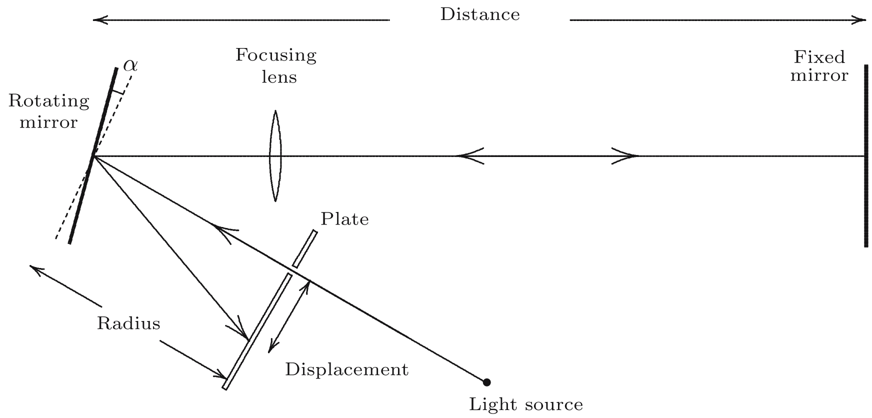


Fig. 1.4. Michelson's experiment.

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 reduce variability. His final dataset consists of 100 separate measurements (see Table 17.1), but each is in fact summarized and averaged from repeated measurements 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% confidence 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

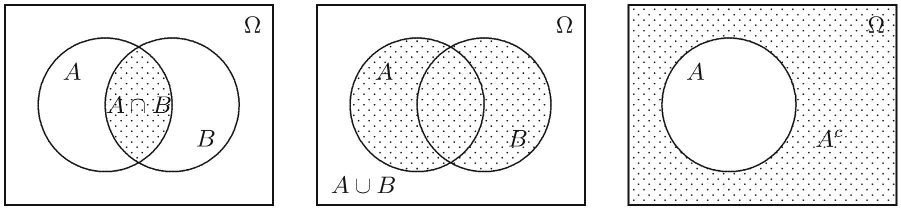
$$L = \{\text{Jan, Mar, May, Jul, Aug, Oct, Dec}\}.$$

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 $R = \{\text{Jan}, \text{Feb}, \text{Mar}, \text{Apr}, \text{Sep}, \text{Oct}, \text{Nov}, \text{Dec}\}$), then the long months that contain the letter r are

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

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.



Intersection $A \cap B$ Union $A \cup B$ Complement A^c

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:

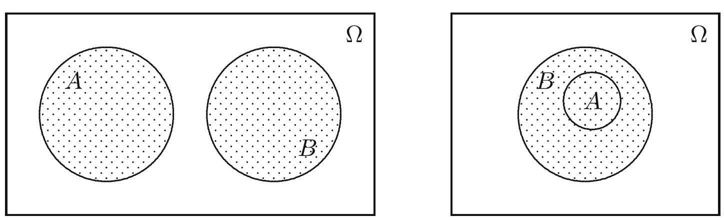
> "It is certainly not true that neither John nor Mary is to blame."

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.



Disjoint sets A and B A subset of B

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 $\mathrm{P}(A)$ in [0,1] such that
> (i) $\mathrm{P}(\Omega) = 1$, and
> (ii) $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$ if A and B are disjoint.
> The number $\mathrm{P}(A)$ 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

$$\mathrm{P}(A \cup B \cup C) = \mathrm{P}(A \cup B) + \mathrm{P}(C) = \mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(C)\,.$$

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

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

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

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

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; $\Omega = \{\omega_1, \omega_2, \ldots\}$. 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

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

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

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

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.,

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

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 $\mathrm{P}(n)$ for each n. Clearly $\mathrm{P}(1) = p$, 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

$$\mathrm{P}(2) = (1-p)p.$$

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,

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

Does this define a probability function on $\Omega = \{1, 2, 3, \ldots\}$? Then we should at least have $\mathrm{P}(\Omega) = 1$. It is not directly clear how to calculate $\mathrm{P}(\Omega)$: 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 $\mathrm{P}(A)$ in $[0, 1]$ such that
(i) $\mathrm{P}(\Omega) = 1$, and
(ii) $\mathrm{P}(A_1 \cup A_2 \cup A_3 \cup \cdots) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \mathrm{P}(A_3) + \cdots$
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

$$\begin{aligned} \mathrm{P}(A_1 \cup A_2) &= \mathrm{P}(A_1 \cup A_2 \cup \emptyset \cup \emptyset \cup \cdots) \\ &= \mathrm{P}(A_1) + \mathrm{P}(A_2) + 0 + 0 + \cdots = \mathrm{P}(A_1) + \mathrm{P}(A_2)\,. \end{aligned}$$

Now we can compute the probability of Ω:

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

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 $\mathrm{P}(\Omega) = p \cdot \dfrac{1}{p} = 1$.

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

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

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

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 $\mathrm{P}(B) = 10(1-p)^3 p^2$.

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 $\mathrm{P}(A) = 2/3$, $\mathrm{P}(B) = 1/6$, and $\mathrm{P}(A \cap B) = 1/9$. What is $\mathrm{P}(A \cup B)$?

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 $\mathrm{P}(C) = 0.3, \mathrm{P}(D) = 0.4$, and $\mathrm{P}(C \cap D) = 0.2$. What is $\mathrm{P}(C^c \cap D)$?

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 $\mathrm{P}(A \cup B \cup C) - \mathrm{P}(B) - \mathrm{P}(C) + \mathrm{P}(B \cap C)$?

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 $\mathrm{P}(A \setminus B) = \mathrm{P}(A) - \mathrm{P}(B)$ if B implies A, i.e., if $B \subset A$.

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

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

2.7 ⊡ Let A and B be two events. Suppose that $\mathrm{P}(A) = 0.4$, $\mathrm{P}(B) = 0.5$, and $\mathrm{P}(A \cap B) = 0.1$. 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: $\mathrm{P}(D_1) \leq 10^{-6}$ *and* $\mathrm{P}(D_2) \leq 10^{-6}$. 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 $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$ 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

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

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 $\mathrm{P}(E \cap F) = \mathrm{P}(F \cap G) = \mathrm{P}(E \cap G) = 1/3$. Can you determine $\mathrm{P}(E)$?
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 $L = \{\text{Jan, Mar, May, Jul, Aug, Oct, Dec}\}$ and $R = \{\text{Jan, Feb, Mar, Apr, Sep, Oct, Nov, Dec}\}$, 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 $R \cap L = \{\text{Jan, Mar, Oct, Dec}\}$—are favorable, so we reassess the probability as $4/7$. We call this the *conditional probability of R given L*, and we write:

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

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

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 $\mathrm{P}(A) = \mathrm{P}(213 \text{ or } 312) = 1/3$; by symmetry, the same is found for $\mathrm{P}(B)$. 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 $\mathrm{P}(A \mid C) = 1/2$. The probability of $A \cap C$ is half of $\mathrm{P}(C)$. No element of C also belongs to B, so $\mathrm{P}(B \mid C) = 0$.

Quick exercise 3.2 Calculate $\mathrm{P}(C \mid A)$ and $\mathrm{P}(C^c \mid A \cup B)$.

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:

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

provided $\mathrm{P}(C) > 0$.

Quick exercise 3.3 Show that $\mathrm{P}(A \mid C) + \mathrm{P}(A^c \mid C) = 1$.

This exercise shows that the rule $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$ also holds for conditional probabilities. In fact, even more is true: if we have a fixed conditioning event C and define $Q(A) = \mathrm{P}(A \mid C)$ 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:

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

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:

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

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:

$$\mathrm{P}(B_3) = \mathrm{P}(A_3 \cap B_2) = \mathrm{P}(A_3 \,|\, B_2)\mathrm{P}(B_2)\,.$$

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

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

and so

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

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:

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

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

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

This can be used to compute the probability for arbitrary n. For example, we find: $\mathrm{P}(B_{22}) = 0.5243$ and $\mathrm{P}(B_{23}) = 0.4927$. In Figure 3.1 the probability

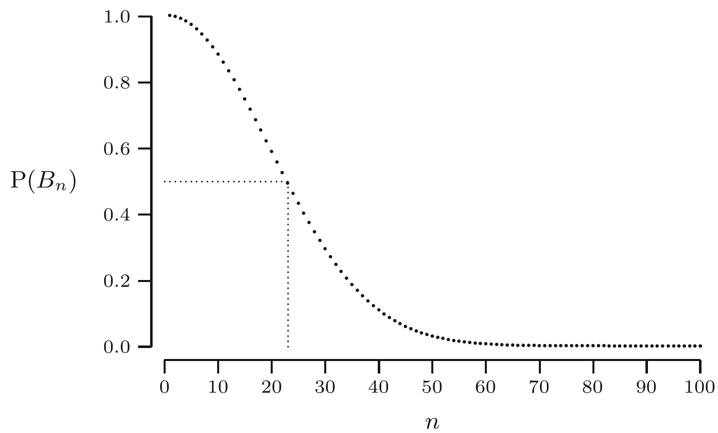


Fig. 3.1. The probability $P(B_n)$ of no coincident birthdays for $n = 1, \ldots, 100$.

$P(B_n)$ 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

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

or we may condition on A and get

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

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

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

but just trying to understand the conditional probability $P(B_2 \mid A_3)$ 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:

$$\begin{aligned} \mathrm{P}(T \mid B) &= 0.70, \\ \mathrm{P}(T \mid B^c) &= 0.10. \end{aligned}$$

Suppose we want to determine the probability $\mathrm{P}(T)$ 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

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

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!):

$$\begin{aligned} \mathrm{P}(T \cap B) &= \mathrm{P}(T \mid B) \cdot \mathrm{P}(B) \\ \mathrm{P}(T \cap B^c) &= \mathrm{P}(T \mid B^c) \cdot \mathrm{P}(B^c) \end{aligned} \tag{3.1}$$

so that

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

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] $\mathrm{P}(B) = 0.02$; then from the last equation we conclude: $\mathrm{P}(T) = 0.02 \cdot 0.70 + (1 - 0.02) \cdot 0.10 = 0.112$.

Quick exercise 3.6 Calculate $\mathrm{P}(T)$ when $\mathrm{P}(T \mid B) = 0.99$ and $\mathrm{P}(T \mid B^c) = 0.05$.

Following is a general statement of the law.

The law of total probability. Suppose $C_1, C_2, \ldots, C_m$ 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:

$$\mathrm{P}(A) = \mathrm{P}(A \mid C_1)\mathrm{P}(C_1) + \mathrm{P}(A \mid C_2)\mathrm{P}(C_2) + \cdots + \mathrm{P}(A \mid C_m)\mathrm{P}(C_m) .$$

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 $\mathrm{P}(A) = \mathrm{P}(A \cap C_1) + \cdots + \mathrm{P}(A \cap C_5)$, and for each i the multiplication rule states $\mathrm{P}(A \cap C_i) = \mathrm{P}(A \mid C_i) \cdot \mathrm{P}(C_i)$.

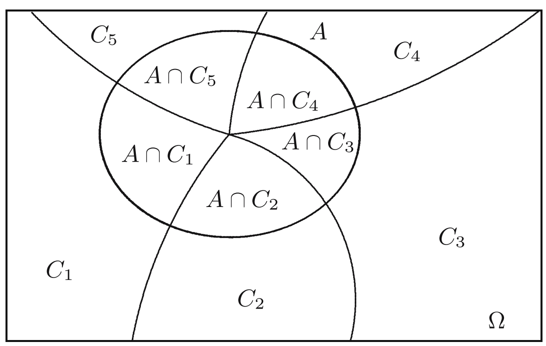


Fig. 3.2. The law of total probability (illustration for $m = 5$).

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 $\mathrm{P}(B \mid T)$. The information we were given is $\mathrm{P}(T \mid B)$, 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 $\mathrm{P}(B) \approx 0.000013$.

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

So with $\mathrm{P}(B) = 0.02$ we find

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

and by a similar calculation: $\mathrm{P}(B \,|\, T^c) = 0.0068$. These probabilities reflect that this Test A is not a very good test; a perfect test would result in $\mathrm{P}(B \,|\, T) = 1$ and $\mathrm{P}(B \,|\, T^c) = 0$. In Exercise 3.4 we redo this calculation, replacing $\mathrm{P}(B) = 0.02$ 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 $C_1, C_2, \ldots, C_m$ 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:
>
> $$\mathrm{P}(C_i \,|\, A) = \frac{\mathrm{P}(A \,|\, C_i) \cdot \mathrm{P}(C_i)}{\mathrm{P}(A \,|\, C_1)\mathrm{P}(C_1) + \mathrm{P}(A \,|\, C_2)\mathrm{P}(C_2) + \cdots + \mathrm{P}(A \,|\, C_m)\mathrm{P}(C_m)}.$$

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

$$\mathrm{P}(C_i \,|\, A) = \frac{\mathrm{P}(A \,|\, C_i) \cdot \mathrm{P}(C_i)}{\mathrm{P}(A)} \tag{3.3}$$

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

Quick exercise 3.7 Calculate $\mathrm{P}(B \,|\, T)$ and $\mathrm{P}(B \,|\, T^c)$ if $\mathrm{P}(T \,|\, B) = 0.99$ and $\mathrm{P}(T \,|\, B^c) = 0.05$.

3.4 Independence

Consider three probabilities from the previous section:

$$\mathrm{P}(B) = 0.02,$$
$$\mathrm{P}(B \,|\, T) = 0.125,$$
$$\mathrm{P}(B \,|\, T^c) = 0.0068.$$

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: $\mathrm{P}(B \mid T) = \mathrm{P}(B)$. In this case we would call B independent of T.

DEFINITION. An event A is called *independent of* B if

$$\mathrm{P}(A \mid B) = \mathrm{P}(A)\,.$$

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

$$A \text{ independent of } B \quad \Leftrightarrow \quad A^c \text{ independent of } B. \tag{3.4}$$

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

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

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

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

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

$$A \text{ independent of } B \quad \Leftrightarrow \quad B \text{ independent of } A. \tag{3.5}$$

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:

$$\mathrm{P}(A \mid B) = \mathrm{P}(A)\,,$$
$$\mathrm{P}(B \mid A) = \mathrm{P}(B)\,,$$
$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)\,,$$

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 $\mathrm{P}(H) = 1/2$. Also, $\mathrm{P}(H \mid R) = 1/2$. So H and R are independent, and we conclude, for example, $\mathrm{P}(R^c \mid H^c) = \mathrm{P}(R^c) = 1 - 8/12 = 1/3$.

We know that $\mathrm{P}(L \cap H) = 1/4$ and $\mathrm{P}(L) = 7/12$. 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
>
> $$\mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_m) = \mathrm{P}(A_1)\,\mathrm{P}(A_2)\cdots\mathrm{P}(A_m)$$
>
> *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, $\mathrm{P}(A) = \mathrm{P}(B) = 1/2$, but also

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

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

$$\mathrm{P}(C \,|\, A) = \mathrm{P}(B \,|\, A) = \mathrm{P}(B) = \frac{1}{2} = \mathrm{P}(C)\,.$$

By symmetry, also $\mathrm{P}(C \,|\, B) = \mathrm{P}(C)$, 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:

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

and

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

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 $\mathrm{P}(N \,|\, L) = 3/7$.

3.2 The event A is contained in C. So when A occurs, C also occurs; therefore $\mathrm{P}(C \,|\, A) = 1$.

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 $\mathrm{P}(C^c \,|\, A \cup B) = 1/2$.

3.3 Using the definition we find:

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

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

$$\mathrm{P}(A \cap C) + \mathrm{P}(A^c \cap C) = \mathrm{P}(C)\,.$$

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:

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

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 $\mathrm{P}(R_3 \cap R_4^c) = \mathrm{P}(R_3) - \mathrm{P}(R_4) = \mathrm{e}^{-3} - \mathrm{e}^{-4}$. Using the complement rule: $\mathrm{P}(R_4^c) = 1 - \mathrm{P}(R_4) = 1 - \mathrm{e}^{-4}$. Together:

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

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

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

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

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

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

3.6 Repeating the calculation we find:

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

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

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

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

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

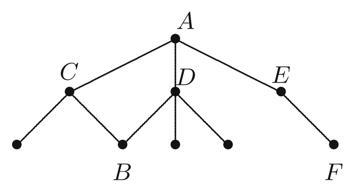
$$\mathrm{P}(B \mid T^c) = \frac{0.0002}{0.9312} = 0.00021.$$

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

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

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 $\mathrm{P}(A \mid B)$.

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 $\mathrm{P}(S_1)$, $\mathrm{P}(S_2 \mid S_1)$, and $\mathrm{P}(S_2 \mid S_1^c)$.

b. Compute $\mathrm{P}(S_2)$ 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 $\mathrm{P}(T \mid B) = 0.70$ and $\mathrm{P}(T \mid B^c) = 0.10$. Assume that the BSE risk for the Netherlands is the same as in 2003, when it was estimated to be $\mathrm{P}(B) = 1.3 \cdot 10^{-5}$. Compute $\mathrm{P}(B \mid T)$ and $\mathrm{P}(B \mid T^c)$.

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 $\mathrm{P}(D_1)$.

b. Compute $\mathrm{P}(D_2 \mid D_1)$ and $\mathrm{P}(D_1 \cap D_2)$.

c. Compute $\mathrm{P}(D_3 \mid D_1 \cap D_2)$ and $\mathrm{P}(D_1 \cap D_2 \cap D_3)$.

d. Continue the procedure to obtain the probability of a "dream draw": $\mathrm{P}(D_1 \cap \cdots \cap D_5)$.

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 $\mathrm{P}(W \mid R)$ and $\mathrm{P}(W \mid R^c)$.

b. Compute $\mathrm{P}(W)$ using the law of total probability.

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

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: $\mathrm{P}(T \mid D) = 0.98$ and $\mathrm{P}(T^c \mid D^c) = 0.95$.

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 $\mathrm{P}(W \mid G)$.

b. Show that $\mathrm{P}(W) = p^2 + 2p(1-p)\mathrm{P}(W \mid D)$ and use $\mathrm{P}(W) = \mathrm{P}(W \mid D)$ (why is this so?) to determine $\mathrm{P}(W)$.

c. Explain why the answers are the same.

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

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

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

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

$$\mathrm{P}(S = k),$$

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

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

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

while

$$\mathrm{P}(S = 13) = \mathrm{P}(\,\emptyset\,) = 0,$$

because 13 is an "impossible outcome."

QUICK EXERCISE 4.2 Use Table 4.1 to determine $\mathrm{P}(S = k)$ 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_1					
ω_2	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

$$p(a) = \mathrm{P}(X = a) \quad \text{for } -\infty < a < \infty.$$

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
>
> $$F(a) = \mathrm{P}(X \le a) \quad \text{for } -\infty < a < \infty.$$

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

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

then

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

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.

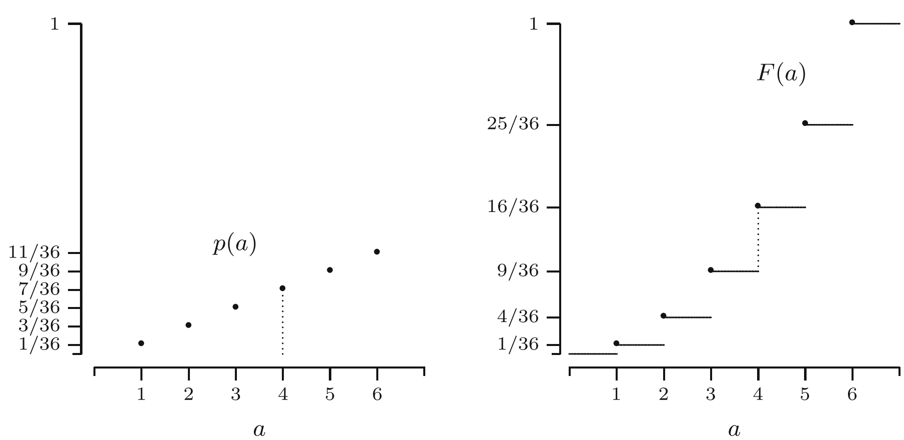


Fig. 4.1. Probability mass function and distribution function of M.

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,

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

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 $F(a) = \mathrm{P}(X < a) + p(a)$.

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
>
> $$p_X(1) = \mathrm{P}(X = 1) = p \quad \text{and} \quad p_X(0) = \mathrm{P}(X = 0) = 1 - p.$$
>
> 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:

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

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

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

Since $\mathrm{P}(X \geq 6) = \mathrm{P}(X = 6) + \cdots + \mathrm{P}(X = 10)$, it is now an easy (but tedious) exercise to determine the probability that you will pass. One finds that $\mathrm{P}(X \geq 6) = 0.0197$. 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 distribution* with parameters n and p, where $n = 1, 2, \ldots$ and $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(k) = \mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n.$$

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.

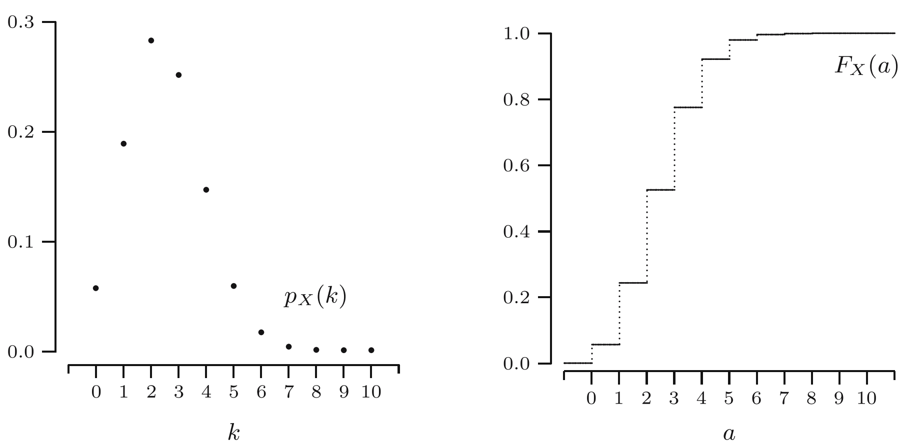


Fig. 4.2. Probability mass function and distribution function of the *Bin*$(10, \frac{1}{4})$ distribution.

4.4 The geometric distribution

In 1986, Weinberg and Gladen [38] investigated the number of menstrual cycles 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 $\mathrm{P}(X = 1) = p$. Due to the independence of consecutive cycles, one finds for $k = 1, 2, \ldots$ that

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

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

$$p_X(k) = \mathrm{P}(X = k) = (1-p)^{k-1}\, p \quad \text{for } k = 1, 2, \ldots .$$

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.

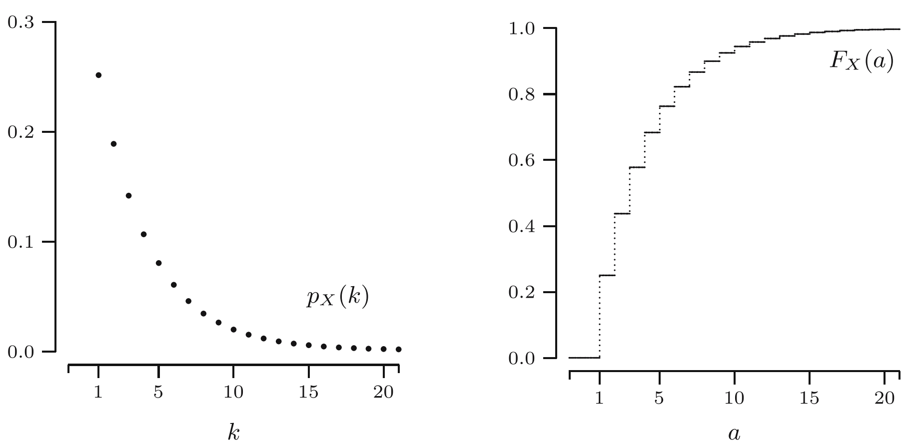


Fig. 4.3. Probability mass function and distribution function of the $Geo(\frac{1}{4})$ distribution.

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

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

$$\mathrm{P}(X > n + k \,|\, X > k) = \mathrm{P}(X > n)\,.$$

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

$$\begin{aligned}
\mathrm{P}(X > n + k \,|\, X > k) &= \frac{\mathrm{P}(\{X > k + n\} \cap \{X > k\})}{\mathrm{P}(X > k)} \\
&= \frac{\mathrm{P}(X > k + n)}{\mathrm{P}(X > k)} = \frac{(1-p)^{n+k}}{(1-p)^k} \\
&= (1-p)^n = \mathrm{P}(X > n)\,.
\end{aligned}$$

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
$\mathrm{P}(S = k)$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

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

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

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 $\mathrm{P}(R_1 = 1, R_2 = 0)$. Due to independence, this is equal to $\mathrm{P}(R_1 = 1)\,\mathrm{P}(R_2 = 0) = \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{16}$.

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

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

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 $\mathrm{P}(X = 2)$, $\mathrm{P}(X = 3)$, and $\mathrm{P}(X = 4)$.

b. Show that $\mathrm{P}(X = n) = (n - 1)p^2(1 - p)^{n-2}$ 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 $a \le b$,

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

The function f has to satisfy $f(x) \ge 0$ 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.

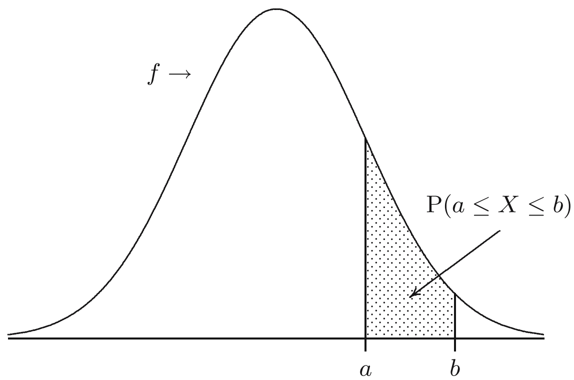


Fig. 5.1. Area under a probability density function f on the interval $[a, b]$.

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

$$\mathrm{P}(a - \varepsilon \leq X \leq a + \varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)\,\mathrm{d}x,$$

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

$$\mathrm{P}(X = a) = 0.$$

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

$$\mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X \leq b) = \mathrm{P}(a < X < b) = \mathrm{P}(a \leq X < b)\,.$$

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

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

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 $x \leq 0$ or $x \geq 1$, 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}.

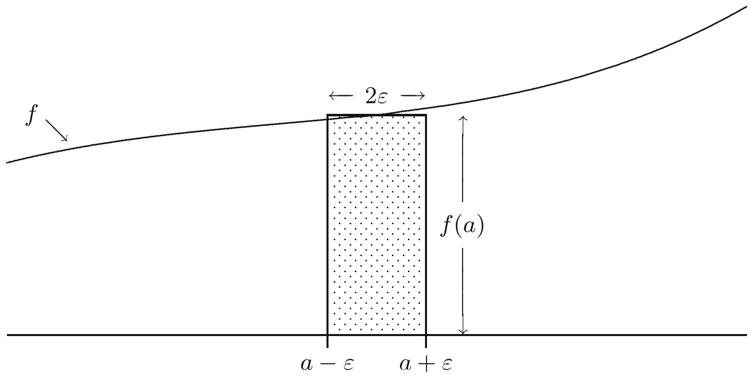


Fig. 5.2. Approximating the probability that X lies ε-close to a.

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.

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

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

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

Quick exercise 5.2 Compute for the darts example the probability that $0 < X \leq r/2$, and the probability that $r/2 < X \leq r$.

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

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

$F(x)=P(U<x)=x$ for $U(0,1)$

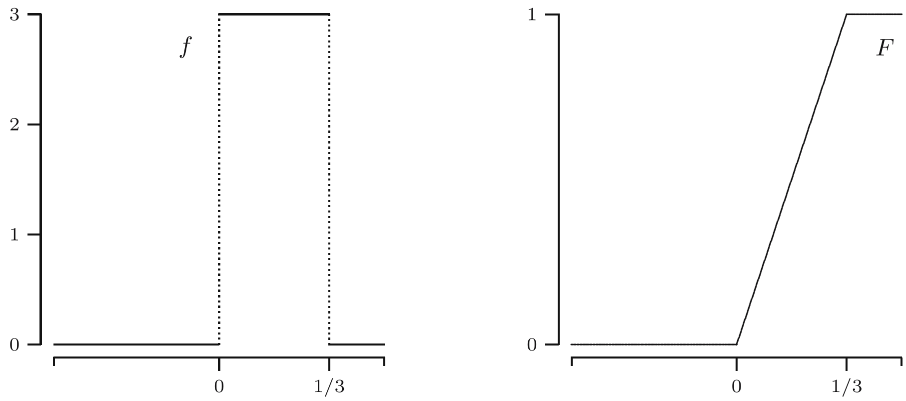


Fig. 5.3. The probability density function and the distribution function of the $U(0, \frac{1}{3})$ distribution.

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 $\mathrm{P}(T > t)$ 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)

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

It follows that the distribution function of T equals $1 - \mathrm{e}^{-\frac{v}{V}t}$, 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

$$f(x) = \lambda \mathrm{e}^{-\lambda x} \qquad \text{for } x \geq 0.$$

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.

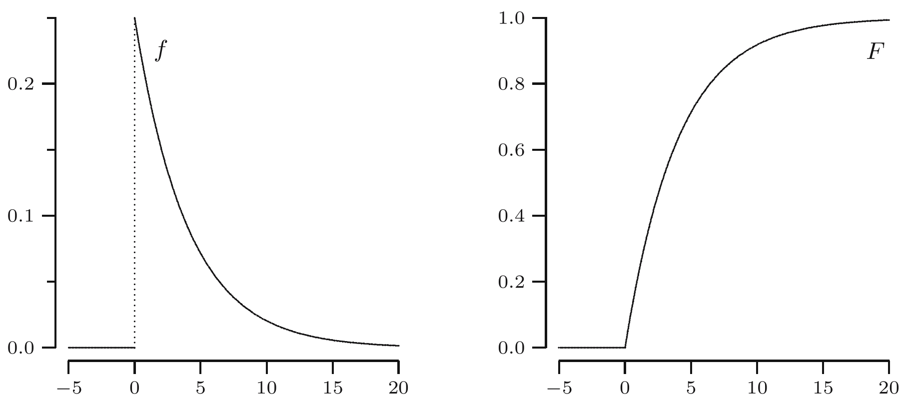


Fig. 5.4. The probability density and the distribution function of the $Exp(0.25)$ distribution.

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

$$f(x) = \frac{\alpha}{x^{\alpha+1}} \qquad \text{for } x \geq 1.$$

We denote this distribution by $Par(\alpha)$.

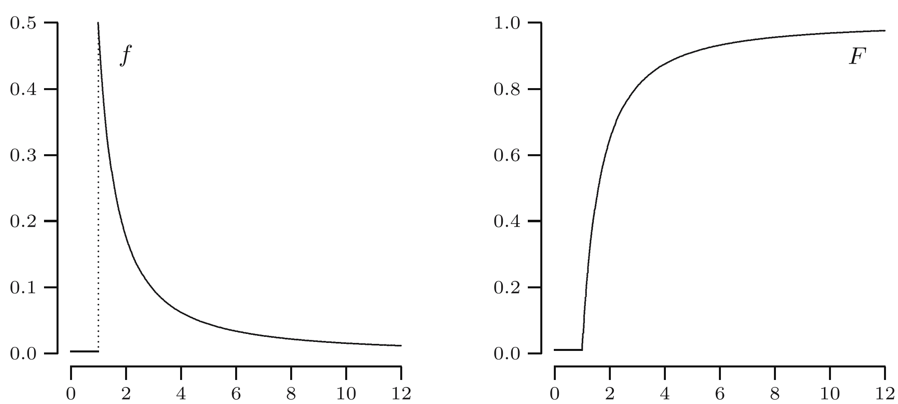


Fig. 5.5. The probability density and the distribution function of the $Par(0.5)$ distribution.

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

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{for } -\infty < x < \infty.$$

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.

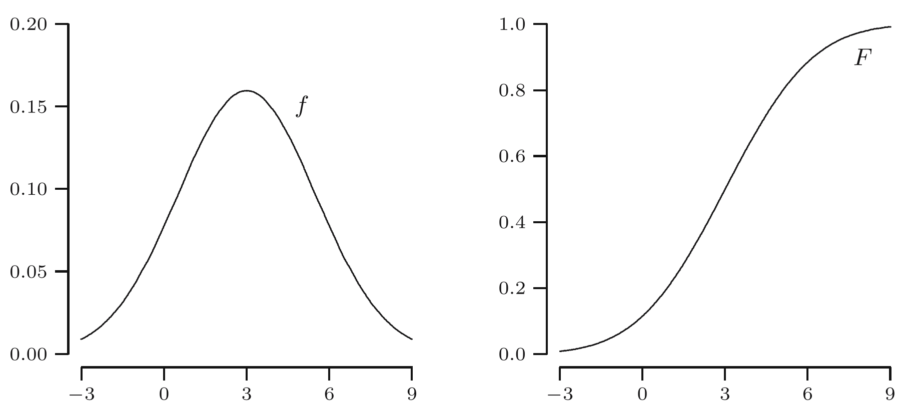


Fig. 5.6. The probability density and the distribution function of the $N(3, 6.25)$ distribution.

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

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

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 100pth *percentile* of the distribution of X is the smallest number q_p such that

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

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.

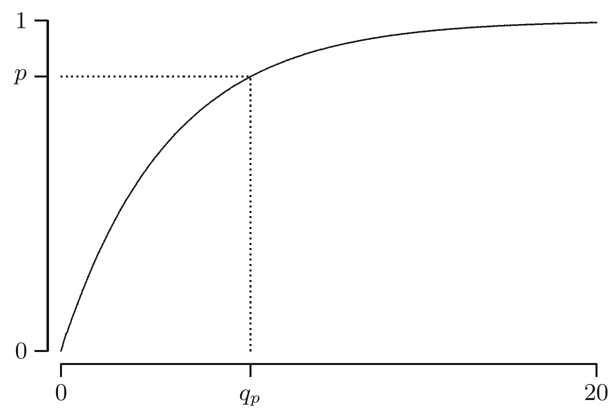


Fig. 5.7. The pth quantile q_p of the *Exp* (0.25) distribution.

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.

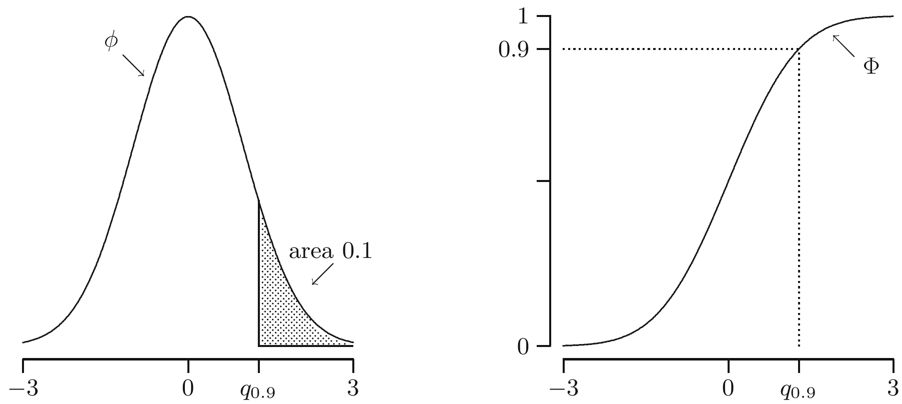


Fig. 5.8. The 90th percentile of the $N(0,1)$ distribution.

5.7 Solutions to the quick exercises

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

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

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

$$\begin{aligned} \mathrm{P}\bigl(10^{-4} \leq X \leq 10^{-2}\bigr) &= \int_{10^{-4}}^{10^{-2}} \frac{1}{2\sqrt{x}}\,\mathrm{d}x \\ &= \sqrt{10^{-2}} - \sqrt{10^{-4}} = 10^{-1} - 10^{-2} = 0.09. \end{aligned}$$

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

5.2 We have $\mathrm{P}(0 < X \leq r/2) = F(r/2) - F(0) = (1/2)^2 - 0^2 = 1/4$, and $\mathrm{P}(r/2 < X \leq r) = F(r) - F(r/2) = 1 - 1/4 = 3/4$, 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

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

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 $\mathrm{P}(X > 5)$. This equals

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

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 $\mathrm{P}(Z \leq a) > \mathrm{P}(Z \leq 0) = 0.5$.

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

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

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

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

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

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

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

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

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

a. Compute c.

b. Compute the distribution function of X.

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

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

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

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 $b \leq 0$, $G(b) = 1$ for $b \geq r$.

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 $A \leq 1/4$?

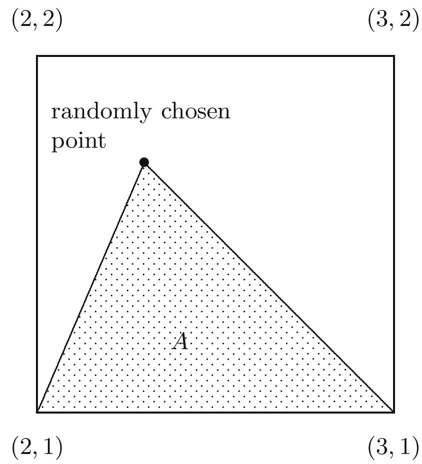


Fig. 5.9. A triangle in a square.

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.