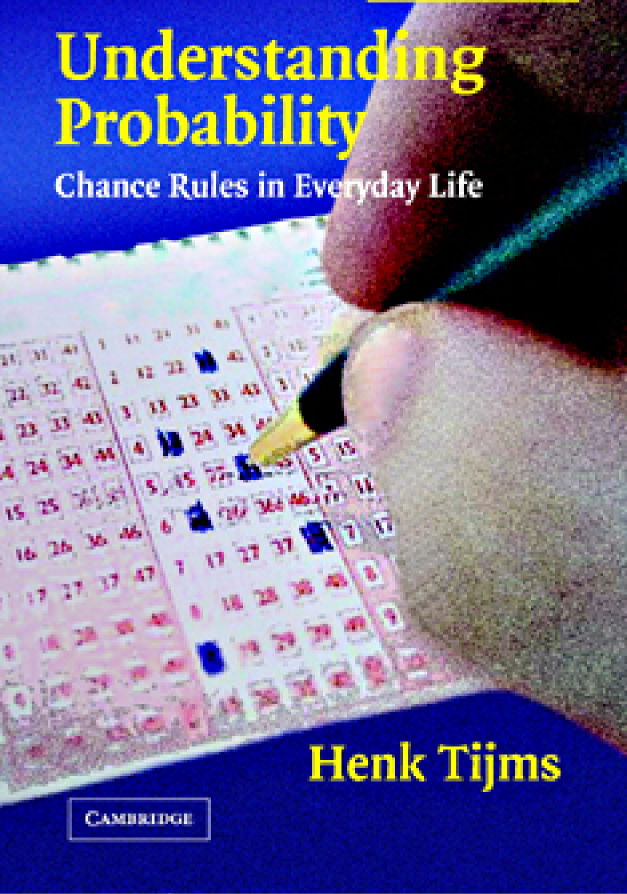
SECOND EDITION

Understanding Probability

Chance Rules in Everyday Life



Henk Tijms

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521701723

This page intentionally left blank

Understanding Probability

Chance events are commonplace in our daily lives. Every day we face situations where the result is uncertain, and, perhaps without realizing it, we guess about the likelihood of one outcome or another. Fortunately, mastering the concepts of probability can cast new light on situations where randomness and chance appear to rule.

In this fully revised second edition of *Understanding Probability*, the reader can learn about the world of probability in an appealing way. The author demystifies the law of large numbers, betting systems, random walks, the bootstrap, rare events, the central limit theorem, the Bayesian approach, and more.

This second edition has wider coverage, more explanations and examples and exercises, and a new chapter introducing Markov chains, making it a great choice for a first probability course. But its easy-going style makes it just as valuable if you want to learn about the subject on your own, and high school algebra is really all the mathematical background you need.

HENK TIJMS is Professor of Operations Research at the Vrije University in Amsterdam. The author of several textbooks, including *A First Course in Stochastic Models*, he is intensively active in the popularization of applied mathematics and probability in Dutch high schools. He has also written numerous papers on applied probability and stochastic optimization for international journals, including *Applied Probability* and *Probability in the Engineering and Informational Sciences*.

Understanding Probability

Chance Rules in Everyday Life

Second Edition

HENK TIJMS
Vrije University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521701723

First published in print format 2007

ISBN-13	978-0-511-34917-1	eBook (EBL)
ISBN-10	0-511-34917-3	eBook (EBL)
ISBN-13	978-0-521-70172-3	paperback
ISBN-10	0-521-70172-4	paperback

Contents

Preface

When I was a student, a class in topology made a great impression on me. The teacher asked me and my classmates not to take notes during the first hour of his lectures. In that hour, he explained ideas and concepts from topology in a nonrigorous, intuitive way. All we had to do was listen in order to grasp the concepts being introduced. In the second hour of the lecture, the material from the first hour was treated in a mathematically rigorous way and the students were allowed to take notes. I learned a lot from this approach of interweaving intuition and formal mathematics.

This book, about probability as it applies to our daily lives, is written very much in the same spirit. It introduces the reader to the world of probability in an informal way. It is not written in a theorem-proof style. Instead, it aims to teach the novice the concepts of probability through the use of motivating and insightful examples. In the book, no mathematics are introduced without specific examples and applications to motivate the theory. Instruction is driven by the need to answer questions about probability problems that are drawn from real-world contexts. Most of the book can easily be read by anyone who is not put off by a few numbers and some high school algebra. The informal yet precise style of the book makes it suited for classroom use, particularly when more self-activation is required from students. The book is organized into chapters that may be understood if read in a nonlinear order. The concepts and the ideas are laid out in the first part of the book, while the second part covers the mathematical background. In the second part of the book, I have chosen to give a short account of the mathematics of the subject by highlighting the essentials in about 200 pages, which I believe better contributes to the understanding of the student than a diffuse account of many more pages. The book can be used for a one-quarter or one-semester course in a wide range of disciplines ranging from social sciences to engineering. Also, it is an ideal book to use as a supplementary text in more mathematical treatments of probability.

The book distinguishes itself from other introductory probability texts by its emphasis on why probability works and how to apply it. Simulation in interaction with theory is the perfect instrument to clarify and to enliven the basic concepts of probability. For this reason, computer simulation is used to give the reader insights into such key concepts as the law of large numbers, which come to life through the results of many simulation trials. The law of large numbers and the central limit theorem are at the center of the book, with numerous examples based on these main themes. Many of the examples deal with lotteries and casino games. The examples help the reader develop a "feel for probabilities." Good exercises are an essential part of each textbook. Much care has been paid to collecting exercises that appeal to the understanding and creativity of the reader rather than requiring the reader to plug numbers into formulas. Several of the examples and exercises in this book are inspired by material from the website of "Chance News." This website contains a wealth of material on probability and statistics. Finally, the text is enlivened with cartoons combining chance and humor, which were supplied by www.cartoonstock.com.

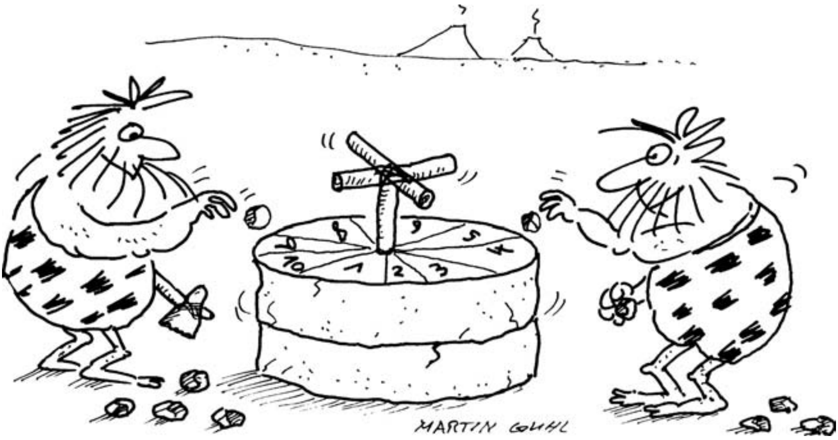
New to this edition

The first edition of the book was very well received, notably by people from outside the field of mathematics. Many readers expressed in their correspondence that they enjoyed the style of the book with its Parts One and Two, where the informal Part One motivates probabilistic thinking through many fascinating examples and problems from the real world and Part Two teaches the more formal mathematics. The comments and recommendations helped me to improve the book further. Part One has remained largely the same, but Part Two has been changed and expanded. The second part has been made self-contained for a first course in probability by adding more explanations and examples in almost every chapter. Also, the second part has been expanded by adding an introductory chapter on Markov chains, particularly suited for students in computer science and engineering. In the same style as the other chapters, the topic of Markov chains is taught by presenting interesting and realistic examples. A solutions manual containing solutions to all of the exercises was prepared for instructors. Finally, educational software supporting this book can be freely downloaded from http://staff.feweb.vu.nl/tijms.

Introduction

It is difficult to say who had a greater impact on the mobility of goods in the preindustrial economy: the inventor of the wheel or the crafter of the first pair of dice. One thing, however, is certain: the genius that designed the first random-number generator, like the inventor of the wheel, will very likely remain anonymous forever. We do know that the first dice-like exemplars were made a very long time ago. Excavations in the Middle East and in India reveal that dice were already in use at least fourteen centuries before Christ. Earlier still, around 3500 B.C., a board game existed in Egypt in which players tossed four-sided sheep bones. Known as the *astragalus*, this precursor to the modern-day die remained in use right up to the Middle Ages.

In the sixteenth century, the game of dice, or craps as we might call it today, was subjected for the first time to a formal mathematical study by the Italian mathematician and physician Gerolamo Cardano (1501–1576). An ardent gambler, Cardano wrote a handbook for gamblers entitled *Liber de Ludo Aleae* (The



Book of Games of Chance) about probabilities in games of chance. Cardano originated and introduced the concept of the set of outcomes of an experiment, and for cases in which all outcomes are equally probable, he defined the probability of any one event occurring as the ratio of the number of favorable outcomes and the total number of possible outcomes. This may seem obvious today, but in Cardano's day such an approach marked an enormous leap forward in the development of probability theory. This approach, along with a correct counting of the number of possible outcomes, gave the famous astronomer and physicist Galileo Galilei the tools he needed to explain to the Grand Duke of Tuscany, his benefactor, why it is that when you toss three dice, the chance of the sum being 10 is greater than the chance of the sum being 9 (the probabilities are $\frac{27}{216}$ and $\frac{25}{216}$, respectively).

By the end of the seventeenth century, the Dutch astronomer Christiaan Huygens (1625–1695) laid the foundation for current probability theory. His text *Van Rekeningh in Spelen van Geluck* (On Reasoning in Games of Chance), published in 1660, had enormous influence on later developments in probability theory (this text had already been translated into Latin under the title *De Ratiociniis de Ludo Aleae* in 1657). It was Huygens who originally introduced the concept of expected value, which plays such an important role in probability theory. His work unified various problems that had been solved earlier by the famous French mathematicians Pierre Fermat and Blaise Pascal. Among these was the interesting problem of how two players in a game of chance should divide the stakes if the game ends prematurely. Huygens' work led the field for many years until, in 1713, the Swiss mathematician Jakob Bernoulli (1654–1705) published *Ars Conjectandi* (The Art of Conjecturing) in which he presented the first general theory for calculating probabilities. Then, in 1812, the great French mathematician Pierre Simon Laplace (1749–1827) published his *Théorie Analytique des Probabilités*. This book unquestionably represents the greatest contribution in the history of probability theory.

Fermat and Pascal established the basic principles of probability in their brief correspondence during the summer of 1654, in which they considered some of the specific problems of odds calculation that had been posed to them by gambling acquaintances. One of the more well known of these problems is that of the Chevalier de Méré, who claimed to have discovered a contradiction in arithmetic. De Méré knew that it was advantageous to wager that a six would be rolled at least one time in four rolls of one die, but his experience as gambler taught him that it was not advantageous to wager on a double six being rolled at least one time in 24 rolls of a pair of dice. He argued that there were six possible outcomes for the toss of a single die and 36 possible outcomes for the toss of a pair of dice, and he claimed that this evidenced a contradiction to the arithmetic law of proportions, which says that the ratio of 4 to 6 should be the same as 24 to 36. De Méré turned to Pascal, who showed him with a few simple calculations that probability does not follow the law of proportions, as De Méré had mistakenly assumed (by De Méré's logic, the probability of at least one head in two tosses of a fair coin would be $2 \times 0.5 = 1$, which we know cannot be true). In any case, De Méré must have been an ardent player in order to have established empirically that the probability of rolling at least one double six in 24 rolls of a pair of dice lies just under one-half. The precise value of this probability is 0.4914. The probability of rolling at least one six in four rolls of a single die can be calculated as 0.5177. Incidentally, you may find it surprising that four rolls of a die are required, rather than three, in order to have about an equal chance of rolling at least one six.

Modern probability theory

Although probability theory was initially the product of questions posed by gamblers about their odds in the various games of chance, in its modern form, it has far outgrown any boundaries associated with the gaming room. These days, probability theory plays an increasingly greater roll in many fields. Countless problems in our daily lives call for a probabilistic approach. In many cases, better judicial and medical decisions result from an elementary knowledge of probability theory. It is essential to the field of insurance.[†] And likewise, the stock market, "the largest casino in the world," cannot do without it. The telephone network with its randomly fluctuating load could not have been economically designed without the aid of probability theory. Call-centers and airline companies apply probability theory to determine how many telephone lines and service desks will be needed based on expected demand. Probability theory is also essential in stock control to find a balance between the stock-out probability and the costs of holding inventories in an environment of uncertain demand. Engineers use probability theory when constructing dikes to calculate the probability of water levels exceeding their margins; this gives them the information they need to determine optimum dike elevation. These examples underline the extent to which the theory of probability has become an integral part of our lives. Laplace was right when he wrote almost 200 years ago in his Théorie Analytique des Probabilités:

> The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel

[†] Actuarial scientists have been contributing to the development of probability theory since its early stages. Also, astronomers have played very important roles in the development of probability theory.

> with a sort of instinct for which ofttimes they are unable to account. . . . It teaches us to avoid the illusions which often mislead us; . . . there is no science more worthy of our contemplations nor a more useful one for admission to our system of public education.

Probability theory and simulation

In terms of practical range, probability theory is comparable with geometry; both are branches of applied mathematics that are directly linked with the problems of daily life. But while pretty much anyone can call up a natural feel for geometry to some extent, many people clearly have trouble with the development of a good intuition for probability. Probability and intuition do not always agree. In no other branch of mathematics is it so easy to make mistakes as in probability theory. The development of the foundations of probability theory took a long time and went accompanied with ups and downs. The reader facing difficulties in grasping the concepts of probability theory might find comfort in the idea that even the genius Gottfried von Leibniz (1646–1716), the inventor of differential and integral calculus along with Newton, had difficulties in calculating the probability of throwing 11 with one throw of two dice. Probability theory is a difficult subject to get a good grasp of, especially in a formal framework. The computer offers excellent possibilities for acquiring a better understanding of the basic ideas of probability theory by means of simulation. With computer simulation, a concrete probability situation can be imitated on the computer. The simulated results can then be shown graphically on the screen. The graphic clarity offered by such a computer simulation makes it an especially suitable means to acquiring a better feel for probability. Not only a didactic aid, computer simulation is also a practical tool for tackling probability problems that are too complicated for scientific solution. Computer simulation, for example, has made it possible to develop winning strategies in the game of blackjack.

An outline

Part One of the book comprises Chapters 1–6. These chapters introduce the reader to the basic concepts of probability theory by using motivating examples to illustrate the concepts. A “feel for probabilities” is first developed through examples that endeavor to bring out the essence of probability in a compelling way. Simulation is a perfect aid in this undertaking of providing insight into the hows and whys of probability theory. We will use computer simulation, when needed, to illustrate subtle issues. The two pillars of probability theory, namely, the *law of large numbers* and the *central limit theorem* receive in-depth treatment. The nature of these two laws is best illustrated through the coin-toss experiment. The law of large numbers says that the percentage of tosses to come out heads will be as close to 0.5 as you can imagine provided that the coin is tossed often enough. How often the coin must be tossed in order to reach a prespecified precision for the percentage can be identified with the central limit theorem.

In Chapter 1, readers first encounter a series of intriguing problems to test their feel for probabilities. These problems will all be solved in the ensuing chapters. In Chapter 2, the law of large numbers provides the central theme. This law makes a connection between the probability of an event in an experiment and the relative frequency with which this event will occur when the experiment is repeated a very large number of times. Formulated by the aforementioned Jakob Bernoulli, the law of large numbers forms the theoretical foundation under the experimental determination of probability by means of computer simulation. The law of large numbers is clearly illuminated by the repeated coin-toss experiment, which is discussed in detail in Chapter 2. Astonishing results hold true in this simple experiment, and these results blow holes in many a mythical assumption, such as the "hot hand" in basketball. One remarkable application of the law of large numbers can be seen in the Kelly formula, a betting formula that can provide insight for the making of horse racing and investment decisions alike. The basic principles of computer simulation will also be discussed in Chapter 2, with emphasis on the subject of how random numbers can be generated on the computer.

In Chapter 3, we will tackle a number of realistic probability problems. Each problem will undergo two treatments, the first one being based on computer simulation and the second bearing the marks of a theoretical approach. Lotteries and casino games are sources of inspiration for some of the problems in Chapter 3.

The binomial distribution, the Poisson distribution, and the hypergeometric distribution are the subjects of Chapter 4. We will discuss which of these important probability distributions applies to which probability situations, and we will take a look into the practical importance of the distributions. Once again, we look to the lotteries to provide us with instructional and entertaining examples. We will see, in particular, how important the sometimes underestimated Poisson distribution, named after the French mathematician Siméon-Denis Poisson (1781–1840), really is.

In Chapter 5, two more fundamental principles of probability theory and statistics will be introduced: the central limit theorem and the normal distribution with its bell-shaped probability curve. The central limit theorem is by far the most important product of probability theory. The names of the mathematicians Abraham de Moivre and Pierre Simon Laplace are inseparably linked to this theorem and to the normal distribution. De Moivre discovered the normal distribution around 1730.[†] An explanation of the frequent occurrence of this distribution is provided by the central limit theorem. This theorem states that data influenced by many small and unrelated random effects are approximately normally distributed. It has been empirically observed that various natural phenomena, such as the heights of individuals, intelligence scores, the luminosity of stars, and daily returns of the S&P, follow approximately a normal distribution. The normal curve is also indispensable in quantum theory in physics. It describes the statistical behavior of huge numbers of atoms or electrons. A great many statistical methods are based on the central limit theorem. For one thing, this theorem makes it possible for us to evaluate how (im)probable certain deviations from the expected value are. For example, is the claim that heads came up 5,250 times in 10,000 tosses of a fair coin credible? What are the margins of errors in the predictions of election polls? The standard deviation concept plays a key roll in the answering of these questions. We devote considerable attention to this fundamental concept, particularly in the context of investment issues. At the same time, we also demonstrate in Chapter 5, with the help of the central limit theorem, how confidence intervals for the outcomes of simulation studies can be constructed. The standard deviation concept also comes into play here. The central limit theorem will also be used to link the random walk model with the Brownian motion model. These models, which are used to describe the behavior of a randomly moving object, are among the most useful probability models in science. Applications in finance will be discussed, including the Black-Scholes formula for the pricing of options.

The probability tree concept is discussed in Chapter 6. For situations where the possibility of an uncertain outcome exists in successive phases, a probability tree can be made to systematically show what all of the possible paths are. Various applications of the probability tree concept will be considered, including the famous Monty Hall dilemma and the test paradox. In addition, we will also look at the Bayes formula in Chapter 6. This formula is a descriptive rule for revising probabilities in light of new information. Among other things, the Bayes rule is used in legal argumentation and in formulating medical diagnoses for specific illnesses. This eighteenth century formula, constructed by the English clergyman Thomas Bayes (1702–1761), laid the foundation for a separate branch of statistics, namely Bayesian statistics. Bayesian probability theory is historically

[†] The French-born Abraham de Moivre (1667–1754) lived most of his life in England. The protestant de Moivre left France in 1688 to escape religious persecution. He was a good friend of Isaac Newton and supported himself by calculating odds for gamblers and insurers and by giving private lessons to students.

the original approach to statistics, predating what is nowadays called classical statistics by a century. Astronomers have contributed much to Bayesian probability theory. In Bayesian probability one typically deals with nonrepeatable chance experiments. Astronomers cannot do experiments on the universe and thus have to make probabilistic inferences from evidence left behind. This is very much the same situation as in forensic science, in which Bayesian probability plays a very important role as well.

Part Two of the book is along the lines of a classical textbook and comprises Chapters 7–15. These chapters are intended for the more mathematically oriented reader. Chapter 7 goes more deeply into the axioms and rules of probability theory. In Chapter 8, the concept of conditional probability and the nature of Bayesian analysis are delved into more deeply. Properties of the expected value are discussed in Chapter 9. Chapter 10 gives an explanation of continuous distributions, always a difficult concept for the beginner to absorb, and provides insight into the most important probability densities. Whereas Chapter 10 deals with the probability distribution of a single random variable, Chapter 11 discusses joint probability distributions for two or more dependent random variables. The multivariate normal distribution is the most important joint probability distribution and is the subject of Chapter 12. Chapter 13 deals with conditional distributions and discusses the law of conditional expectations. In Chapter 14, we deal with the method of moment-generating functions. This powerful method enables us to analyze many applied probability problems. Also, the method is used to provide proofs for the strong law of large numbers and the central limit theorem. In the final Chapter 15, we introduce a random process, known as a Markov chain, which can be used to model many real-world systems that evolve dynamically in time in a random environment.

PART ONE

Probability in action

1
Probability questions

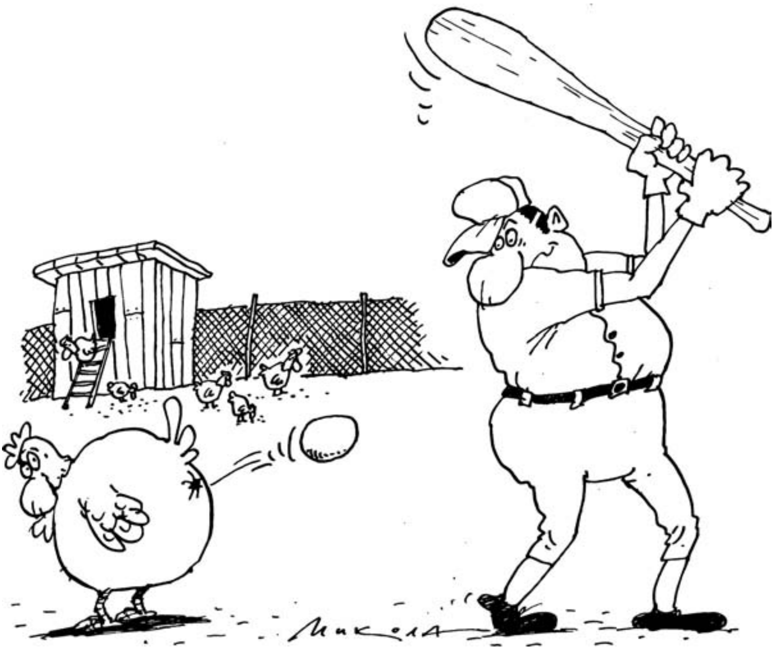
In this chapter, we provide a number of probability problems that challenge the reader to test his or her feeling for probabilities. As stated in the Introduction, it is possible to fall wide of the mark when using intuitive reasoning to calculate a probability, or to estimate the order of magnitude of a probability. To find out how you fare in this regard, it may be useful to try one or more of these 12 problems. They are playful in nature but are also illustrative of the surprises one can encounter in the solving of practical probability problems. Think carefully about each question before looking up its solution. All of the solutions to these problems can be found scattered throughout the ensuing chapters.

Question 1. A birthday problem (§3.1, §4.2.3)

You go with a friend to a football (soccer) game. The game involves 22 players of the two teams and one referee. Your friend wagers that, among these 23 persons on the field, at least two people will have birthdays on the same day. You will receive ten dollars from your friend if this is not the case. How much money should you, if the wager is to be a fair one, pay out to your friend if he is right?

Question 2. Probability of winning streaks (§2.1.3, §5.9.1)

A basketball player has a 50% success rate in free throw shots. Assuming that the outcomes of all free throws are independent from one another, what is the probability that, within a sequence of 20 shots, the player can score five baskets in a row?



Question 3. A scratch-and-win lottery (§4.2.3)

A scratch-and-win lottery dispenses 10,000 lottery tickets per week in Andorra and ten million in Spain. In both countries, demand exceeds supply. There are two numbers, composed of multiple digits, on every lottery ticket. One of these numbers is visible, and the other is covered by a layer of silver paint. The numbers on the 10,000 Andorran tickets are composed of four digits and the numbers on the ten million Spanish tickets are composed of seven digits. These numbers are randomly distributed over the quantity of lottery tickets, but in such a way that no two tickets display the same open or the same hidden number. The ticket holder wins a large cash prize if the number under the silver paint is revealed to be the same as the unpainted number on the ticket. Do you think the probability of at least one winner in the Andorran Lottery is significantly different from the probability of at least one winner in Spain? What is your estimate of the probability of a win occurring in each of the lotteries?

Question 4. A lotto problem (§4.2.3)

In each drawing of Lotto 6/45, six distinct numbers are drawn from the numbers $1, \ldots, 45$. In an analysis of 30 such lotto drawings, it was apparent that some numbers were never drawn. This is surprising. In total, $30 \times 6 = 180$ numbers were drawn, and it was expected that each of the 45 numbers would be chosen about four times. The question arises as to whether the lotto numbers were drawn according to the rules, and whether there may be some cheating occurring. What is the probability that, in 30 drawings, at least one of the numbers $1, \ldots, 45$ will not be drawn?

Question 5. Hitting the jackpot (Appendix)

Is the probability of hitting the jackpot (getting all six numbers right) in a 6/45 Lottery greater or lesser than the probability of throwing heads only in 22 tosses of a fair coin?

Question 6. Who is the murderer? (§8.2)

A murder is committed. The perpetrator is either one or the other of the two persons X and Y. Both persons are on the run from authorities, and after an initial investigation, both fugitives appear equally likely to be the perpetrator. Further investigation reveals that the actual perpetrator has blood type A. Ten percent of the population belongs to the group having this blood type. Additional inquiry reveals that person X has blood type A, but offers no information concerning the blood type of person Y. What is your guess for the probability that person X is the perpetrator?

Question 7. A coincidence problem (§4.3)

Two people, perfect strangers to one another, both living in the same city of one million inhabitants, meet each other. Each has approximately 500 acquaintances in the city. Assuming that for each of the two people, the acquaintances represent a random sampling of the city's various population sectors, what is the probability of the two people having an acquaintance in common?

Question 8. A sock problem (Appendix)

You have taken ten different pairs of socks to the laundromat, and during the washing, six socks are lost. In the best-case scenario, you will still have seven matching pairs left. In the worst-case scenario, you will have four matching pairs left. Do you think the probabilities of these two scenarios differ greatly?

Question 9. A statistical test problem (§3.6)

Using one die and rolling it 1,200 times, someone claims to have rolled the points 1, 2, 3, 4, 5, and 6 for a respective total of 196, 202, 199, 198, 202, and 203 times. Do you believe that these outcomes are, indeed, the result of coincidence or do you think they are fabricated?

Question 10. The best-choice problem (§2.3)

Your friend proposes the following wager: 20 people are requested, independently of one another, to write a number on a piece of paper (the papers should be evenly sized). They may write any number they like, no matter how high.

You fold up the 20 pieces of paper and place them randomly onto a tabletop. Your friend opens the papers one by one. Each time he opens one, he must decide whether to stop with that one or go on to open another one. Your friend's task is to single out the paper displaying the highest number. Once a paper is opened, your friend cannot go back to any of the previously opened papers. He pays you one dollar if he does not identify the paper with the highest number on it, otherwise you pay him five dollars. Do you take the wager? If your answer is no, what would you say to a similar wager where 100 people were asked to write a number on a piece of paper and the stakes were one dollar from your friend for an incorrect guess against ten dollars from you if he guesses correctly?

Question 11. The Monty Hall dilemma (§6.1)

A game-show climax draws nigh. A drum-roll sounds. The game show host leads you to a wall with three closed doors. Behind one of the doors is the automobile of your dreams, and behind each of the other two is a can of dog food. The three doors all have even chances of hiding the automobile. The host, a trustworthy person who knows precisely what is behind each of the three doors, explains how the game will work. First, you will choose a door without opening it, knowing that after you have done so, the host will open one of the two remaining doors to reveal a can of dog food. When this has been done, you will be given the opportunity to switch doors; you will win whatever is behind the door you choose at this stage of the game. Do you raise your chances of winning the automobile by switching doors?

Question 12. A daughter-son problem (§2.9, §6.1)

You are told that a family, completely unknown to you, has two children and that one of these children is a daughter. Is the chance of the other child also being a daughter equal to $\frac{1}{2}$ or $\frac{1}{3}$? Are the chances altered if, aware of the fact that the family has two children only, you ring their doorbell and a daughter opens the door?

The psychology of probability intuition is a main feature of some of these problems. Consider the birthday problem: how large must a group of randomly chosen people be such that the probability of two people having birthdays on the same day will be at least 50%? The answer to this question is 23. Almost no one guesses this answer; most people name much larger numbers. The number 183 is very commonly suggested on the grounds that it represents half the number of days in a year. A similar misconception can be seen in the words of a lottery official regarding his lottery, in which a four-digit number was drawn daily from the 10,000 number sequence 0000, 0001, . . . , 9999. On the second anniversary of the lottery, the official deemed it highly improbable that any of the 10,000 possible numbers had been drawn two or more times in the last 625 drawings. He added that this could only be expected after approximately half of the 10,000 possible numbers had been drawn. The lottery official was wildly off the mark: the probability that some number will not be drawn two or more times in 625 drawings is inconceivably small and is of the order of magnitude of 10^{-9}. This probability can be calculated by looking at the problem as a "birthday problem" with 10,000 possible birthdays and a group of 625 people (see §3.1 in Chapter 3). Canadian lottery officials, likewise, had no knowledge of the birthday problem and its treacherous variants when they put this idea into play: they purchased 500 automobiles from nonclaimed prize monies, to be raffled off as bonus prizes among their 2.4 million registered subscribers. A computer chose the winners by selecting 500 subscriber numbers from a pool of 2.4 million registered numbers without regard for whether or not a given number had already appeared. The unsorted list of the 500 winning numbers was published and to the astonishment of lottery officials, one subscriber put in a claim for two automobiles. Unlike the probability of a given number being chosen two or more times, the probability of some number being chosen two or more times is not negligibly small in this case; it is in the neighborhood of 5%! The Monty Hall dilemma – which made it onto the front page of the *New York Times* in 1991 – is even more interesting in terms of the reactions it generates. Some people vehemently insist that it does not matter whether a player switches doors at the end of the game, whereas others confidently maintain that the player must switch. We will not give away the answer here, but suffice it to say that many a mathematics professor gets this one wrong. These types of examples demonstrate that, in situations of uncertainty, one needs rational methods in order to avoid mental pitfalls.[†] Probability theory provides us with these methods. In the chapters that follow, you will journey through the fascinating world of probability theory. This journey will not take you over familiar, well-trodden territory; it will provide you with interesting prospects.

[†] An interesting article on mistakes in reasoning in situations of uncertainty is K. McKean, "Decisions, decisions, . . . ," *Discover*, June 1985, 22–31. This article is inspired by the standard work of D. Kahneman, P. Slovic and A. Tversky, *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, 1982.

2

The law of large numbers and simulation

In the midst of a coin-tossing game, after seeing a long run of tails, we are often tempted to think that the chances that the next toss will be heads must be getting larger. Or, if we have rolled a die many times without seeing a six, we are sure that finally we will roll a six. These notions are known as the *gambler's fallacy*. Of course, it is a mistake to think that the previous tosses will influence the outcome of the next toss: a coin or die has no memory. With each new toss, each of the possible outcomes remains equally likely. Irregular patterns of heads and tails are even characteristic of tosses with a fair coin. Unexpectedly long runs of heads or tails can already occur with a relatively few number of tosses. To see five or six heads in a row in 20 tosses is not exceptional. It is the case, however, that as the number of tosses increases, the fractions of heads and tails should be about equal, but that is guaranteed only *in the long run*. In the theory of probability, this fact is known as *the law of large numbers*. Just as the name implies, this law only says something about the game after a large number of tosses. This law does not imply that the absolute difference between the numbers of heads and tails should oscillate close to zero. On the contrary. For games of chance, such as coin-tossing, it is even typical, as we shall see, that for long time periods, either heads or tails will remain constantly in the lead, with the absolute difference between the numbers of heads and tails tending to become larger and larger. The course of a game of chance, although eventually converging in an average sense, is a whimsical process. What else would you have expected?

In this chapter, *the law of large numbers* will play the central role. Together with the central limit theorem from Chapter 5, this law forms the fundamental basis for probability theory. With the use of some illustrative examples – especially coin-tossing – and the use of simulation of chance experiments on the computer, we hope to provide the reader with a better insight into the law of large numbers, and into what this law says, and does not say, about the properties of random processes. To clarify and illustrate probability concepts, the simulation approach has some advantages over the formal, purely theoretical approach: it allows us to almost *instantly* simulate chance experiments, and present the results in a clear and graphic form. A picture is worth a thousand words! In this chapter our first goal is to help the reader develop “a feel for probabilities.” Then, the theory will be gradually introduced to enable us to calculate probabilities in concrete situations, using a clear and systematic approach.

2.1 The law of large numbers for probabilities

Suppose that the weather statistics over the last 200 years show that, on average, it rained 7 of 30 days in June, with no definite pattern for which particular days it rained. Assuming things do not change, then the probability of rain on June 15 the following year has the numerical value $\frac{7}{30}$. In this case, the past relative frequency of rainy days in June is used to assign a numerical value to the probability of rain on a given day in June during the following year. Put another way, the so-called empirical law of large numbers suggests the choice of $\frac{7}{30}$ for the probability of rain on any given day. We can shed further light on this law by considering repeated tosses of a fair coin. If after each toss you observe the percentage of heads up to that point, then you will see that in the beginning this percentage can fluctuate considerably, but eventually it settles down near 50% as the number of tosses increases. In general, suppose that a certain chance experiment will be carried out a large number of times under exactly the same conditions, and in a way so that the repetitions of the experiment are independent of each other. Let A be a given event in the experiment. For example, A is the event that in a randomly selected group of 23 people, two or more people have the same birthday. The *relative frequency* of the event A in n repetitions of the experiment is defined as

$$f_n(A) = \frac{n(A)}{n},$$

where $n(A)$ is the number of times that event A occurred in the n repetitions of the experiment. The relative frequency is a number between 0 and 1. Intuitively, it is clear that

> **the relative frequency with which event A occurs will fluctuate less and less as time goes on, and will approach a limiting value as the number of repetitions increases without bound.**

This phenomenon is known as *the empirical law of large numbers*. Intuitively, we would like to define the probability of the occurrence of the event A in a single repetition of the experiment as the limiting number to which the relative frequency $f_n(A)$ converges as n increases. Introducing the notion of probability this way bypasses several rather serious obstacles. The most serious obstacle is that, for relative frequency, the standard meaning of the notion of a limit cannot be applied (because you cannot assume *a priori* that the limiting number will be the same each time). For the foundations of probability theory, a different approach is followed. The more formal treatise is based on the concepts of sample space and probability measure. A *sample space* of a chance experiment is a set of elements that is in a one-to-one correspondence with the set of all possible outcomes of the experiment. On the sample space, a so-called probability measure is defined that associates to each subset of the sample space a numerical probability. The probability measure must satisfy a number of basic principles (axioms), which we will go into in Section 2.2 and in Chapter 7. These principles are otherwise motivated by properties of relative frequency. After we accept that the relative frequency of an event gives a good approximation for the probability of the event, then it is reasonable to let probabilities satisfy the same relations as relative frequencies. From these basic principles, if theoretical results can be derived that agree with our experience in concrete probability situations, then we know that the basic principles chosen are reasonable. Indeed, the so-called *theoretical law of large numbers* can be derived from the basic principles of probability theory. This theoretical law makes mathematically precise what the empirical law of large numbers tries to express. The theoretical law of large numbers can best be understood in the context of a random process where a fair coin is tossed an unlimited number of times. An outcome of this random process can be described by an infinite sequence of H's and T's, recording whether a head or tail turns up with each toss. The symbol ω is used to designate an outcome of the random process. For each conceivable outcome ω, we define the number $K_n(\omega)$ as

$$K_n(\omega) = \text{the number of heads in the first } n \text{ tosses in outcome } \omega.$$

For example, with the outcome $\omega = (H, T, T, H, H, H, T, H, H, \ldots)$, we have $K_5(\omega) = 3$ and $K_8(\omega) = 5$. Intuitively, we expect that "nature" will guarantee that ω will satisfy

$$\lim_{n\to\infty} K_n(\omega)/n = 1/2.$$

There are many conceivable sequences ω for which $K_n(\omega)/n$ does not converge to $\frac{1}{2}$ as $n \to \infty$, such as sequences containing only a finite number of H's. Nevertheless, "nature" chooses only sequences ω for which there is convergence to $\frac{1}{2}$. The theoretical law of large numbers says that the set of outcomes for which $K_n(\omega)/n$ does not converge to $\frac{1}{2}$ as $n \to \infty$ is "negligibly small" in a certain measure-theoretic sense. In probability theory we say that the fraction of tosses that come up heads *converges with probability one* to the constant $\frac{1}{2}$ (see also Chapter 7).

To give a mathematical formulation of *the theoretical law of large numbers*, advanced mathematics is needed. In words, we can formulate this law as follows:

> **If a certain chance experiment is repeated an unlimited number of times under exactly the same conditions, and if the repetitions are independent of each other, then the fraction of times that a given event A occurs will converge with probability 1 to a number that is equal to the probability that A occurs in a single repetition of the experiment.**

This strong law of large numbers corresponds directly to our world of experience. This result is also the mathematical basis for the widespread application of computer simulations to solve practical probability problems. In these applications, the (unknown) probability of a given event in a chance experiment is estimated by the relative frequency of occurrence of the event in a large number of computer simulations of the experiment. The application of simulations is based on the elementary principles of probability; it is a powerful tool with which extremely complicated probability problems can be solved.

The mathematical basis for the theoretical (strong) law of large numbers was given for the first time by the famous Russian mathematician A.N. Kolmogorov in the twentieth century.[†] A so-called weak version of the law of large numbers had already been formulated several centuries earlier by the Swiss mathematician Jakob Bernoulli in his masterpiece *Ars Conjectandi* that was published posthumously in 1713. In that book, which was partially based on Christiaan Huygens' work, Bernoulli was the first to make the mathematical connection between the probability of an event and the relative frequency with which the event occurs. It is important to bear in mind that the law of large numbers says nothing about the outcome of a single experiment. But what can be predicted with 100% certainty from this law is the long-run behavior of the system in the hypothetical situation of an unlimited number of independent repetitions of the experiment. Not only is the method of computer simulation based on this fact, but also the profit-earning capacities of insurance companies and casinos are based on the strong law of large numbers.

2.1.1 Coin-tossing

How can you better illustrate the law of large numbers than with the experiment of tossing a coin? We will do this experiment for both fair and unfair coins. We let p designate the probability that one toss of the coin shows "heads." For a fair coin, clearly $p = \frac{1}{2}$. Define the variables

$K_n =$ the total number of heads that will appear in the first n tosses

and

$f_n =$ the relative frequency with which heads will appear in the first n tosses.

Clearly, it follows that $f_n = K_n/n$. Even more interesting than K_n is the variable $K_n - np$, the difference between the actual number of heads and the expected number of heads. Table 2.1 gives the simulated values of $K_n - np$ for 30,000 tosses of a coin for a number of intermediate values of n. This is done for both a fair coin ($p = \frac{1}{2}$) and for an unfair coin ($p = \frac{1}{6}$). The numbers in Table 2.1 are the outcome of a particular simulation study. Any other simulation study will produce different numbers. It is worthwhile to take a close look at the results in Table 2.1. You see that the realizations of the relative frequency, f_n, approach

[†] Andrey Nikolayevich Kolmogorov (1903–1987) was active in many fields of mathematics and is considered one of the greatest mathematicians of the twentieth century. He is credited with the axiomatic foundation of probability theory.

Table 2.1. *Results of coin-toss simulations.*

	Fair coin ($p = \frac{1}{2}$)		Unfair coin ($p = \frac{1}{6}$)	
n	$K_n - np$	f_n	$K_n - np$	f_n
10	1.0	0.6000	0.33	0.2000
25	1.5	0.5600	1.83	0.2400
50	2.0	0.5400	2.67	0.2200
100	2.0	0.5200	3.33	0.2040
250	1.0	0.5040	5.33	0.1880
500	−2.0	0.4960	4.67	0.1760
1,000	10.0	0.5100	−3.67	0.1630
2,500	12.0	0.5048	−15.67	0.1604
5,000	−9.0	0.4982	−5.33	0.1656
7,500	11.0	0.5015	21.00	0.1659
10,000	24.0	0.5024	−33.67	0.1633
15,000	40.0	0.5027	−85.00	0.1610
20,000	91.0	0.5045	−17.33	0.1658
25,000	64.0	0.5026	−30.67	0.1654
30,000	78.0	0.5026	−58.00	0.1647

the true value of the probability p in a rather irregular manner. This is a typical phenomenon (try it yourself with your own simulations!). You see the same sort of phenomenon in lists that lottery companies publish of the relative frequencies of the different numbers that have appeared in past drawings. Results like those in Table 2.1 make it clear that fluctuations in the relative frequencies of the numbers drawn are nothing other than "natural" turns of fortune. In Table 2.1, it also is striking that the relative frequency f_n converges more slowly to the true value of the probability p than most of us would expect intuitively. The smaller the value of p, the more simulation effort is needed to ensure that the empirical relative frequency is close to p. In Chapter 5, we will see that the simulation effort must be increased about a hundredfold in order to simulate an unknown probability with one extra decimal place of precision. Thus, in principle, you should be suspicious of simulation studies that consist of only a small number of simulation runs, especially if they deal with small probabilities!

2.1.2 Random walk

Let's go back to the experiment of the fair coin-toss. Many people mistakenly think that a number of tosses resulting in heads will be followed by a number of tosses resulting in tails, such that both heads and tails will turn up approximately the same number of times. In the world of gambling, many gamblers make use of a system that is based on keeping track of the number of heads and tails that turn up as a game progresses. This is often described as the *gambler's fallacy*. Alas, it is absolute folly to think that a system of this kind will help. A coin simply does not have a memory and will therefore exhibit no compensatory behavior. In order to stimulate participation in lotteries, lottery sponsors publish lists of so-called hot and cold numbers, recording the number of wins for each number and the number of drawings that have taken place since each number was last drawn as a winning number. Such a list is often great fun to see, but will be of no practical use whatsoever in the choosing of a number for a future drawing. Lottery balls have no memory and exhibit no compensatory behavior.

For example, suppose a fair coin is tossed 100 times, resulting in heads 60 times. In the next 100 tosses, the absolute difference between the numbers of heads and tails can increase, whereas the relative difference declines. This would be the case, for example, if the next 100 tosses were to result in heads 51 times. In the long run, "local clusters" of heads or tails are *absorbed* by the average. It is certain that the relative frequencies of heads and tails will be the same over the long run. There is simply no law of averages for the absolute difference between the numbers of heads and tails. Indeed, the absolute difference between the numbers of heads and tails tends to become larger as the number of tosses increases. This surprising fact can be convincingly demonstrated using computer simulation. The graph in Figure 2.1 describes the path of the *actual* number of heads turned up minus the *expected* number of heads when simulating 2,000 tosses of a fair coin. This process is called a *random walk*, based on the analogy of an indicator that moves one step higher if heads is thrown and one step lower otherwise. A little bit of experimentation will show you that results such as those shown in Figure 2.1 are not exceptional. On the contrary, in fair coin-tossing experiments, it is typical to find that, as the number of tosses increases, the fluctuations in the random walk become larger and larger and a return to the zero-level becomes less and less likely. Most likely you will see the actual difference in the number of heads and tails grow and grow. For instance, the chance of getting a split somewhere between 45 and 55 with 100 tosses is almost 73%. But for the difference of the number of heads and tails to be within a range of $+5$ to -5 after 10,000 tosses is much less likely, about 9%; and even quite unlikely, about 0.9%, after 1,000,000 tosses. The appearance of the growing fluctuations in the random walk can be clarified by looking at the central limit theorem, which will be discussed in Chapter 5. In that chapter, we demonstrate how the range of the difference between the actual number of heads and the expected number has a tendency to grow proportionally with $\sqrt{n}$ as n ($=$ the number of tosses) increases. This result is otherwise not in conflict

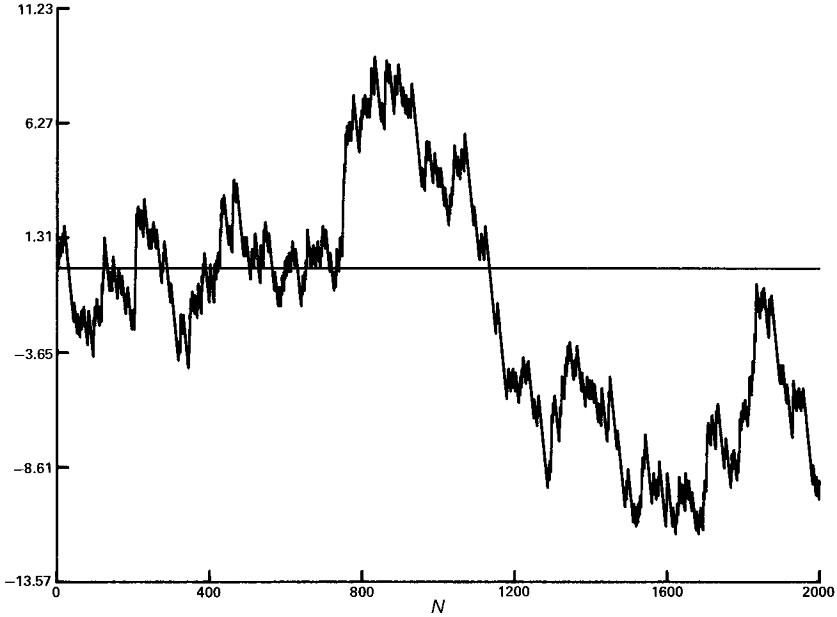


Fig. 2.1. A random walk of 2,000 coin tosses.

with the law of large numbers, which says that $\frac{1}{n}$ × (actual number of heads in n tosses minus $\frac{1}{2}n$) goes to 0 when $n \to \infty$. It will be seen in Section 5.8.1 that the probability distribution of the proportion of heads in n tosses becomes more and more concentrated around the 50 : 50 ratio as n increases and has the property that its spread around this ratio is on the order of $\frac{1}{\sqrt{n}}$.

2.1.3 The arc-sine law[†]

The random walk resulting from the repeated tossing of a fair coin is filled with surprises that clash with intuitive thinking. We have seen that the random walk exhibits ever-growing fluctuations and that it returns to zero less and less frequently. Another characteristic of the fair coin-toss that goes against intuition is that in the vast majority of cases, the random walk tends to occur on one side of the axis line. To be precise, suppose that the number of tosses to be done is *fixed* in advance. Intuitively, one would expect that the most likely value of the percentage of total time the random walk occurs on the positive side of the axis will be somewhere near 50%. But, quite the opposite is true, actually. This

[†] This specialized section may be omitted at first reading.

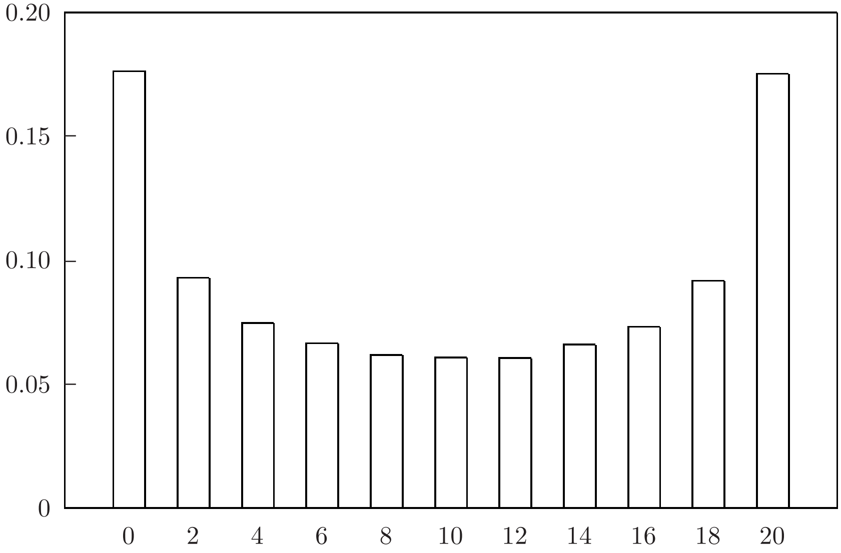


Fig. 2.2. Simulated distribution for 20 tosses.

is illustrated by the simulation results in Figure 2.2. For this figure, we have simulated 100,000 repetitions of a match between two players A and B. The match consists of a series of 20 tosses of a fair coin, where player A scores a point each time heads comes up and player B scores a point each time tails comes up. Figure 2.2 gives the simulated distribution of the number of times that player A is in the lead during a series of 20 tosses. The height of the bar on each base point k gives the simulated value of the probability that player A is k times in the lead during the 20 tosses. Here the following convention is made. If there is a tie after the final toss, the final toss gets assigned as leader the player who was in the lead after the penultimate toss (on the basis of this convention, the number of times that player A is in the lead is always even).

Looking at the simulation results, it appears that player A has a probability of 17.5% of being in the lead during the whole match. Put differently, the player in the lead after the first toss has approximately a 35% probability of remaining in the lead throughout the 20-toss match. In contrast to this is the approximately 6% probability that player A will lead for half of the match and player B will lead for the other half. This specific result in the case of 20 matches can be more generally supported by the *arc-sine law*, given here without proofs. If the number of tosses in a match between players A and B is *fixed in advance*, and if this number is sufficiently large, then the following approximation formula

Table 2.2. *Probability $P(\alpha, \beta)$ in the arc-sine law.*

(α, β)	$P(\alpha, \beta)$	(α, β)	$P(\alpha, \beta)$
$(0.50, 0.505)$	0.0064	$(0.995, 1)$	0.0901
$(0.50, 0.510)$	0.0127	$(0.990, 1)$	0.1275
$(0.50, 0.525)$	0.0318	$(0.975, 1)$	0.2022
$(0.50, 0.550)$	0.0638	$(0.950, 1)$	0.2871
$(0.50, 0.600)$	0.1282	$(0.900, 1)$	0.4097

holds true

$$P\,(\text{player } A \text{ is at least } 100x\% \text{ of time in the lead}) \approx 1 - \frac{2}{\pi}\arcsin(\sqrt{x})$$

for each x satisfying $0 < x < 1$. From this approximation formula, it can be deduced that, for all α, β with $\frac{1}{2} \leq \alpha < \beta < 1$, it is true that

$$P(\text{one of the two players is in the lead for somewhere between}\\ 100\alpha\% \text{ and } 100\beta\% \text{ of the time}) \approx \frac{4}{\pi}\left\{\arcsin(\sqrt{\beta}) - \arcsin(\sqrt{\alpha})\right\}.$$

Use $P(\alpha, \beta)$ to abbreviate $\frac{4}{\pi}\{\arcsin(\sqrt{\beta}) - \arcsin(\sqrt{\alpha})\}$. In Table 2.2 we give the value of $P(\alpha, \beta)$ for various values of α and β. The table shows that, in approximately one of five matches, one of the two players is in the lead for more than 97.5% of the time. It also shows that in one of 11 matches, one player is in the lead for more than 99.5% of the time. A fair coin, then, will regularly produce results that show no change in the lead for very long, continuous periods of time. Financial markets analysts would do well to keep these patterns in mind when analyzing financial markets. However controversial the assertion, some prominent economists claim that financial markets have no memory and behave according to a random walk. Their argument is quite simple: if share prices were predictable, then educated investors would buy low and sell high, but it would not be long before many others began to follow their lead, causing prices to adjust accordingly and to return to random behavior. This assertion is still extremely controversial because psychological factors (herd instinct) have a large influence on financial markets.[†]

Figure 2.2 and Table 2.2 demonstrate that the percentage of time of a random walk occurring on the positive side of the axis is much more likely to be near 0% or 100% than it is to be near the "expected" value of 50% (the assumption of a

[†] see also Richard H. Thaler, *The Winner's Curse, Paradoxes and Anomalies in Economic Life*, Princeton University Press, 1992.

predetermined number of steps is crucial for this fact). At first glance, most people cannot believe this to be the case. It is, however, true, and can be demonstrated with simulation experiments. These same simulations also demonstrate that the manner in which heads and tails switch off in a series of tosses with a fair coin is extremely irregular: surprisingly long series of heads or tails alone can occur. For example, in an experiment consisting of 20 tosses of a fair coin, simulation allows one to determine that the probability of a coin turning up heads five or more times in a row is approximately 25%, and that the probability of the coin landing on the same side, whether heads or tails, five or more times in a row is approximately 46%. On the grounds of this result, one need not be surprised if a basketball player with a free-throw success rate of 50% scores five or more baskets in a row in a series of 20 shots.

2.2 Basic probability concepts

This section deals with some of the fundamental theoretical concepts in probability theory. Using examples, these concepts will be introduced. The *sample space* of an experiment has already been defined as a set of elements that is in a one-to-one correspondence with the set of all possible outcomes of the experiment. Any subset of the sample space is called an *event*. That is, an event is a set consisting of possible outcomes of the experiment. If the outcome of the experiment is contained in the set E, it is said that the event E has occurred. A sample space in conjunction with a probability measure is called a *probability space*. A *probability measure* is simply a function P that assigns a numerical probability to each subset of the sample space. A probability measure must satisfy a number of consistency rules that will be discussed later.

Let's first illustrate a few things in light of an experiment that children sometimes use in their games to select one child out of the group. Three children simultaneously present their left or right fist to the group. If one of the children does not show the same fist as the other two, that child is "out." The sample space of this experiment can be described by the set $S = \{RRR, RRL, RLR, RLL, LLL, LLR, LRL, LRR\}$ consisting of eight elements, where $R(L)$ stands for a right (left) fist. The first letter of every element indicates which fist the first child shows, the second letter indicates the fist shown by the second child, and the third letter indicates the fist of the third child. If we assume that the children show the fists independently of one another, and each child chooses a fist randomly, then each of the outcomes is equally probable and we can assign a probability of $\frac{1}{8}$ to each outcome. The outcome subset $\{RRL, RLR, RLL, LLR, LRL, LRR\}$ corresponds with the event that one of the children is declared "out." We assign a probability of $\frac{6}{8}$ to this event.

Another interesting application is Efron's dice game. Let us first consider the situation of two players A and B each having a symmetric die. The six faces of the die of player A have the numbers $5, 5, 5, 1, 1, 1$ and the numbers on the six faces of the die of player B are $4, 4, 4, 4, 0, 0$. The players roll simultaneously their dice. What is the probability of A getting a higher number than B? To answer this question, we choose as sample space the set $\{(5, 4), (5, 0), (1, 4), (1, 0)\}$, where the first component of each outcome (i, j) indicates the score of player A and the second component indicates the score of player B. It is reasonable to assign the probability $\frac{1}{2} \times \frac{2}{3} = \frac{1}{3}$ to the outcome $(5, 4)$, the probability $\frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$ to the outcome $(5, 0)$, the probability $\frac{1}{2} \times \frac{2}{3} = \frac{1}{3}$ to the outcome $(1, 4)$ and the probability $\frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$ to the outcome $(1, 0)$. The subset $\{(5, 4), (5, 0), (1, 0)\}$ corresponds to the event that A gets a higher score than B. Thus, the probability of A beating B is $\frac{1}{3} + \frac{1}{6} + \frac{1}{6} = \frac{2}{3}$. In Efron's dice game, there are two other players C and D having the symmetric dice with the numbers $3, 3, 3, 3, 3, 3$ and $6, 6, 2, 2, 2, 2$, respectively. The probability of C beating D is $\frac{2}{3}$. Surprisingly enough, the probability of the underdog B of the players A and B beating the favorite C of the players C and D is $\frac{2}{3}$, and the probability of the favorite A of the players A and B beating the underdog D of the players C and D is $\frac{1}{3}$. It is left to the reader to verify this result. The result will not surprise sports enthusiasts.

2.2.1 Random variables

In many chance experiments, we are more interested in some function of the outcome of the chance experiment than in the actual outcomes. A *random variable* is simply a function that is defined on the sample space of the experiment and assigns a numerical value to each possible outcome of the experiment. For example, in the experiment that consists of tossing a fair coin three times, the random variable X could be defined as the number of times the coin turns up heads. Or, in the experiment consisting of the simultaneous rolling of a pair of dice, the random variable X could be defined as the sum of the values rolled, or as the greater of the two values rolled. The concept of random variable is always a difficult concept for the beginner. For an intuitive understanding, the best way is to view a random variable as a variable that takes on its values by chance. A random variable gets its value only after the underlying chance experiment has been performed. It is common to use uppercase letters such as X, Y, and Z to denote random variables, and lowercase letters x, y, and

z to denote their possible numerical values. In many applications the random variable X can take on only a finite number of possible values or values from a countably infinite set, such as the set of all nonnegative integers. In such a case, the random variable X is said to be a *discrete* random variable. In the first part of this book, we are mainly concerned with discrete random variables that take on a finite number of values. Let us assume that X can only take on values from the finite set $I = \{x_1, \ldots, x_M\}$. The event $\{X = x_j\}$ is defined as the set of those outcomes for which the random variable X takes on the value x_j. The probability of the event $\{X = x_j\}$ is thus defined as the sum of the probabilities of the individual outcomes for which X takes on the value x_j. This probability is denoted by $P(X = x_j)$. The function $p_j = P(X = x_j)$ for $j = 1, \ldots, M$ is called the *probability mass function* of X. The possible values $x_1, \ldots, x_M$ are called mass points of X.

Example 2.1 John and Mary each roll one die. What is the probability mass function of the largest of the two scores?

Solution. Let the random variable X denote the largest of the two scores. This random variable has $I = \{1, \ldots, 6\}$ as its set of possible values. To find the distribution of X, you will need the sample space of the experiment. A logical choice is the set

$$S = \{(1, 1), \ldots, (1, 6), (2, 1), \ldots, (6, 1), \ldots, (6, 6)\},$$

where the outcome (i, j) corresponds with the event that the score of John is i dots and the score of Mary is j dots. Each of the 36 possible outcomes is equally probable with fair dice. One translates this fact by assigning an equal probability of $\frac{1}{36}$ to each outcome. The random variable X assumes the value $\max(i, j)$ for outcome (i, j). For example, X assumes the value 3 for each of the five outcomes $(1, 3), (3, 1), (2, 3), (3, 2)$ and $(3, 3)$. Consequently, $P(X = 3) = \frac{5}{36}$. In this way one finds

$$P(X = 1) = \frac{1}{36}, \quad P(X = 2) = \frac{3}{36}, \quad P(X = 3) = \frac{5}{36},$$
$$P(X = 4) = \frac{7}{36}, \quad P(X = 5) = \frac{9}{36}, \quad P(X = 6) = \frac{11}{36}.$$

In the following example we discuss an experiment for which not every element of the sample space is equally probable. This example involves a so-called *compound* experiment. A compound experiment is one that is based on a sequence of elementary experiments. When the outcomes of the elementary experiments are independent of one another, then the probabilities assigned in the compound experiment are based on the multiplication of the probabilities of the outcomes in the individual elementary experiments. The theoretical construct for this *product rule* is discussed in Chapter 7.

Example 2.2 Two desperados A and B are playing Russian roulette, and they have agreed that they will take turns pulling the trigger of a revolver with six cylinders and one bullet. This dangerous game ends when the trigger has been pulled six times without a fatal shot occurring (after each attempt the magazine is spun to a random position). Desperado A begins. What is the probability mass function of the number of times desperado A pulls the trigger?

Solution. The sample space for this experiment is taken as

$$S = \{F, GF, GGF, GGGF, GGGGF, GGGGGF, GGGGGG\},$$

where an F stands for an attempt resulting in a fatal shot, and G stands for an attempt that has a good ending. The results of the consecutive attempts are independent from one another. On these grounds, we will assign the probabilities

$$\frac{1}{6}, \frac{5}{6} \times \frac{1}{6}, \left(\frac{5}{6}\right)^2 \times \frac{1}{6}, \left(\frac{5}{6}\right)^3 \times \frac{1}{6}, \left(\frac{5}{6}\right)^4 \times \frac{1}{6}, \left(\frac{5}{6}\right)^5 \times \frac{1}{6} \text{ and } \left(\frac{5}{6}\right)^6$$

to the consecutive elements of the sample space. The random variable X will be defined as the number of times that desperado A pulls the trigger. The random variable X takes on the value 1 for outcomes F and GF, the value 2 for outcomes GGF and $GGGF$, and the value 3 for all other outcomes. This gives

$$P(X = 1) = \frac{1}{6} + \frac{5}{6} \times \frac{1}{6} = 0.30556,$$

$$P(X = 2) = \left(\frac{5}{6}\right)^2 \times \frac{1}{6} + \left(\frac{5}{6}\right)^3 \times \frac{1}{6} = 0.21219,$$

$$P(X = 3) = \left(\frac{5}{6}\right)^4 \times \frac{1}{6} + \left(\frac{5}{6}\right)^5 \times \frac{1}{6} + \left(\frac{5}{6}\right)^6 = 0.48225.$$

2.2.2 Probability in finite sample spaces

We constructed a probability model for the various situations occurring in the above examples. The ingredients necessary for the making of a model are a sample space and the probabilities assigned to the elements of the sample space. These ingredients are part of a translation process from a physical context into a mathematical framework. The probabilities assigned to the outcomes of the chance experiment do not just appear out of nowhere, we must choose them. Naturally, this must be done in such a way that the model is in agreement with reality. In most cases, when the experiment can be repeated infinitely under stable conditions, we have the empirical relative frequencies of the outcomes in mind along with the assignment of probabilities to the possible outcomes. For the case of a chance experiment with a *finite* sample space, it suffices to assign a probability to each individual element of the sample space. These elementary probabilities must naturally meet the requirement of being greater than or equal to 0 and adding up to 1. An event in the experiment corresponds with a subset in the sample space. It is said that an *event* A occurs when the outcome of the experiment belongs to the *subset* A of the sample space. A numerical value $P(A)$ is assigned to each subset A of the sample space. This numerical value $P(A)$ tells us how likely the event A is to occur. The probability function $P(A)$ is logically defined as

$P(A)$ **is the sum of the probabilities of the individual outcomes in the set** A**.**

For the special case in which all outcomes are equally probable, $P(A)$ is found by dividing the number of outcomes in set A by the total number of outcomes. The model with equally probable outcomes is often called the *Laplace model*. This basic model shows up naturally in many situations.

The function P that assigns a numerical probability $P(A)$ to each subset A of the sample space is called a *probability measure*. A sample space in conjunction with a probability measure is called a *probability space*. The probability measure P must satisfy the axioms of modern probability theory

Axiom 1. $P(A) \geq 0$ *for every event* A.
Axiom 2. $P(A) = 1$ *when* A *is equal to the sample space.*
Axiom 3. $P(A \cup B) = P(A) + P(B)$ *for disjoint events* A *and* B.

Events A and B are said to be *disjoint* when the subsets A and B have no common elements. It is important to keep in mind that these axioms only provide us with the conditions that the probabilities must satisfy; they do not tell us how to assign probabilities in concrete cases. They are either assigned on the basis of relative frequencies (as in a dice game) or on the grounds of subjective consideration (as in a horse race). In both of these cases, the axioms are natural conditions that must be satisfied. The third axiom says that the probability of event A or event B occurring is equal to the sum of the probability of event A and the probability of event B, when these two events cannot occur simultaneously.[†] In the case of a nonfinite sample space, the *addition rule* from the third axiom must be modified accordingly. Rather than going into the details of such a modification here, we would direct interested readers to Chapter 7. The beauty of mathematics can be

[†] The choice of the third axiom can be reasoned by the fact that relative frequency has the property $f_n(A \cup B) = f_n(A) + f_n(B)$ for disjoint events A and B, as one can directly see from the definition of relative frequency in Section 2.1 $\left(\frac{n(A)+n(B)}{n} = \frac{n(A)}{n} + \frac{n(B)}{n}\right)$.

seen in the fact that these simple axioms suffice to derive such profound results as the theoretical law of large numbers. Compare this with a similar situation in geometry, where simple axioms about points and lines are all it takes to establish some very handsome results.

To illustrate, take another look at the above Example 2.2. Define A as the event that desperado A dies with his boots on and B as the event that B dies with his boots on. Event A is given by $A = \{F, GGF, GGGGF\}$. This gives

$$P(A) = \frac{1}{6} + \left(\frac{5}{6}\right)^2 \frac{1}{6} + \left(\frac{5}{6}\right)^4 \frac{1}{6} = 0.3628.$$

Likewise, one also finds that $P(B) = 0.3023$. The probability $P(A \cup B)$ represents the probability that one of the two desperados will end up shooting himself. Events A and B are disjoint and so

$$P(A \cup B) = P(A) + P(B) = 0.6651.$$

2.3 Expected value and the law of large numbers

The concept of expected value was first introduced into probability theory by Christiaan Huygens in the seventeenth century. Huygens established this important concept in the context of a game of chance, and to gain a good understanding of precisely what the concept is, it helps to retrace Huygens' footsteps. Consider a casino game where the player has a 0.70 probability of losing 1 dollar and probabilities of 0.25 and 0.05 of winning 2 and 3 dollars, respectively. A player who plays this game a large number of times reasons intuitively as follows in order to determine the average win per game in n games. In approximately $0.70n$ repetitions of the game, the player loses 1 dollar per game and in approximately $0.25n$ and $0.05n$ repetitions of the game, the player wins 2 and 3 dollars, respectively. This means that the total win in dollars is approximately equal to

$$(0.70n) \times (-1) + (0.25n) \times 2 + (0.05n) \times 3 = -(0.05)n,$$

or the average win per game is approximately -0.05 dollars (meaning that the average "win" is actually a loss). If we define the random variable X as the win achieved in just a single repetition of the game, then the number -0.05 is said to be the expected value of X. The expected value of X is written as $E(X)$. In the casino game $E(X)$ is given by

$$E(X) = (-1) \times P(X = -1) + 2 \times P(X = 2) + 3 \times P(X = 3).$$

The general definition of expected value is reasoned out in the example above. Assume that X is a random variable with a discrete probability distribution $p_j = P(X = x_j)$ for $j = 1, \ldots, M$. The *expected value* or *expectation* of the random variable X is then defined by

$$E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_M p_M.$$

Invoking the commonly used summation sign $\sum$, we get

$$E(X) = \sum_{j=1}^{M} x_j p_j.$$

Stating this formula in words, $E(X)$ is a weighted average of the possible values that X could assume, where each value is weighted with the probability that X would assume the value in question. The term "expected value" can be misleading. It must not be confused with the "most probable value." An insurance agent who tells a 40-year-old person that he/she can expect to live another 37 years naturally means that you come up with 37 more years when you multiply the possible values of the person's future years with the corresponding probabilities and then add the products together. The expected value $E(X)$ is not restricted to values that the random variable X could possibly assume. For example, let X be the number of points accrued in one roll of a fair die. Then

$$E(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3\frac{1}{2}.$$

The value $3\frac{1}{2}$ can never be the outcome of a single roll with the die. When we are taking a very large number of rolls of the die, however, it does appear that the average value of the points will be close to $3\frac{1}{2}$. One can look into this empirical result intuitively with the law of large numbers for probabilities. This law teaches us that, when you have a very large number of rolls with a die, the fraction of rolls with j points is closely equal to $\frac{1}{6}$ for every $j = 1, \ldots, 6$. From here it follows that the average number of points per roll is close to $\frac{1}{6}(1 + 2 + \cdots + 6) = 3\frac{1}{2}$.

The empirical finding that the average value of points accrued in the rolls of a fair die gets ever closer to $3\frac{1}{2}$ as the number of rolls increases can be placed in a more general framework. Consider therefore a chance experiment that can be repeatedly performed under exactly the same conditions. Let X be a random variable that is defined on the probability space of the experiment. In order to keep the train of thought running smoothly, it is helpful to suppose that the experiment is a certain (casino) game and that X is the random payoff of the game. Suppose the game is carried out a large number of times under exactly the same conditions, and in a way such that the repetitions of the game are independent of each other. It would appear, then, that

> **the average payment per game will fluctuate less and less as time goes on, and will approach a limiting value as the number of repetitions of the game increases without bound.**

This empirical result has a mathematical counterpart that stems from probability theory axioms. If we define the random variable X_k as the payoff in the kth repetition of the game, then the *theoretical law of large numbers for expected value* can be stated as

> **the average payment $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ over the first n repetitions of the game will converge with probability 1 to a constant as $n \to \infty$ and this constant is equal to the expected value $E(X)$.**

Intuitively, under convergence with probability 1, "nature" assures that the random process of repeated games always produces a realization for which the long-term actual average payment per game assumes the numerical value $E(X)$ (see also Section 7.2). In many practical problems, it is helpful to interpret the expected value of a random variable as a long-term average. The law of large numbers justifies this intuitive interpretation.

Example 2.3 In the game "Unders and Overs" two dice are rolled and you can bet whether the total of the two dice will be under 7, over 7, or equal to 7.[†] The gambling table is divided into three sections marked as "Under 7," "7," and "Over 7." The payoff odds for a bet on "Under 7" are 1 to 1; for a bet on "Over 7," are 1 to 1; and for a bet on "7," are 4 to 1 (payoffs of r to 1 mean that you get $r + 1$ dollars back for each dollar bet if you win; otherwise, you lose your stake). Each player can put chips on one or more sections of the gambling table. Your strategy is to bet one chip on "Under 7" and one chip on "7" each time. What is your average win or loss per round if you play the game over and over?

Solution. Let the random variable X denote the number of chips you get back in any given round. The possible values of X are 0, 2, and 5. The random variable X is defined on the sample space consisting of the 36 equiprobable outcomes $(1, 1), (1, 2), \ldots, (6, 6)$. Outcome (i, j) means that i points turn up on the first die and j points on the second die. The total of the two dice is 7 for the six outcomes $(1, 6), (6, 1), (2, 5), (5, 2), (3, 4)$, and $(4, 3)$. Thus $P(X = 5) = \frac{6}{36}$.

[†] In the old days the game was often played at local schools in order to raise money for the school.

Similarly, $P(X = 0) = \frac{15}{36}$ and $P(X = 2) = \frac{15}{36}$. This gives

$$E(X) = 0 \times \frac{15}{36} + 2 \times \frac{15}{36} + 5 \times \frac{6}{36} = 1\frac{2}{3}.$$

You bet two chips each round. Thus, your average loss is $2 - 1\frac{2}{3} = \frac{1}{3}$ chip per round when you play the game over and over.

Expected value and risk

In the case that the random variable X is the random payoff in a game that can be repeated many times under identical conditions, the expected value of X is an informative measure on the grounds of the law of large numbers. However, the information provided by $E(X)$ is usually not sufficient when X is the random payoff in a nonrepeatable game. Suppose your investment has yielded a profit of \$3,000 and you must choose between the following two options: the first option is to take the sure profit of \$3,000 and the second option is to reinvest the profit of \$3,000 under the scenario that this profit increases to \$4,000 with probability 0.8 and is lost with probability 0.2. The expected profit of the second option is $0.8 \times \$4{,}000 + 0.2 \times \$0 = \$3{,}200$ and is larger than the \$3,000 from the first option. Nevertheless, most people would prefer the first option. The downside risk is too big for them. A measure that takes into account the aspect of risk is the variance of a random variable. This concept will be discussed in detail in Chapter 5.

2.3.1 Best-choice problem

In order to answer Question 10 from Chapter 1, you must know which strategy your friend is using to correctly identify the piece of paper with the largest number. Suppose your friend allows the first half of the papers to pass through his hands, but keeps a mental note of the highest number that appears. As he opens and discards the papers in the subsequent group, he stops at the appearance of the first paper showing a number higher than the one he took note of earlier. Of course, this paper will only appear if the ultimate highest number was not among the first ten papers opened. Let p represent the (unknown) probability that your friend will win the contest using this simple strategy. Imagine that you will have to pay five dollars to your friend if he wins and that otherwise, you receive one dollar. The expected value of your net win in a given contest is then

$$(1 - p) \times 1 - p \times 5 = 1 - 6p.$$

The contest is unfavorable to you if $p > \frac{1}{6}$. With a simple model not only can you show that this is the case, but also that p is actually greater than $\frac{1}{4}$. Now, try to visualize that the paper with the highest number has a 1 stamped on it in invisible ink, that the paper with the next-highest number has a 2 stamped on it, etc. Then imagine that the 20 pieces of paper are randomly lined up. The relative ranking of the numbers on the 20 papers corresponds to a permutation (ordered sequence) of the numbers $1, \ldots, 20$. This suggests a sample space consisting of all the possible permutations $(i_1, i_2, \ldots, i_{20})$ of the numbers $1, \ldots, 20$. The outcome $(i_1, i_2, \ldots, i_{20})$ corresponds to the situation in which i_1 is stamped in invisible ink on the outside of the first paper your friend chooses, i_2 on the second paper your friend chooses, etc. The total number of permutations of the integers $1, \ldots, 20$ is $20 \times 19 \times \ldots \times 1$. The notation $n!$ is used for the product $1 \times 2 \times \ldots \times n$ (see the Appendix). Thus, the sample space consists of 20! different elements. Each element is assigned the same probability $\frac{1}{20!}$. Let A represent the event that the second highest number is among the first ten papers, but that the highest number is not. In any case, your friend will win the contest if event A occurs. In order to find $P(A)$, one must count the number of elements $(i_1, i_2, \ldots, i_{20})$ where one of the numbers $i_1, \ldots, i_{10}$ is equal to 2 and one of the numbers $i_{11}, \ldots, i_{20}$ is equal to 1. This number is equal to $10 \times 10 \times 18!$. Thus,

$$P(A) = \frac{10 \times 10 \times 18!}{20!} = \frac{100}{19 \times 20} = 0.263.$$

The probability p that your friend will win the contest is greater than $P(A)$ and is then, indeed, greater than 25%. Using this reasoning you will also come to the same conclusion if 100 people or even one million people write down a random number on a piece of paper and your friend allows half of the pieces to go by without choosing one. Using computer simulation, it can be verified that the simple strategy of letting the first half of the pieces of paper go by gives your friend the probabilities 0.359 and 0.349 of winning when the number of people participating is 20 and 100, respectively. On the computer, the contest can be played out a great many times. You would take the fraction of contests won by your friend as an estimate for the probability p of your friend winning. In order to simulate the model on the computer, you need a procedure for generating a random permutation of the numbers $1, \ldots, n$ for a given value of n. Such a procedure is discussed in Section 2.9.

In Problem 3.24 of Chapter 3, we come back to the best-choice problem, and you may be surprised by the solution here. When we speak of n papers with n being high (say, $n \geq 100$), then the maximum probability of winning is approximately equal to $\frac{1}{e} = 0.368$, irrespective of the value of n. The optimal strategy is to open the first $\frac{n}{e}$ papers and then to choose the next paper to appear with a number higher than those contained in all of the previous papers. This strategy might guide you when you are looking for a restaurant in a city you visit for the first time! You sample for a while in order to improve your knowledge of what's available. The original version of the best-choice problem is the Sultan's dowry problem, which was first stated by Martin Gardner in 1960.

2.4 The drunkard's walk

The drunkard's walk is named for the drunkard exiting a pub who takes a step to the right with a probability of $\frac{1}{2}$ or a step to the left with a probability of $\frac{1}{2}$. Each successive step is executed independently of the others. The following questions arise: what is the probability that the drunkard will ever return to his point of origin, and what is the expected distance back to the point of origin after the drunkard has taken many steps? These questions seemingly fall into the category of pure entertainment, but, in actuality, nothing could be further from the truth. The drunkard's walk has many important applications in physics, chemistry, astronomy, and biology. These applications usually consider two- or three-dimensional representations of the drunkard's walk. The biologist looks at the transporting of molecules through cell walls. The physicist looks at the electrical resistance of a fixed particle. The chemist looks for explanations for the speed of chemical reactions. The climate specialist looks for evidence of global warming. The model of the drunkard's walk is extremely useful for this type of research.† We first look at the model of the drunkard walking along a straight line. Plotting the path of the drunkard's walk along a straight line is much the same as tracing the random walk of the fair-coin toss. Imagine a drunkard at his point of origin. His steps are of unit length, and there is a probability of $\frac{1}{2}$ that in any given step he will go to the right and a probability of $\frac{1}{2}$ that he will go to the left. The drunkard has no memory, i.e., the directions of the man's successive steps are independent of one another. Define the random variable D_m as

$$D_m = \text{the drunkard's distance from his point of origin after } m \text{ steps.}$$

Obviously, the quadratic distance D_m^2 is also a random variable. It holds that the expected value of the *quadratic* distance of the drunkard from his point of origin after m steps is given by

$$E\left(D_m^2\right) = m$$

for every value of m. A proof of this result will be outlined in Problem 9.18 in Chapter 9. For now, it is worth noting that the result does *not* allow us

† See G.H. Weiss, "Random walks and their applications," *American Scientist*, January-February 1983, **71**, 65–70.

to conclude that $E(D_m)$ is equal to $\sqrt{m}$, although this erroneous conclusion is often cited as true. Rather, the actual answer for $E(D_m)$ is that $\sqrt{m}$ must be amended by a factor of less than 1. For m large, this correction factor is approximately equal to 0.798. The following can then be said

$$E(D_m) \approx \sqrt{\frac{2}{\pi} m},$$

where the symbol $\approx$ stands for "is approximately equal to." This result will be explained in Section 5.8, with the help of the central limit theorem.

2.4.1 The drunkard's walk in higher dimensions

For the drunkard's walk on the two-dimensional plane, the expected value of the distance of the drunkard from his point of origin after taking m steps is given approximately by

$$E(D_m) \approx \frac{1}{2}\sqrt{\pi m}.$$

This approximation formula is applicable both in the case where the drunkard leaves from point (x, y) with equal probability $\frac{1}{4}$ towards each of the four bordering grid points $(x+1, y)$, $(x-1, y)$, $(x, y+1)$, and $(x, y-1)$ and in the case where the drunkard takes steps of unit length each time in a randomly chosen direction between 0 and 2π. The approximation formula for the drunkard's walk in three-dimensional space is

$$E(D_m) \approx \sqrt{\frac{8}{3\pi} m}.$$

We delve into these approximations further on in Chapter 12. The approximation for $E(D_m)$ has many applications. How long does it take a photon to travel from the Sun's core to its surface? The answer is that it takes approximately 10 million years, and it is found by using the model of the drunkard's walk. A photon has a countless number of collisions on its way to the Sun's surface. The distance traveled by a photon between two collisions can be measured as 6×10^{-6} mm. The Sun's radius measures 70,000 km. A photon travels at a speed of 300,000 km per second. Taking into consideration that 70,000 km is equal to 7×10^{10} mm, the equality

$$\sqrt{\frac{8}{3\pi} m} = \frac{7 \times 10^{10}}{6 \times 10^{-6}}$$

shows that the average number of collisions that a photon undergoes before reaching the Sun's surface is approximately equal to $m = 1.604 \times 10^{32}$. The speed of light is 300,000 km per second, meaning that the travel time of a photon between two collisions is equal to $(6 \times 10^{-6})/(3 \times 10^{11}) = 2 \times 10^{-17}$ seconds. The average travel time of a photon from the Sun's core to its surface is thus approximately equal to 3.208×10^{15} seconds. If you divide this by $365 \times 24 \times 3{,}600$, then you find that the average travel time is approximately 10 million years. A random walk is not a very fast way to get anywhere! Once it reaches the surface of the Sun, it takes a photon only 8 minutes to travel from the surface of the Sun to the Earth (the distance from the Sun to the Earth is 149,600,000 km).

2.4.2 The probability of returning to the point of origin

The drunkard's walk provides surprising results with regard to the probability of the drunkard returning to his point of origin if he keeps at it long enough. This probability is equal to 1 both for the drunkard's walk on the line and the drunkard's walk in two dimensions, but it is less than 1 for the drunkard's walk in the third dimension, assuming that the drunkard travels over a discrete grid of points. In the third dimension, the probability of ever returning to the point of origin is 0.3405.[†] To make it even more surprising, the drunkard will eventually visit every grid point with probability 1 in the dimensions 1 and 2, but the expected value of the number of necessary steps back to his point of origin is infinitely large. An advanced knowledge of probability theory is needed to verify the validity of these results.

2.5 The St. Petersburg paradox

In 1738 Daniel Bernoulli (1700–1782), one of the many mathematicians of the famous Bernoulli family, presented before the Imperial Academy of Sciences in St. Petersburg a classic paper on probability,[‡] in which he discussed the following problem. In a certain casino game, a fair coin is tossed successively until the moment that heads appears for the first time. The casino payoff is two dollars if heads turns up in the first toss, four dollars if heads turns up for the first time in the second toss, etc. In general, the payoff is 2^n dollars if heads turns up for the first time in the nth toss. Thus, with each additional toss the

[†] On Earth all roads lead to Rome, but not in space!

[‡] D. Bernoulli,"Specimen theoriae novae de mensura sortis," *Commentarii Academiae Scientiarum Imperalis Petropolitanea* V (1738): 175–192 (translated and republished as "Exposition of a new theory on the measurement of risk," *Econometrica* **22** (1954): 23–36).

payoff of the casino is doubled. What amount must the casino require the player to stake such that, over the long term, the game will not be a losing endeavor for the casino? To answer this question, we need to calculate the expected value of the casino payoff for a single repetition of the game. The probability of getting heads in the first toss is $\frac{1}{2}$, the probability of getting tails in the first toss and heads in the second toss is $\frac{1}{2} \times \frac{1}{2}$, etc., and the probability of getting tails in the first $n-1$ tosses and heads in the nth toss is $\left(\frac{1}{2}\right)^n$. The expected value of the casino payoff for a single repetition of the game is thus equal to

$$\frac{1}{2} \times \$2 + \frac{1}{4} \times \$4 + \cdots + \frac{1}{2^n} \times \$2^n + \cdots .$$

In this infinite series, a figure equal to \$1 is added to the sum each time. In this way, the sum exceeds every conceivable large value and mathematicians would say that the sum of the infinite series is infinitely large. The expected value of the casino payoff for a single repetition of the game is thus an infinitely large dollar amount. This means that casino owners should not allow this game to be played, whatever amount a player is willing to stake. However, no player in his right mind would be prepared to stake, say, 10 million dollars for the opportunity to play this game. The reality of the situation is that the game is simply not worth that much. In Bernoulli's day, a heated discussion grew up around this problem. Some of those involved even began to question whether there was not a problem with the mathematics. But no, the math was good, and the mathematical model for the game is calculated correctly. The trouble is that the model being used simply does not provide a good reflection of the actual situation in this case! For one thing, the model implicitly suggests that the casino is always in a position to pay out, whatever happens, even in the case of a great number of tosses being executed before the first heads appears, which adds up to a dazzlingly high payoff. The practical reality is that the casino is only in possession of a limited amount of capital and cannot pay out more than a limited amount. The paradox can be explained thus: if the mathematical model does not provide a good reflection of reality, the conclusion it forms will have no practical relevance.[†]

The problem does become more realistic when the following modification is made to the game. The casino can only pay out up to a limited amount. To simplify the matter, let's assume that the maximum payoff is a given multiple of 2. Let the maximum casino payoff per game be equal to 2^M dollars for some given integer M (e.g., $M = 15$ would correspond with a maximum payoff of

[†] An interesting discussion of the St. Petersburg paradox is given in the article "The St. Petersburg paradox and the crash of high-tech stocks in 2000," by G. Székely and D. Richards, in *The American Statistician*, **58** (2004): 225–231.

\$32,768). In every repetition of the game a fair coin is tossed until either heads appears for the first time or M tosses are executed without heads appearing. The casino pays the player 2^k dollars when heads appears for the first time in the kth toss and pays nothing if tails is tossed M times in a row. What must the player's minimum stake be such that the game will not be a loss for the casino over the long term? The same reasoning we used before says that the expected value of the casino payoff for a single execution of the game is equal to

$$\frac{1}{2} \times \$2 + \frac{1}{4} \times \$4 + \cdots + \frac{1}{2^M} \times \$2^M = \$M.$$

This means that the modified game is profitable for the casino if the player's stake is above M dollars. It is instructive to look at how the average payoff per game converges to the theoretical expected value M if the game is executed a great number of times. In Figure 2.3, we give the simulated results for 10,000 repetitions of the game both for $M = 10$ and $M = 20$. From these results it appears that as the value of M increases, many more plays are necessary before the average payoff per play converges to the theoretical expected value. The explanation for the slow convergence when M is large lies in the fact that very large payoffs occurring with a very small probability contribute a nonnegligible amount to the expected value. The simulation confirms this. In situations where a very small probability plays a nonnegligible role, very long simulations are required in order to get reliable estimates. The lesson to be gained here is that, in situations of this kind, it is especially dangerous to conclude results from simulations that are "too short." In addition, this underscores the importance of evaluating the reliability of results gained through the process of simulation. Such evaluation can only be achieved with help from a concept called the confidence interval, which will be discussed in Chapter 5.

2.6 Roulette and the law of large numbers

The law of large numbers is the basis for casino profits. If a sufficient number of players stake money at the gaming tables (and stake amounts are limited to a given maximum), then the casino will not operate at a loss and in fact will be ensured of steadily growing profits.[†]

[†] Blackjack (or twenty-one) is the only casino game in which the player has a theoretical advantage over the casino. Around 1960, computer-simulated winning blackjack strategies were developed. Casinos can be glad that these strategies are not only difficult to put into practice, but also provide only a small advantage to players. Players with large bankrolls attempting to use this system usually find either that small changes in game rules thwart their attempt or that they are simply escorted from the premises.

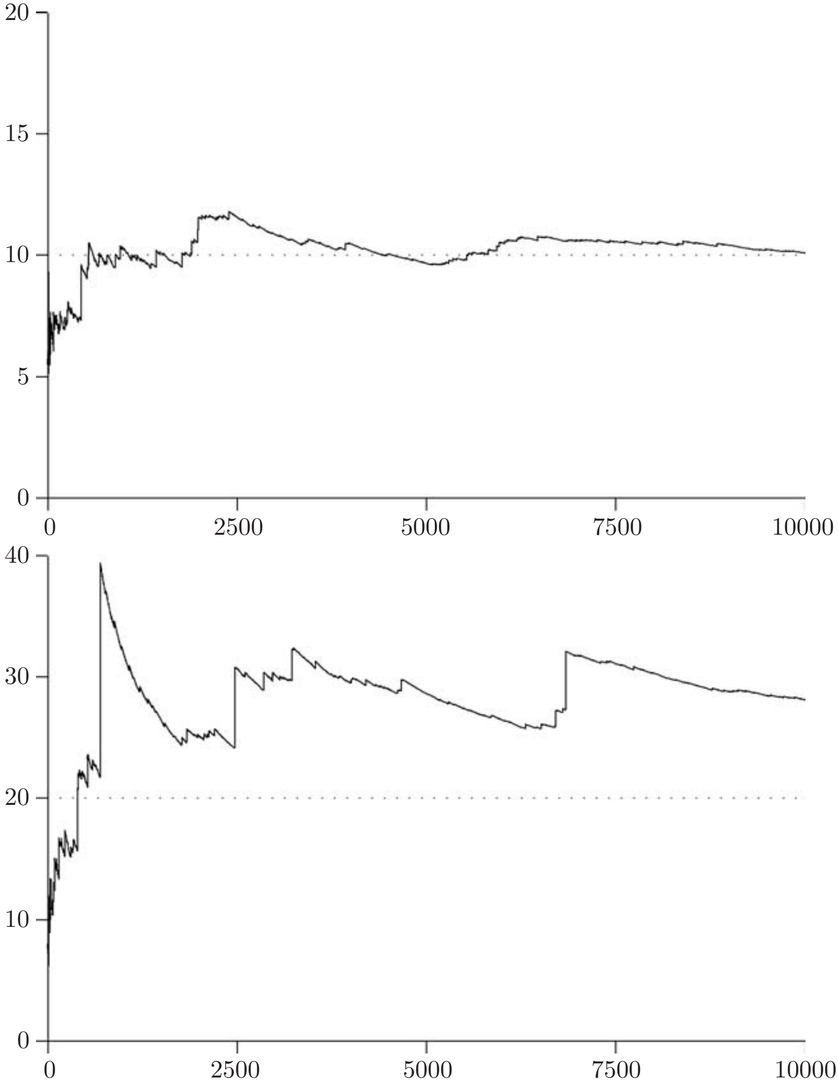


Fig. 2.3. Average payoff in St. Petersburg game.

Roulette is one of the oldest casino games. The basis for this game was established at the beginning of the seventeenth century by French mathematician Blaise Pascal. The most common version of roulette uses the numbers 0, 1, . . . , 36, where the number 0 is always reserved as winning number for the house (European roulette). Players bet against the house on a number to emerge when the roulette wheel stops spinning. Bets may be placed on either single numbers or combinations of numbers. A winning bet placed on a combination of k numbers earns $\frac{36}{k} - 1$ times the amount staked plus the initial stake itself in winnings. For each separate bet, the expected value of the casino take for each dollar staked is equal to

$$1 \times \left(1 - \frac{k}{37}\right) - \left(\frac{36}{k} - 1\right) \times \frac{k}{37} = \frac{1}{37} \text{ dollars.}$$

In other words, casinos get 2.7 cents for every dollar staked over the long run, and this translates into a *house percentage* of 2.70%. In American roulette, which differs from the European version in that the roulette wheel has a "house double-zero" (00) in addition to the single (house) zero (0), the house percentage for each bet is 5.26%, except for five-number combination bets; these offer an even higher house percentage of 7.89%. It is impossible to win this game over the long run. No matter what betting system you use, you can count on giving away 2.7 cents for every dollar you stake. It is impossible to make a winning combination of bets when every individual bet is a losing proposition. Betting systems are only of interest for their entertainment and excitement value. Betting systems that are much in use are the Big–Martingale system and the D'Alembert system. In both systems, the game is played according to simple probability patterns with 18 numbers (always betting on red, for example), where the payoff equals twice the amount staked. The Big–Martingale system works thus: the amount of your initial stake is one chip. If you lose, your next stake will be twice your previous stake plus one chip. If you win, your next stake will be one chip. Should you score your first win after four attempts, your first four stake amounts will have been 1, 3, 7, and 15 chips, and after four turns you will have gained $30 - (1 + 3 + 7 + 15) = 4$ chips. In the D'Alembert system the amount of your initial stake will also be one chip. After a loss you raise your stake with one chip, and after a win you decrease your stake by one chip. Engaging as these systems may be, they, too, will result in a loss over the long run of 2.7 cents for every dollar staked. Attempting to influence your average loss in roulette by using a betting system is as nonsensical as it was, long ago, for a despot to try to influence the ratio of newborn boys to girls by prohibiting women from bearing any more children as soon as they gave birth to a boy. The latter merely the folly of the gambler dressed up in different clothes!

Betting systems for roulette that claim to be winners, whether in book form or on the Internet, represent nothing more than charlatanism. To underline the fact that one betting system is no better than another in roulette, we chart the results of a simulation study that compares the Big–Martingale system with the flat system, which calls for a stake of one chip on each round. The study was composed of one million simulated repetitions of the game for both systems. For each repetition the initial playing capital consisted of 100 chips, and a maximum of 100 bets were made, always on red. Under the flat system, one chip was staked on each spin of the wheel. Under the Big–Martingale system, 100 bets or less were made, depending on how long the chips lasted. The following total scores were found for the one million simulation runs:

$$\begin{aligned} &\text{Flat system:}\\ \text{total amount staked} &= 100{,}000{,}000\\ \text{total loss} &= 2{,}706{,}348\\ \text{loss per unit staked} &= 0.0271 \end{aligned}$$

$$\begin{aligned} &\text{Big–Martingale system:}\\ \text{total amount staked} &= 384{,}718{,}672\\ \text{total loss} &= 10{,}333{,}828\\ \text{loss per unit staked} &= 0.0269. \end{aligned}$$

As you can see, the quotient of the total loss and the total amount staked, in both cases, lies near the house advantage of 0.027. It is also interesting to note the probability distribution of the number of chips that are won or lost at the end of one repetition of the game. We give the simulated probability distribution for the flat system in Figure 2.4 and for the Big–Martingale system in Figure 2.5. In Figure 2.5, a logarithmic scale is used. As might have been expected, the probability distribution for the Big–Martingale system is much more strongly concentrated at the outer ends than the distribution for the flat system.

2.7 The Kelly betting system[†]

You are playing a game where you have an edge. How should you bet to manage your money in a good way? The idea is always to bet a *fixed proportion* of your present bankroll. When your bankroll decreases you bet less, as it increases you bet more. This strategy is called the Kelly system, after the American mathematician J.F. Kelly, Jr., who published this system in 1956.[‡] The objective of Kelly betting is to maximize the long-run rate of growth of your bankroll.

[†] This paragraph can be skipped at first reading.

[‡] However, many years before Kelly's publication, W.A. Whitworth already proposed this system in his book *Chance and Choice*, 3rd edition, Deighton Bell, Cambridge, 1886. In fact, the basic idea of Kelly betting goes back to Daniel Bernoulli. In his famous 1738 article he suggested that when you have a choice of bets or investments you should use that with the highest geometric mean of outcomes. The geometric mean of positive numbers $a_1, a_2, \ldots, a_n$ is defined as $(a_1 \times a_2 \times \cdots \times a_n)^{1/n}$, which is equivalent to $\exp(\frac{1}{n}\sum_{i=1}^{n} \ln(a_i))$.

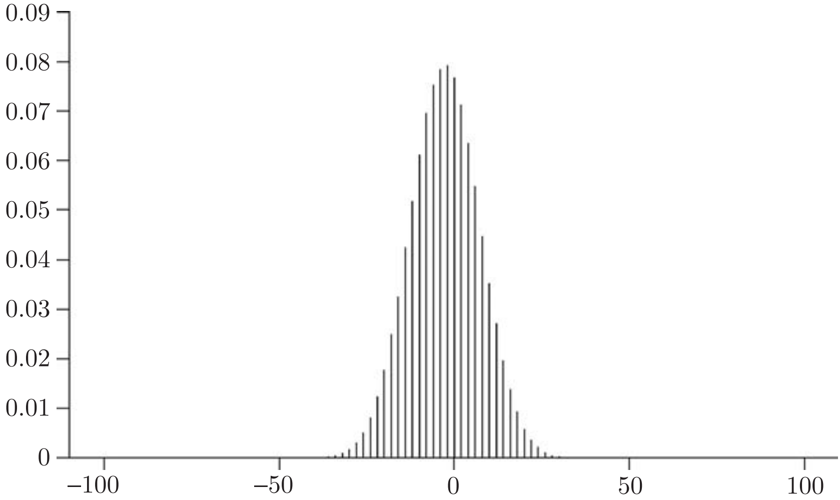


Fig. 2.4. Win/loss calculations for the flat system.

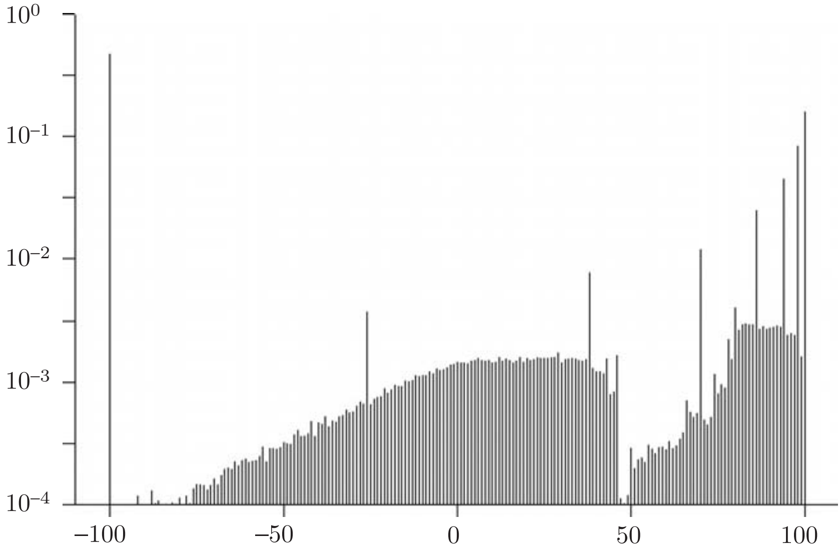


Fig. 2.5. Win/loss calculations for the Big–Martingale system.

The optimal value of the fraction to bet can be found by simple arguments based on the law of large numbers.

Suppose you are offered a sequence of bets, each bet being a losing proposition with probability 0.6 and paying out three times your stake with probability 0.4. How to gamble if you must? Note that each bet is favorable, because the expected net payoff is positive ($0.4 \times 3 - 1 > 0$). However, it is not wise to bet your whole bankroll each time; if you do, you will certainly go bankrupt after a while. Indeed, it is better to bet 10% of your current bankroll each time. This strategy maximizes the long-run rate of growth of your bankroll and achieves an effective rate of return of 0.98% over the long run. To derive this result, it is helpful to use a general notation. Let's assume that the *payoff odds* are $f - 1$ to 1 for a given $f > 1$. That is, in case of a win, you get a payoff of f times the amount bet; otherwise, you lose the amount bet. Letting p denote the probability of the player winning the bet, it is assumed that $0 < p < 1$ and $pf > 1$ (favorable bet).

Assuming that your starting bankroll is V_0, define the random variable V_n as

$$V_n = \text{the size of your bankroll after } n \text{ bets},$$

when you bet a fixed fraction α ($0 < \alpha < 1$) of your current bankroll each time. Here it is supposed that winnings are reinvested and that your bankroll is infinitely divisible. It is not difficult to show that

$$V_n = (1 - \alpha + \alpha R_1) \times \cdots \times (1 - \alpha + \alpha R_n)\, V_0 \qquad \text{for } n = 1, 2, \ldots,$$

where the random variable R_k is equal to the payoff factor f if the kth bet is won and is otherwise equal to 0. Evidence of this relationship appears at the end of this section. In mathematics, a growth process is most often described with the help of an exponential function. This motivates us to define the exponential growth factor G_n via the relationship

$$V_n = V_0 e^{nG_n},$$

where $e = 2.718\ldots$ is the base of the natural logarithm. If you take the logarithm of both sides of this equation, you see that the definition of G_n is equivalent to

$$G_n = \frac{1}{n} \ln\left(\frac{V_n}{V_0}\right).$$

If you apply the above product formula for V_n and use the fact that $\ln(ab) = \ln(a) + \ln(b)$, then you find

$$G_n = \frac{1}{n}\left[\ln(1 - \alpha + \alpha R_1) + \cdots + \ln(1 - \alpha + \alpha R_n)\right].$$

The law of large numbers applies to the growth rate G_n if n (= the number of bets) is very large. Indeed, the random variables $X_i = \ln(1 - \alpha + \alpha R_i)$ form a sequence of independent random variables having a common distribution. If you apply the law of large numbers, you find that

$$\lim_{n \to \infty} G_n = E\left[\ln(1 - \alpha + \alpha R)\right],$$

where the random variable R is equal to f with probability p and is equal to 0 with probability $1 - p$. This leads to

$$\lim_{n \to \infty} G_n = p \ln(1 - \alpha + \alpha f) + (1 - p) \ln(1 - \alpha).$$

Under a strategy that has a fixed betting fraction α, the long-run growth factor of your bankroll is thus given by

$$g(\alpha) = p \ln(1 - \alpha + \alpha f) + (1 - p) \ln(1 - \alpha).$$

It is not difficult to verify that an α_0 with $0 < \alpha_0 < 1$ exists such that the long-run growth factor $g(\alpha)$ is positive for all α with $0 < \alpha < \alpha_0$ and negative for all α with $\alpha_0 < \alpha < 1$. Choose a betting fraction between 0 and α_0 and your bankroll will ultimately exceed every large level if you simply keep playing for a long enough period of time. It is quite easy to find the value of α for which the long-run growth factor of your bankroll is maximal. Toward that end, set the derivative of the function $g(\alpha)$ equal to 0. This leads to $p(f - 1)/(1 - \alpha + f\alpha) - (1 - p)/(1 - \alpha) = 0$. Hence, the optimal value of α is given by

$$\alpha^* = \frac{pf - 1}{f - 1}.$$

This is the famous formula for the *Kelly betting fraction*. This fraction can be interpreted as the ratio of the expected net payoff for a one-dollar bet and the payoff odds. The Kelly system is of little use for casino games, but may be useful for the situation of investment opportunities with positive expected net payoff. In such situations, it may be more appropriate to use a modification of the Kelly formula that takes into account the interest accrued on the noninvested part of your bankroll. In Problem 2.9, the reader is asked to modify the Kelly formula when a fixed interest is attached to a player's nonstaked capital.

2.7.1 Long-run rate of return

For the strategy under which you bet the same fraction α of your bankroll each time, define the return factor γ_n by

$$V_n = (1 + \gamma_n)^n V_0.$$

The random variable γ_n gives the rate of return on your bankroll over the first n bets. It follows from the relationship $V_n = e^{nG(n)}V_0$ that

$$\gamma_n = e^{G(n)} - 1.$$

Earlier, we saw that the random variable $G(n)$ converges to the constant $g(\alpha) = p\ln(1 - \alpha + \alpha f) + (1 - p)\ln(1 - \alpha)$ as $n \to \infty$. This means that γ_n converges to the constant $\gamma_{\text{eff}}(\alpha) = e^{g(\alpha)} - 1$ as $n \to \infty$. The constant $\gamma_{\text{eff}}(\alpha)$ gives the effective rate of return for the long run if you bet the same fraction α of your bankroll each time. Substituting the expression for $g(\alpha)$ and using $e^{b\ln(a)} = a^b$, you find that the long-run rate of return is given by

$$\gamma_{\text{eff}}(\alpha) = (1 - \alpha + \alpha f)^p(1 - \alpha)^{1-p} - 1.$$

As an illustration, consider the data

$$\begin{aligned} &p(= \text{win probability}) = 0.4, \quad && f(= \text{payoff factor}) = 3, \\ &n(= \text{number of bets}) = 100, \quad && V_0(= \text{starting capital}) = 1. \end{aligned}$$

Under the Kelly system, a fraction $\alpha^* = 0.1$ of your current bankroll is bet each time. Let us compare this strategy with the alternative strategy under which the fixed fraction $\alpha = 0.25$ of your bankroll is bet each time. The comparison is done by executing a simulation experiment where both strategies are exposed to the same experimental conditions. The results of this simulation are given in Figure 2.6. The simulation outcomes confirm that, in the long run, the Kelly strategy is superior with respect to the growth rate. From the formula for $\gamma_{\text{eff}}(\alpha)$, it follows that the Kelly betting strategy with $\alpha = 0.1$ has an effective rate of return of 0.98% over the long run, whereas the betting strategy with $\alpha = 0.25$ has an effective rate of return of -1.04% over the long run. In Chapter 5, we come back to another property of the Kelly strategy: it minimizes the expected time needed to reach a specified, but large value for your bankroll.

2.7.2 Fractional Kelly

As you can see from Figure 2.6, the Kelly growth rate curve gives you a roller coast ride. Most of us would not be able to sleep at night while our investment is on such a ride. If you wish to reduce your risk, you are better off using a fractional Kelly strategy. Under such a strategy you always bet the same fraction $c\alpha^*$ of your bankroll for a constant c with $0 < c < 1$. The increase in safety is at the expense of only a small decrease in the growth rate of your bankroll. The reduction in the long-term rate of return can be quantified by the approximate

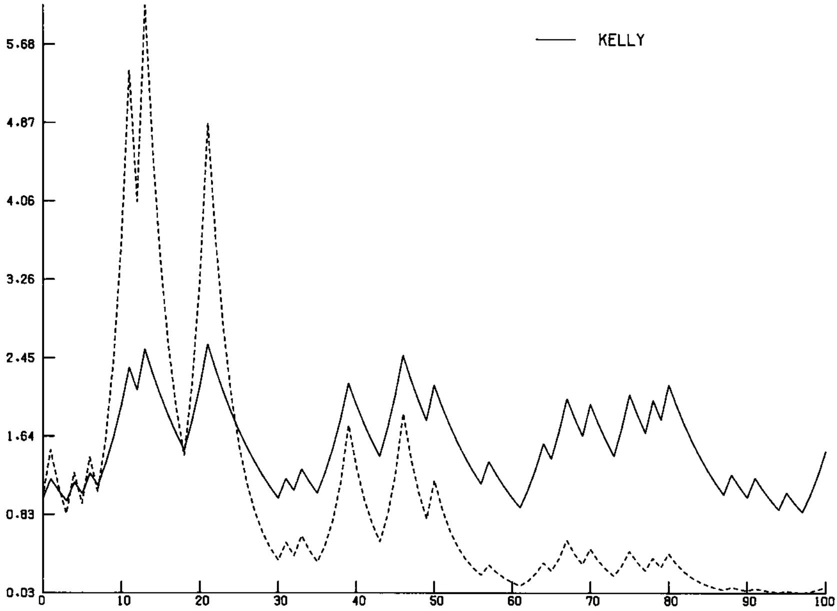


Fig. 2.6. Kelly strategy and an alternative.

relation

$$\frac{\gamma_{\mathrm{eff}}(c\alpha^*)}{\gamma_{\mathrm{eff}}(\alpha^*)} \approx c(2-c).$$

Thus, "half Kelly" has approximately $\frac{3}{4}$ of the long-run rate of return of the Kelly strategy. The increased safety of the fractional Kelly strategy ($\alpha = c\alpha^*$) can be quantified by the approximate relation[†]

$$\begin{aligned} &P(\text{reaching a bankroll of } aV_0 \text{ without falling down first to } bV_0)\\ &\quad\approx \frac{1-b^{2/c-1}}{1-(b/a)^{2/c-1}} \end{aligned}$$

for any $0 < b < 1 < a$. For example, by betting only half of the Kelly fraction, you give up one-quarter of your maximum growth rate, but you increase the probability of doubling your bankroll without having it halved first from 0.67 to 0.89. This probability is about 98% for the fractional Kelly strategy with $c = 0.3$. The value $c = 0.3$ is a recommended value for fractional Kelly.

[†] The approximate relations are taken from Edward O. Thorp (1998), The Kelly criterion in blackjack, sports betting, and the stock market: www.bjmath.com. Simulation studies reveal that the approximations are very accurate for all cases of practical interest.

2.7.3 Derivation of the growth rate

Proof of the relationship

$$V_n = (1-\alpha+\alpha R_1)\times\cdots\times(1-\alpha+\alpha R_n)V_0 \qquad \text{for } n = 1, 2, \ldots$$

is as follows. If you invest a fraction α of the capital you possess every time, then

$$V_k = (1-\alpha)V_{k-1} + \alpha V_{k-1}R_k \qquad \text{for } k = 1, 2, \ldots .$$

At this point, we apply the mathematical principle of induction to prove the product formula for V_n. This formula is correct for $n = 1$ as follows directly from $V_1 = (1-\alpha)V_0 + \alpha V_0 R_1$. Suppose the formula is proven for $n = j$. It would then be true for $n = j + 1$ that

$$\begin{aligned} V_{j+1} &= (1-\alpha)V_j + \alpha V_j R_{j+1} = (1-\alpha+\alpha R_{j+1})V_j \\ &= (1-\alpha+\alpha R_{j+1})(1-\alpha+\alpha R_1)\times\cdots\times(1-\alpha+\alpha R_j)V_0 \\ &= (1-\alpha+\alpha R_1)\times\cdots\times(1-\alpha+\alpha R_{j+1})V_0, \end{aligned}$$

which proves the assertion for $n = j + 1$ and so the induction step is complete.

2.8 Random-number generator

Suppose you are asked to write a long sequence of H's and T's that would be representative of the tossing of a fair coin, where H stands for heads and T for tails. You may not realize just how incredibly difficult a task this is. Virtually no one is capable of writing down a sequence of H's and T's such that they would be statistically indistinguishable from a randomly formed sequence of H's and T's. Anyone endeavoring to accomplish this feat will likely avoid clusters of H's and T's. But such clusters do appear with regularity in truly random sequences. For example, as we saw in Section 2.1, the probability of tossing heads five successive times in 20 tosses of a fair coin is not only nonnegligible, but also it actually amounts to 25%. A sequence of H's and T's that does not occasionally exhibit a long run of H's or a long run of T's cannot have been randomly generated. In probability theory, access to random numbers is of critical importance. In the simulation of probability models, a random-number generator, as it is called, is simply indispensable.

A *random-number generator* produces a sequence of numbers that are picked at random between 0 and 1 (excluding the values 0 and 1). It is as if fate falls on a number between 0 and 1 by pure coincidence. When we speak of generating a random number between 0 and 1, we assume that the probability of the generated number falling in any given subinterval of the unit interval (0,1) equals the length of that subinterval. Any two subintervals of the same length have equal probability of containing the generated number. In other words, the probability distribution of a random number between 0 and 1 is the so-called *uniform* distribution on (0,1). This is a continuous distribution, which means that it only makes sense to speak of the probability of a randomly chosen number falling in a *given interval*. It makes no sense to speak of the probability of an individual value. The probability of each individual outcome is zero. The amount of probability assigned to an interval gets smaller and smaller as the interval shrinks and becomes zero when the interval has shrunk to zero. For example, a randomly chosen number between 0 and 1 will fall in the interval (0.7315, 0.7325) with a probability of 0.001. The probability that a randomly chosen number will take on a *prespecified* value, say 0.732, is equal to 0. A random-number generator immediately gives us the power to simulate the outcome of a fair-coin toss without actually having to toss the coin. The outcome is heads if the random number lies between 0 and $\frac{1}{2}$ (the probability of this is 0.5), and otherwise the outcome is tails.

Producing random numbers is not as easily accomplished as it seems, especially when they must be generated quickly, efficiently, and in massive amounts.[†] For occasional purposes, the use of a watch might be suitable if it were equipped with a stopwatch that could precisely measure time in tenths of seconds. Around 1920, crime syndicates in New York City's Harlem used the last five digits of the daily published US treasury balance to generate the winning number for their illegal "Treasury Lottery." But this sort of method is of course not really practical. Even for simple simulation experiments the required amount of random numbers runs quickly into the tens of thousands or higher. Generating a very large amount of random numbers on a one-time only basis, and storing them up in a computer memory, is also practically infeasible. But there is a solution to this kind of practical hurdle that is as handsome as it is practical. Instead of generating *truly* random numbers, a computer can generate *pseudo-random* numbers, as they are known, and it achieves this with the help of a nonrandom procedure. This procedure is iterative by nature and is determined by a suitably chosen function f. Starting with an arbitrary number z_0, the numbers $z_1, z_2, \ldots$ are successively generated by

$$z_1 = f(z_0), z_2 = f(z_1), \ldots, z_n = f(z_{n-1}), \ldots .$$

[†] An interesting account of the history of producing random numbers can be found in D.J. Bennett's *Randomness*. Cambridge, MA: Harvard University Press, 1999.

We refer to the function f as a random-number generator and it must be chosen such that the series $\{z_i\}$ is indistinguishable from a series of truly random numbers. In other words, the output of function f must be able to stand up to a great many statistical tests for "randomness." When this is the case, we are in command of a simple and efficient procedure to produce random numbers. An added advantage is that a series of numbers generated by a random-number generator is *reproducible* by beginning the procedure over again with the same seed number z_0. This can come in very handy when you want to make a simulation that compares alternative system designs: the comparison of alternative systems is purest when it can be achieved (to the extent that it is possible) under identical experimental conditions.

In practice, most random-number generators in use can be referred to as a *multiplicative* generator

$$z_n = az_{n-1} \text{ (modulo } m),$$

where a and m are fixed positive integers. For the seed number z_0, a positive integer must always be chosen. The notation $z_n = az_{n-1}$ (modulo m) means that z_n represents the whole remainder of az_{n-1} after division by m; for example, 17 (modulo 5) = 2. This scheme produces one of the numbers $0, 1, \ldots, m-1$ each time. It takes no more than m steps until some number repeats itself. Whenever z_n takes on a value it has had previously, exactly the same sequence of values is generated again, and this cycle repeats itself endlessly. When the parameters a and m are suitably chosen, the number 0 is not generated and each of the numbers $1, \ldots, m-1$ appears exactly once in each cycle. In this case the parameter m gives the length of the cycle. This explains why a very large integer should be chosen for m. The number z_n determines the random number u_n between 0 and 1 by $u_n = \frac{z_n}{m}$. The quality of the generator is strongly dependent on the choice of parameters a and m (a much used generator is characterized as $a = 630{,}360{,}016$ and $m = 2^{31} - 1$). We will not delve into the theory behind this. An understanding of the theory is not necessary in order to use this random-number generator on your computer. Today, most computers come equipped with a good random-number generator (this was not the case in days of yore). The simulation programs listed at the end of this chapter show very clearly how to use the random-number generator.

2.8.1 Pitfalls encountered in randomizing

The development of a good random-number generator must not be taken lightly. It is foolish, when using a multiplicative generator, to choose parameters for *a* and *m* oneself, or to piece together a patchwork algorithm by combining fragments from a number of existing methods, for example. That something is wild or complicated does not automatically mean that it is also random. The task of mixing objects together (lotto balls, for example) through physical means, such that we can say that the result is a random mix, is even more difficult than making a good random-number generator. A useful illustration of the difficulties involved in this undertaking can be seen in the example of the drafting of soldiers into the U.S. Armed Forces during the period of the Vietnam War. In 1970, widely varying drafting programs that had been run by individual states were scrapped in favor of a national lottery. The framework of the lottery was built on a plan to use birthdays as a means of choosing the young men to be drafted. Preparations for the drawing were made as follows. First, the 31 days of January were recorded on pieces of paper that were placed into capsules, and these, in turn, were placed into a large receptacle. After that, the 29 days of February (including February 29) were recorded, placed into capsules and added to the receptacle. At this point, the January and February capsules were mixed. Next, the 31 days of March were recorded, encapsulated and mixed through the January/February mixture, and the days of all the other months were treated similarly. When it was time for the drawing, the first capsule to be drawn was assigned the number 1, the second capsule drawn was assigned a number 2, and so on until all capsules had been drawn and assigned a number between 1 and 366. The men whose birth dates were contained in capsules receiving low-end numbers were called up first. Doubts about the integrity of the lottery were raised immediately following the drawing. Statistical tests demonstrated, indeed, that the lottery was far from random (see also Section 3.5). The failure of the lottery is easily traced to the preparatory procedures that occurred prior to the drawing. The mixing of the capsules was inadequately performed: the January capsules were mixed through the others 11 times, wheras the December capsules were mixed only once. In addition, it appeared that, during the public drawing, most capsules were chosen from the top of the receptacle. Preparations for the 1971 drawing were made with a great deal more care, partly because statisticians were called in to help. This time, two receptacles were used: one with 366 capsules for the days of the year, and another with 366 capsules for the numbers 1 through 366. One capsule was chosen from each receptacle in order to couple the days and numbers. The biggest improvement was that the *order* in which the capsules with the 366 days and the 366 numbers went into their respective receptacles was determined *beforehand* by letting a computer generate two random permutations of the integers $1, \ldots, 366$. The random permutations governed the order in which the capsules containing the days of the year and the lottery numbers were put into the receptacles. Next, a physical hand-mixing of the capsules took place. In fact, the physical mixing was not necessary but served as a public display of what people think of as random. The actual mixing took place through the random permutations. In Section 2.9, it is shown how the computer generates a random permutation of the integers $1, \ldots, 366$.

2.8.2 The card shuffle

Another example of how informal procedures will not lead to a random mix can be seen in the shuffling of a deck of cards. Most people will shuffle a pack of 52 cards three or four times. This is completely inadequate to achieve a random mix of the cards. A card mix is called random when it can be said that each card in the deck is as likely to turn up in any one given position as in any other. For a pack of 52 cards, it is reasonable to say that *seven* "riffle shuffles" are needed to get a mix of cards that, for all practical purposes, is sufficiently random. In a riffle shuffle the deck of cards is divided into two more or less equal stacks that are then intermingled (riffled) to form one integrated stack. It is assumed that the riffle shuffle is imperfect and thus contains a nonnegligible element of chance (in the perfect riffle shuffle, the deck is exactly halved and every single card is interwoven back and forth; it can be mathematically demonstrated that, after eight such perfect riffle shuffles, a new deck of 52 cards will have returned to its original order). Some experienced bridge players are capable of taking advantage of situations in which the cards are shuffled such that their resulting distribution is not random. In order to raise their chances of winning, some regular casino gamblers make use of the knowledge that cards are usually not shuffled to a random mix (one deck should be shuffled seven times, two decks should be shuffled nine times, and six packs should be shuffled twelve times). In professional bridge tournaments and in casinos, computers are being used more and more to ensure a random mix of cards. It took advanced mathematics to explain the fact that only after seven or more imperfect riffle shuffles could one expect to find a more or less random mix of a deck of 52 cards.[†] The mix of cards resulting from seven riffle shuffles is sufficiently random for common card games such as bridge, but it is not really random in the mathematical sense. This can be seen in Peter Doyle's fascinating card game called "Yin Yang Solitaire." To play this game, begin with a new deck of cards. In the United States, a new deck of cards comes in the order specified here: with the deck laying face-down,

[†] See D.J. Aldous and P. Diaconis, "Shuffling cards and stopping times," *The American Mathematical Monthly* **93** (1986): 333–348.

you will have ace, two, . . . , king of hearts, ace, two, . . . , king of clubs, king, . . . , two, ace of diamonds, and king, . . . , two, ace of spades. Hearts and clubs are yin suits, and diamonds and spades are yang suits. The deck is shuffled seven times, cut, and placed face down on the table. The player's goal is to make a stack of cards for each suit. This is achieved by removing each card from the top of the deck and turning it over to reveal its face value. A stack for a suit is started as soon as the ace of that suit appears. Cards of the same suit are added to the stack according to the rule that they must be added in the order ace, two, . . . , king. A single pass through the deck is normally not enough to complete the stack for all four suits. Having made a pass, the remaining deck is turned back over and another pass is made. The game is over as soon as either the two yin-suit stacks or the two yang-suit stacks are complete. Yin wins if the two yin suits are completed first. If the deck has been thoroughly permuted (by being put through a clothes dryer cycle, say), the yins and yangs will be equally likely to be completed first. But it turns out that, after seven ordinary riffle shuffles and a cut, it is significantly more likely that the yins will be completed before the yangs. The probability of yin winning is about 81% in this case. This demonstrates the difficulty of getting a fully random mix of the cards by hand. Finally, it is interesting to note that the probability of yin winning is approximately equal to 67%, 59%, and 54%, respectively, after eight, nine, and ten riffle shuffles. Only after 15 riffle shuffles can we speak of a nearly 50% probability of yin winning.

2.9 Simulating from probability distributions

A random-number generator for random numbers between 0 and 1 suffices for the simulation of random samples from an arbitrary probability distribution. A handsome theory with all kinds of efficient methods has been developed for this purpose; however, we will confine ourselves to mentioning just the few basic methods that serve our immediate purposes.

2.9.1 Simulating from an interval

You want to surprise some friends by arriving at their party at a completely random moment in time between 2:30 and 5:00. How can you determine that moment? You must generate a random number between $2\frac{1}{2}$ and 5. How do you blindly choose a number between two given real numbers a and b when $a < b$? First, you have your computer generate a random number u between 0 and 1.

Then, you find a random number between a and b by

$$a + (b - a)u.$$

2.9.2 Simulating from integers

How can you designate one fair prize-winner among the 725 people who correctly answered a contest question? You achieve this by numbering the correct entries as $1, 2, \ldots, 725$ and generating randomly an integer out of the integers $1, 2, \ldots, 725$. How can you blindly choose an integer out of the integers $1, \ldots, M$? First, have your computer generate a random number u between 0 and 1. Then, using the notation $\lfloor f \rfloor$ for the integer that results by rounding down the number f, the integer

$$1 + \lfloor Mu \rfloor$$

can be considered as a random integer sampled from the integers $1, \ldots, M$. One application is the simulation of the outcome of a roll of a fair die ($M = 6$). For example, the random number $u = 0.428\ldots$ leads to the outcome 3 $(= 1 + \lfloor 6u \rfloor)$ of the roll of the die. In general, letting u denote a random number between 0 and 1, a random integer from the integers $a, a + 1, \ldots, b$ is given by

$$a + \lfloor (b - a + 1)u \rfloor.$$

A nice illustration of the procedure of drawing a random integer is provided by simulating the famous lost boarding pass puzzle. One hundred people line up to board an airplane with 100 passenger seats. Each passenger gets on one at a time to select his or her assigned seat. The first passenger in line has lost his boarding pass and takes a random seat instead. Each subsequent passenger takes his or her assigned seat if available, otherwise a random unoccupied seat. You are the last passenger. What is the probability that you can get your own seat? In simulating this problem, it is convenient to number the passengers in line as $1, 2, \ldots, 100$ and to number their assigned seats accordingly. A simulation run is started by drawing a random integer from the integers $1, 2 \ldots, 100$, say the integer s. If $s = 1$ or $s = 100$ the simulation run can be stopped: the last passenger in line takes his or her own seat if $s = 1$ and does not take the originally assigned seat if $s = 100$. In case $1 < s < 100$, then passengers $2, \ldots, s - 1$ take their own seats and passenger s takes a random seat from the seats $1, s + 1 \ldots, 100$. In fact we then have the lost boarding pass problem with $100 - s + 1$ seats rather than 100 seats. Renumber the seats $1, s + 1 \ldots, 100$ as $1, 2 \ldots, 100 - s + 1$ and draw a random integer from the integers $1, 2 \ldots, 100 - s + 1$, say the integer t. The last passenger gets his or her own seat if $t = 1$ and does not get the assigned seat if $t = 100 - s + 1$. It will be obvious how to proceed the simulation run if $1 < t < 100 - s + 1$. By making a large number of simulation runs, you can estimate the desired probability by dividing the number of successful runs by the total number of runs. The answer is surprising and many people will bang their forehead when they see the answer.

2.9.3 Simulating from a discrete distribution

For a football pool, how can you come up with a replacement outcome for a cancelled football match for which a group of experts has declared a home-win with 50% probability, a visitor's win with 15% probability, and a tie game with 35% probability? You can do this by simulating from a distribution that has assigned probabilities 0.50, 0.15 and 0.35 to the numbers 1, 2, and 3, respectively. How do you simulate from a discrete distribution of a random variable X that assumes a finite number of values $x_1, \ldots, x_M$ with corresponding probabilities $p_1, \ldots, p_M$? This is very simple for the special case of a two-point distribution in which the random variable X can only assume the values x_1 and x_2. First, you have your computer generate a random number u between 0 and 1. Next, for the random variable X you find the simulated value x_1 if $u \leq p_1$ and the value x_2 otherwise. You generate in this way the value x_1 with probability p_1 and the value x_2 with probability $p_2 = 1 - p_1$ (why?). In particular, the outcome of "heads or tails" in the toss of a fair coin can be simulated in this way. Generate a random number u between 0 and 1. If $u \leq \frac{1}{2}$, then the outcome is "heads" and otherwise the outcome is "tails." The inversion method for simulating from a two-point distribution can also be extended to that of a general discrete distribution, but this leads to an inefficient approach for the general case of $M > 2$. A direct search for the index l satisfying $p_1 + \cdots + p_{l-1} < u \leq p_1 + \cdots + p_l$ is too time-consuming for simulation purposes when M is not small. A direct search for the index l satisfying $p_1 + \cdots + p_{l-1} < u \leq p_1 + \cdots + p_l$ is too time-consuming for simulation purposes when M is not small. An ingenious method has been designed to circumvent this difficulty. We briefly discuss this method. The reader may skip this discussion without loss of continuity. The key idea is to split the total probability mass 1 of the points $x_1, x_2, \ldots, x_M$ in B equal portions of $\frac{1}{B}$, where B is a sufficiently large integer with $B > M$ (e.g., $B = 2M$). In each of the B buckets $b = 1, 2, \ldots, B$, you put a probability mass of $\frac{1}{B}$. Also, you assign to each bucket one or more of the mass points x_j for a total mass of $\frac{1}{B}$. How the mass points are precisely assigned to each bucket will be explained in a moment. As a consequence of the fact that B is sufficiently large, only a few of the points x_j will be assigned to any bucket. Once this preparatory work is done, you can simulate from the probability mass function. You first choose at random one of the B buckets. Next, you determine within this bucket the mass point x_l for which $p_1 + \cdots + p_{l-1} < u \leq p_1 + \cdots + p_l$. This requires very little computing time, since the bucket contains only a few of the x_j. How to assign mass points x_j to each bucket is best explained by an example. Suppose that the random variable X can take on $M = 4$ values and that its probability mass function $p_j = P(X = x_j)$ for $j = 1, \ldots, 4$ is given by

$$p_1 = 0.30, \;\; p_2 = 0.20, \;\; p_3 = 0.35, \;\; p_4 = 0.15.$$

Let us take $B = 5$ buckets. Each bucket represents a probability mass of 0.2. In bucket 1 this probability mass is obtained by assigning the mass point x_1 to this bucket for 0.2 of its probability mass 0.3. The point x_1 is assigned to bucket 2 as well, but for the remaining 0.1 of its probability mass. Also, point x_2 is assigned to bucket 2 for a probability mass of 0.1 to get a total probability mass of 0.2 in bucket 2. Continuing in this way, the points x_2 (for a mass of 0.1) and x_3 (for a mass of 0.1) are assigned to bucket 3, the point x_3 (for a mass of 0.2) is assigned to bucket 4, and the points x_3 (for a mass of 0.05) and x_4 (for a mass of 0.15) are assigned to bucket 5. Then, the simulation from the discrete random variable X proceeds as follows:

Step 1 Generate a random number u between 0 and 1.
Step 2 Choose at random a bucket b according to $b := 1 + \lfloor Bu \rfloor$.
Step 3 Search in bucket b for the point x_l with $\sum_{j=1}^{l-1} p_j < u \leq \sum_{j=1}^{l} p_j$.

The last step can be very efficiently implemented (by $\frac{b-1}{B} < u \leq \frac{b}{B}$, the point x_l can be found in bucket b). The point x_l obtained in Step 3 is a random sample from the discrete random variable X. As illustration, suppose that the random number $u = 0.8201\ldots$ is generated in Step 1. Then, bucket $b = 5$ is chosen in Step 2 and the point x_3 results from Step 3.

In case each of the probabilities $p_j = P(X = x_j)$ is given only in a few decimals, then there is a very simple but useful method called the *array method*. As a means of understanding this method, consider the case in which each probability p_j is given in precisely two decimals. That is, p_j can be represented by $k_j/100$ for some integer k_j with $0 \leq k_j \leq 100$ for $j = 1, \ldots, M$. You then form an array `A[i]`, $i = 1, \ldots, 100$, by setting the first k_1 elements equal to x_1, the next k_2 elements equal to x_2, etc., and the last k_M elements equal to x_M. To illustrate, take again the probability mass function $p_j = P(X = x_j)$ for $j = 1, \ldots, 4$ with $p_1 = 0.30$, $p_2 = 0.20$, $p_3 = 0.35$ and $p_4 = 0.15$. You then have

$$\begin{aligned} &\texttt{A[1]} = \cdots = \texttt{A[30]} = x_1, && \texttt{A[31]} = \cdots = \texttt{A[50]} = x_2, \\ &\texttt{A[51]} = \cdots = \texttt{A[85]} = x_3, && \texttt{A[86]} = \cdots = \texttt{A[100]} = x_4. \end{aligned}$$

Now have your computer generate a random number u between 0 and 1. Calculate the integer $m = 1 + \lfloor 100u \rfloor$. This simulated integer m is a randomly chosen integer from the integers $1, \ldots, 100$. Next, take $\texttt{A}[m]$ as the simulated value of the random variable X. For example, suppose that the random number $u = 0.8201\ldots$ has been generated. This gives $m = 83$ and thus the simulated value x_3 for the random variable X. It will be clear that the array method applies with an array of one thousand elements when each probability p_j is given to exactly three decimal places.

2.9.4 Random permutation

How can you randomly assign numbers from the integers $1, \ldots, 10$ to ten people such that each person gets a different number? This can be done by making a random permutation of the integers $1, \ldots, 10$. A random permutation of the integers $1, \ldots, 10$ is a sequence in which the integers $1, \ldots, 10$ are put in random order. The following algorithm generates a random permutation of $1, \ldots, n$ for a given positive integer n.

Algorithm for random permutation

```
(i)   Initialize $t := n$ and $a[j] := j$ for $j = 1, \ldots, n$.
(ii)  Generate a random number $u$ between 0 and 1.
(iii) Set $k := 1 + \lfloor tu \rfloor$ (random integer from the indices $1, \ldots, t$).
      Interchange the current values of $a[k]$ and $a[t]$.
(iv)  Let $t := t - 1$. If $t > 1$, return to step (ii); otherwise, stop and the desired
      random permutation $(a[1], \ldots, a[n])$ is obtained.
```

The idea of the algorithm is first to randomly choose one of the integers $1, \ldots, n$ and to place that integer in position n. Next, you randomly choose one of the remaining $n - 1$ integers and place it in position $n - 1$, and so on. For the simulation of many probability problems, this is a very useful algorithm. A nice illustration of the procedure of generating a random permutation is provided by the simulation of the best-choice problem from Section 2.3.1. For the case of 20 slips of paper, let us simulate the probability of getting the slip of paper with the largest number when the strategy is to let pass the first L slips of paper and then pick the first one with the highest number so far. Here L is a given value with

$1 \leq L < 20$. In the simulation, it is convenient to assign the rank number 1 to the slip with the highest number, the rank number 2 to the slip with the second highest number, etc. In a simulation run you first generate a random permutation $a[1], \ldots, a[20]$ of the integers $1, \ldots, 20$. Then, you determine the smallest value among $a[1], \ldots, a[L]$, say the value m. Next, pick the first $f > L$ with $a[f] < m$ if such a f exists, otherwise let $f = 20$. The simulation run is said to be successful if $a[f] = 1$. By making a large number of simulation runs, you can estimate the desired probability by dividing the number of successful runs by the total number of runs. Repeating the simulation experiment for several values of the critical level L leads to the optimal value of L.

2.9.5 Simulating a random subset of integers

How does a computer generate the Lotto 6/45 "Quick Pick," that is, six different integers from the integers $1, \ldots, 45$? More generally, how does the computer generate randomly r different integers from the integers $1, \ldots, n$? This is accomplished by following the first r iteration steps of the above algorithm for a random permutation until the positions $n, n-1, \ldots, n-r+1$ are filled. The elements $a[n], \ldots, a[n-r+1]$ in these positions constitute the desired random subset.

2.9.6 Simulation programs

In the field of physics, it is quite common to determine the values of certain constants in an experimental way. Computer simulation makes this kind of approach possible in the field of mathematics too. For example, the value of π can be estimated with the help of some basic principles of simulation (of course, this is not the simplest method for the calculation of π). This is the general idea: take a unit circle (radius = 1) with the origin (0, 0) as middle point. In order to generate random points inside the circle, position the unit circle in a square that is described by the four corner points $(-1, 1)$, $(1, 1)$, $(1, -1)$, and $(-1, -1)$. The area of the unit circle is π and the area of the square is equal to 4. Now, generate a large number of points that are randomly spread out over the surface of the square. Next, count the number of points that fall within the surface of the unit circle. If you divide this number of points by the total number of generated points, you get an estimate for $\frac{\pi}{4}$. You can identify a blindly chosen point (x, y) in the square by generating two random numbers u_1 and u_2 between 0 and 1 and then taking $x = -1 + 2u_1$ and $y = -1 + 2u_2$. Point (x, y), then, only belongs to the unit circle if $x^2 + y^2 \leq 1$. The *hit-or-miss method* used to generate random points inside the circle can also be used to generate random points in any given bounded region in the plane or in other higher-dimensional spaces. The idea of the hit-or-miss method was introduced to the statistics community by physicists N. Metropolis and S. Ulam in their article “The Monte Carlo method,” *Journal of the American Statistical Association* **44** (1949): 335–341. In this article, Metropolis and Ulam give a classic example of finding the volume of a 20-dimensional region within a unit cube when the required multiple integrals are intractable. Taking a large number of points at random inside this cube and counting how many of these points satisfy all the given inequalities that defined the region, they estimate the volume of the region.

Below, for the interested reader, we have listed a Pascal program for simulating the value of π. Pascal was chosen as the programming language because of its clarity and readability. In Pascal, all information occurring between brackets { } is commentary included for instructional purposes and is not read by the computer.

```
PROGRAM SimulatePi(Input, Output);

CONST
  n = 10000; { number of runs }
VAR
  k : Integer; { current run }
  DartsInCircle : Integer; { total number of hits }
  Fraction : Real; { proportion of hits }
  Pi : Real; { estimate of pi }

PROCEDURE ThrowDart;
  VAR
    x, y : Real; {(x,y) coordinates of the position of
                  the dart }
  BEGIN { x and  y are randomly sampled from [-1,1] }
    x := -1+2*Random;
    y := -1+2*Random;
    { Test whether (x,y) falls within the unit circle }
    IF (Sqr(x) + Sqr(y) <= 1) THEN
       DartsInCircle := DartsInCircle + 1;
  END;

BEGIN { main program }
  Randomize; { initializes the random-number generator
               on your computer}
  DartsInCirkel := 0;
  { execute the simulation }
  for k := 1 to n do
    ThrowDart;
  { compute the desired results}
```

```
    Fraction := DartsInCircle / n;
    Pi := 4*Fraction;
    WriteLn('the simulated value of pi is:',Pi);

  END.
```

This simulation program requires only a minor adjustment to verify experimentally that the volume of a sphere with radius r is equal to $4\pi r^3/3$ (try it!). In general, computer simulation can be used for the numerical evaluation of (complicated) integrals. It suffices to have a random-number generator. Also, the reader might find computer simulation a quick and useful approach to solve geometric probability problems that are otherwise not easily amenable to an analytical approach. Several examples of such problems are given in Problem 2.21.

We close this chapter with a simulation program for the daughter-son problem from Chapter 1. This problem can lead to heated discussions over the right answer. The right answer to the first question posed in this problem is $\frac{1}{3}$, but the probability in question changes to $\frac{1}{2}$ in the second situation discussed in the problem. In Chapter 6, we give a probabilistic derivation of these answers based on the assumption that the probability of a newborn infant being a girl is the same as the probability of its being a boy. In answering the second question, we have also made an assumption that, randomly, one of the two children will open the door. If it is assumed that when a family has one boy and one girl, the girl always opens the door, the situation changes dramatically. We give, here, a simulation program that is bound to convince mathematicians and non-mathematicians alike of the correctness of the answer. Writing a simulation program forces you to make implicit assumptions explicit and, in every step, to indicate precisely what you mean, unambiguously. In this way, you fall into fewer mental traps than would be the case if you were giving a purely verbal account.

```
  PROGRAM Family(Input, Output);

  CONST
    number_runs = 10000; { total number of simulation
                           runs}

  VAR
    number_d : Integer; { total number of runs with 1 or
                          2 daughters }
    number_dd : Integer; { total number of runs with
                           2 daughters }
    number_d_opens : Integer;
    { total number of runs in which a daughter opens
      the door }
  k : integer; { index of the current run }

PROCEDURE init;
  BEGIN
     number_d := 0;
     number_dd := 0;
     number_d_opens := 0;
  END;

PROCEDURE run;
  VAR
    composition: (DD, DS, SD, SS);
        { composition of the family, D=daughter,
          S=son }
    opens_door : (D, S};
        { daughter or son opens the door}
    drawing : Integer; { random drawing from 1,2,3,4 }
  BEGIN
    { determine the composition of the family: draw
      from 1,2,3,4 }
     drawing := 1 + trunc(4*Random);
     CASE drawing OF
       1 : composition := DD;
       2 : composition := DS;
       3 : oomposition := SD;
       4 : composition := SS;
     END;

     { determine whether daughter or son opens the
       door }
     IF ( composition = DD) THEN
       opens_door := D
     ELSE IF ( composition = SS) THEN
       opens_door := S
     ELSE BEGIN { composition is DS of SD: additional
                  random number is needed to determine
                  who opens the door, D or S; each
                  possibility has a probability
                  of 0.5 }
       IF (Random < 0.5) THEN
          opens_door := D
       ELSE
          opens_door := S;
     END;
```

```
      { update counters }
      IF (composition <> SS) THEN
        number_d := number_d + 1;
      IF (composition = DD) THEN
        number_dd := number_dd + 1;
      IF (opens_door = D) then
        number_d_opens := number_d_opens + 1;

   END;

 BEGIN { main program }
    Randomize; { initialize the  random-number generator
                 of your computer }
    { execute the simulation }
    init;
    for k := 1 to number_runs do
       run;
    { results }
    Writeln('probability of two daughters given one
             daughter :',
       number_dd/number_d);
    Writeln('probability of two daughters given a
             daughter opens:',
       number_dd/number_d_opens);
 END.
```

2.10 Problems

2.1 On a modern die the face value 6 is opposite to the face value 1, the face value 5 to the face value 2, and the face value 4 to the face value 3. In other words, by turning a die upside down, the face value k is changed into $7 - k$. This fact may be used to explain why when rolling three dice the totals 9 and 12 ($= 3 \times 7 - 9$) are equally likely. Old Etruscan dice show 1 and 2, 3 and 4, and 5 and 6 on opposite sides. Would the totals 9 and 12 remain equally likely when rolling three Etruscan dice?

2.2 In the television program “Big Sisters,” 12 candidates remain. The public chooses four candidates for the final round. Each candidate has an equal probability of being chosen. The Gotham Echo reckons that the local heroine, Stella Stone, has a probability of 38.5% of getting through to the final: they give her a $\frac{1}{12}$ probability of being chosen first, a $\frac{1}{11}$ probability of being chosen second, a $\frac{1}{10}$ probability of being chosen third, and a $\frac{1}{9}$ probability of being chosen fourth. Is this calculation correct?

2.3 A dog has a litter of four puppies. Set up a probability model to answer the following question. Can we correctly say that the litter more likely consists of three puppies of one gender and one of the other than that it consists of two puppies of each gender?

2.4 Answer each of the following four questions by choosing an appropriate sample space and assigning probabilities to the various elements of the sample space.

(a) In Leakwater township, there are two plumbers. On a particular day three Leakwater residents call village plumbers independently of each other. Each resident randomly chooses one of the two plumbers. What is the probability that all three residents will choose the same plumber?

(b) You roll a fair die three times in a row. What is the probability that the second roll will deliver a higher point count than the first roll and the third roll a higher count than the second?

(c) Two players A and B each roll one die. The absolute difference of the outcomes is computed. Player A wins if the difference is 0, 1, or 2; otherwise, player B wins. Is this a fair game?

2.5 Use an appropriate sample space with equiprobable elements to answer the following question. You enter a grand-prize lottery along with nine other people. Ten numbered lots, including the winning lot, go into a box. One at a time, participants draw a lot out of the box. Does it make a difference to your chance of winning whether you are the first or the last to draw a lot?

2.6 In the daily lottery game "Guess 3," three different numbers are picked randomly from the numbers $0, 1, \ldots, 9$. The numbers are picked in order. To play this game, you must choose between "Exact order" and "Any order" on the entry form. In either case, the game costs \$1 to play. Should you choose to play "Exact order," you must tick three different numbers in the order you think they will be picked. If those numbers are picked in that order, you win a \$360 payoff. Should you opt to play "Any order," you tick three numbers without regard for their order of arrangement. You win a \$160 payoff if those three numbers are picked. Set up a probability model to calculate the expected value of the payoff amount for both options.

2.7 In the dice game known as "seven," two fair dice are rolled and the sum of scores is counted. You bet on "manque" (that a sum of 2, 3, 4, 5 or 6 will result) or on "passe" (that a sum of 8, 9, 10, 11 or 12 will result). The sum of 7 is a fixed winner for the house. A winner receives a payoff that is double the amount staked on the game. Nonwinners forfeit the amount staked. Define an appropriate probability space for this experiment. Then calculate the expected value of the payoff per dollar staked.

2.8 Sic Bo is an ancient Chinese dice game that is played with three dice. There are many possibilities for betting on this game. Two of these are "big" and "small." When you bet "big," you win if the total points rolled equals 11, 12, 13, 14, 15, 16 or 17, except when three 4's or three 5's are rolled. When you bet "small," you win if the total points rolled equals 4, 5, 6, 7, 8, 9 or 10, except when three 2's or three 3's are rolled. Winners of "big" and "small" alike receive double the amount staked on the game. Calculate the house percentage for each of these betting formats.

2.9 Consider the Kelly betting model from Section 2.7. In addition to the possibility of investing in a risky project over a large number of successive periods, you can get a fixed interest rate at the bank for the portion of your capital that you do not invest. You can reinvest your money at the end of each period. Let the interest rate be r, i.e., every dollar you do not invest in a certain period will be worth $1 + r$ dollars at the end of the period. The expected value of the payoff of the risky project satisfies $pf > 1 + r$.

(a) For the growth factor G_n in the representation $V_n = e^{nG_n} V_0$ show that it holds true that

$$\lim_{n\to\infty} G_n = p \ln[(1-\alpha)(1+r) + \alpha f] + (1-p) \ln[(1-\alpha)(1+r)].$$

Verify that this expression is maximal for $\alpha^* = \frac{pf-(1+r)}{f-(1+r)}$.

(b) Suppose you are faced with a 100%-safe investment returning 5% and a 90%-safe investment returning 25%. Calculate how to invest your money using the Kelly strategy. Calculate also the effective rate of return on your investment over the long-term.

2.10 A particular game pays f_1 times the amount staked with a probability of p and f_2 times the amount staked with a probability of $1-p$, where $f_1 > 1, 0 \leq f_2 < 1$ and $pf_1 + (1-p)f_2 > 1$. You play this game a large number of times and each time you stake the same fraction α of your bankroll. Verify that the Kelly fraction is given by

$$\alpha^* = \min\left(\frac{pf_1 + (1-p)f_2 - 1}{(f_1 - 1)(1 - f_2)}, 1\right)$$

with $(1-\alpha^* + \alpha^* f_1)^p (1-\alpha^* + \alpha^* f_2)^{1-p} - 1$ as the corresponding effective rate of return over the long-term.

2.11 In a group of 25 people, a person tells a rumor to a second person, who in turns tells it to a third person, and so on. Each person tells the rumor to just one of the people chosen at random, excluding the person from whom he/she heard the rumor. The rumor is told 10 times. What is the probability that the rumor will not be repeated to any one person once more? What is the probability that the rumor will not return to the originator? Use simulation to find the expected value of the number of persons having knowledge of the rumor.

2.12 Three players, A, B, and C, each put ten dollars into a pot with a list on which they have, independently of one another, predicted the outcome of three successive tosses of a fair coin. The fair coin is then tossed three times. The player having most correctly predicted the three outcomes gets the contents of the pot. The contents are to be divided if multiple players guess the same number of correct outcomes.

(a) Calculate the expected value of the amount that player A will get.

(b) Suppose that players A and B decide to collaborate, unbeknownst to player C. The collaboration consists of the two players agreeing that the list of player B will always be a mirror image of player A's list (should player A predict an outcome of HTT, for example, then player B would predict TTH). Calculate the expected value of the amount that player A will receive.

2.13 The following game is played in a particular carnival tent. The carnival master has two covered beakers, each containing one die. He shakes the beakers thoroughly, removes the lids and peers inside. You have agreed that whenever at least one of the two dice shows an even number of points, you will bet with even odds that the other die will also show an even number of points. Is this a fair bet?

2.14 Three players enter a room and are given a red or a blue hat to wear. The color of each hat is determined by a fair coin-toss. Players cannot see the color of their own hats, but do see the color of the other two players' hats. The game is won when at least one of the players correctly guesses the color of his own hat and no player gives an incorrect answer. In addition to having the opportunity to guess a color, players may also pass. Communication of any kind between players is not permissible after they have been given hats; however, they may agree on a group strategy beforehand. Verify that there is a group strategy that results in a $\frac{3}{4}$ probability of winning. (This puzzle was discussed in the *New York Times* of April 10, 2001. The hat problem with many players is related to problems in coding theory. The strategy gets far more complicated for larger numbers of players. In the game with $2^m - 1$ players, there is a strategy for which the group is victorious with a probability of $(2^m - 1)/2^m$).

2.15 At a completely random moment between 6:30 and 7:30 a.m., the morning newspaper is delivered to Mr. Johnson's residence. Mr. Johnson leaves for work at a completely random moment between 7:00 and 8:00 a.m. regardless of whether the newspaper has been delivered. What is the probability that Mr. Johnson can take the newspaper with him to work? Use computer simulation to find the probability.

2.16 You choose three points at random inside a square. Then choose a fourth point at random inside the square. What is the probability that the triangle formed by the first three points is obtuse? What is the probability that the fourth point will fall inside this triangle? What are the probabilities when the points are chosen at random inside a circle?

2.17 Use computer simulation to find the probability that the quadratic equation $Ax^2 + Bx + C = 0$ has real roots when A, B, and C are chosen at random from the interval $(-q, q)$, independently of each other. Also, use simulation to find this probability when A, B, and C are nonzero integers that are chosen at random between $-q$ and q, independently of each other. Vary q as 1, 10, 100, 1,000, and 10,000.

2.18 Solve Problem 2.17 again for the situation in which the coefficient A is fixed at the value 1.

2.19 Use computer simulation to find the probability that the triangle OAB has an angle larger than 90° when A and B are randomly chosen points within the unit circle having the point O as center. What is this probability if the unit sphere is taken instead of the unit circle? Also, simulate the probabilities of getting a triangle with an obtuse angle from three random points in the unit circle and from three random points in the unit sphere.

2.20 A stick is broken at random into two pieces. You bet on the ratio of the length of the longer piece to the length of the smaller piece. You receive \$$k$ if the ratio is between k and $k + 1$ for some $1 \leq k \leq m - 1$, while you receive \$$m$ if the ratio is larger than m. Here m is a given positive integer. Using computer simulation, verify that your expected payoff is approximately equal to $\$2[\ln(m + 1) - 0.4228 + 2/(m + 1)]$. Do you see a resemblance with the St. Petersburg paradox?

2.21 Use computer simulation to find

(a) The expected value of the distance between two points that are chosen at random inside the interval (0, 1).

(b) The expected value of the distance between two points that are chosen at random inside the unit square.

(c) The expected value of the distance between two points that are chosen at random inside the unit circle.

(d) The expected value of the distance between two points that are chosen at random inside an equilateral triangle with sides of unit length.

2.22 A millionaire plays European roulette every evening for pleasure. He begins every time with $A = 100$ chips of the same value and plays on until he has gambled away all 100 chips. When he has lost his 100 chips for that evening's entertainment, he quits. Use computer simulation to find the average number of times the millionaire will play per round, for the Big–Martingale betting system and for the D'Alembert betting system. Also determine the probability that on a given evening the millionaire will acquire $B = 150$ chips before he is finished playing. Do the same for $A = 50$ and $B = 75$. Can you give an intuitive explanation for why the average value of the total number of chips the millionaire stakes per evening is equal to $37A$ over the long-term, regardless of the betting system he uses?

2.23 You decide to bet on ten spins of the roulette wheel in European roulette and to use the double-up strategy. Under this strategy, you bet on red each time and you double your bet if red does not come up. If red comes up, you go back to your initial bet of 1 euro. Use computer simulation to find the expected value of your loss after a round of ten bets and to find the expected value of the total amount bet during a round. Can you explain why the ratio of these two expected values is equal to $\frac{1}{37}$?

2.24 Seated at a round table, five friends are playing the following game. One of the five players opens the game by passing a cup to the player seated either to his left or right. That player, in turn, passes the cup to a player on his left or right and so on until the cup has progressed all the way around the table. As soon as one complete round has been achieved, the player left holding the cup pays for a round of drinks. A coin-toss is performed before each turn in order to determine whether the cup will go to the left or right. Use computer simulation to find, for each player, the probability that the player will have to buy a round of drinks.

2.25 What is the probability that any two adjacent letters are different in a random permutation of the 11 letters of the word Mississippi? What is the probability that in a thoroughly shuffled deck of 52 cards no two adjacent cards are of the same rank? Use computer simulation.

2.26 You have been asked to determine a policy for accepting reservations for an airline flight. This particular flight uses an aircraft with 15 first-class seats and 75 economy-class seats. First-class tickets on the flight cost \$500 and economy-class tickets cost \$250. The number of individuals who seek to reserve seats takes on the equally likely values of $10, 11, \ldots, 20$ in first class and the equally likely values $40, 41, \ldots, 120$ in economy class. The demand for first-class seats and that for economy-class seats are independent. The airline allows itself to sell somewhat more tickets than it has seats. This is a common practice called overbooking. You are asked to analyze the four possible booking policies which permit the overbooking of either up to 0, or up to 3 first-class seats and the overbooking of either up to 5, or up to 10 economy seats. Each passenger who buys a first-class seat has a 10% probability of not showing up for the flight. The probability of not showing up is 5% for economy-class passengers. Passengers decide whether to show up independently of each other. First-class passengers who do not show up can return their unused tickets for a full refund. No-shows in the economy class are not entitled to any refund. Any first-class passengers who show up for the flight but cannot be seated in the first class are entitled to a full refund plus $400 compensation. If there are free seats in first class and economy class is full, economy-class passengers can be seated in first class. If an economy-class passenger shows up and is denied a seat, however, they get a full refund plus $200 compensation. Use computer simulation to find for each of the four possible overbooking policies both the expected value of the net profit for the flight and the probabilities of the net profit falling in each of the ranges [$15,000, $16,000], . . . ,[$27,000, $28,000].

2.27 The card game called Ace-Jack-Two is played between one player and the bank. It goes this way: a deck of 52 cards is shuffled thoroughly, after which the bank repeatedly reveals three cards next to each other on a table. If an ace, jack or two is among the three cards revealed, the bank gets a point. Otherwise, the player gets a point. The points are tallied after 17 rounds are played. The one with the most points is the winner. Use computer simulation to determine the probability of the bank winning and the average number of points that the bank will collect per game.

2.28 Consider the best-choice problem from Section 2.3 with 100 slips of paper. You let the first 30 slips of paper go by and then pick the first one to come along thereafter that contains a higher number than was seen in the first 30 slips. Use computer simulation to find the probability of obtaining either the largest or the second largest possible number. What is the probability of getting one of the three largest numbers?

2.29 Two candidates A and B remain in the finale of a television game show. At this point, each candidate must spin a wheel of fortune. The 20 numbers 5, 10, . . . , 95, 100 are listed on the wheel and when the wheel has stopped spinning, a pointer randomly stops on one of the numbers. Each candidate has a choice of spinning the wheel one or two times, whereby a second spin must immediately follow the first. The goal is to reach a total closest to but not exceeding 100 points. The winner is the candidate who gets the highest score. Should there be a tie, then the candidate to spin the wheel first is the winner. The candidate who spins second has the advantage of knowing what the score of the first candidate was. Lots are drawn to determine which player begins. Suppose that candidate A has to spin first. His/her strategy is to stop after the first spin if this spin gives a score larger than a certain level L and otherwise to continue for a second spin. Use computer simulation to find the optimal value of the stopping level L and the maximal probability of candidate A winning.

2.30 Reconsider Problem 2.29 with three candidates A, B, and C. Candidate A spins first; candidate B, second; and candidate C, last.

(a) Use the optimal stopping rule found in Problem 2.29 to describe the optimal strategy of candidate B.

(b) Use the result of (a) and computer simulation to determine the optimal stopping rule for candidate A. What is the maximal probability of candidate A winning and what is the maximal probability of candidate B winning?

2.31 Using five dice, you are playing a game consisting of accumulating as many points as possible in five rounds. After each round you may “freeze” one or more of the dice, i.e., a frozen die will not be rolled again in successive rounds, but the amount of points showing will be re-counted in successive rounds. You apply the following strategy: if there are still i rounds to go, you freeze a die only when it displays more than α_i points, where $\alpha_4 = 5$, $\alpha_3 = 4$, $\alpha_2 = 4$, $\alpha_1 = 3$ and $\alpha_0 = 0$. A grand total of s points results in a payoff of $s - 25$ dollars if $s \geq 25$, and a forfeiture of $25 - s$ dollars if $s < 25$. Use computer simulation to find the expected value of the payoff. What is the probability that your grand total will be 25 or more points.

2.32 Solve the following problems for the coin-tossing experiment:

(a) Use computer simulation to find the probability that the number of heads ever exceeds twice the number of tails if a fair coin is tossed 5 times. What is the probability if the coin is tossed 25 times. What is the probability if the coin is tossed 50 times? Verify experimentally that the probability approaches the value $\frac{1}{2}(\sqrt{5} - 1)$ if the number of tosses increases.

(b) A fair coin is tossed no more than n times, where n is fixed in advance. After each toss, you can decide to stop the coin-toss experiment. Your payoff is 1,000 dollars multiplied by the proportion of heads at the moment the experiment is stopped. Your strategy is to stop as soon as the proportion of heads exceeds $\frac{1}{2}$ or as soon as n tosses are done, whichever occurs first. Use computer simulation to find your expected payoff for $n = 5$, 10, and 25. Verify experimentally that your expected payoff approaches the value $\frac{1}{4}\pi$ times \$1,000 if n becomes large. Can you devise a better strategy than the one proposed?

2.33 In a TV program, the contestant can win one of three prizes. The prizes consist of a first prize and two lesser prizes. The dollar value of the first prize is a five-digit number and begins with 1, whereas the dollar values of the lesser prizes are three-digit numbers. There are initially four unexposed digits in the value of first prize and three in each of the values of the other two prizes. The game involves trying to guess the digits in the dollar value of the first prize before guessing the digits in either of the dollar values of the other two prizes. Each of the digits 0–9 is used only once among the three prizes. The contestant chooses one digit at a time until all of the digits in the dollar value of one of the three prizes have been completed. What is the probability that the contestant will win the first price? Use computer simulation to find this probability.

2.34 A random sequence of 0's and 1's is generated by tossing a fair coin N times. A 0 corresponds to the outcome heads and a 1 to the outcome tails. A run is an uninterrupted sequence of 0's or 1's only. Use computer simulation to verify experimentally that the length of the longest run exhibits little variation and has its probability mass concentrated around the value $\log_2(N) - \frac{2}{3}$ when N is sufficiently large.

2.35 You are playing the following game: a fair coin is tossed until it lands heads three times in a row. You get 12 dollars when this occurs, but you must pay one dollar for each toss. Use computer simulation to find out whether this is a fair contest.

2.36 A drunkard is standing in the middle of a very large town square. He begins to walk. Each step he takes is a unit distance in a randomly chosen direction. The direction for each step taken is chosen independently of the direction of the others. Suppose that the drunkard takes a total of n steps for a given value of n. Verify experimentally that the expected value of the quadratic distance between the starting and ending points is equal to n, whereas the expected value of the distance between starting and ending points is approximately equal to $0.886\sqrt{n}$ if n is sufficiently large. Also, for $n = 25$ and $n = 100$, find the probability that the maximal distance of the drunkard to his starting point during the n steps will exceed $1.18\sqrt{n}$.

2.37 A particle moves over the flat surface of a grid such that an equal unit of distance is measured with every step. The particle begins at the origin (0,0). The first step may be to the left, right, up or down, with equal probability $\frac{1}{4}$. The particle cannot move back in the direction that the previous step originated from. Each of the remaining three directions has an equal probability of $\frac{1}{3}$. Suppose that the particle makes a total of n steps for a given value of n. Verify experimentally that the expected value of the distance between the particle's starting and ending points is approximately equal to $1.25\sqrt{n}$ if n is sufficiently large. Also, for $n = 25$ and $n = 100$, find the probability that the maximal distance of the particle to its starting point during the n steps will exceed $1.65\sqrt{n}$.

2.38 You have received a reliable tip that in the local casino the roulette wheel is not exactly fair. The probability of the ball landing on the number 13 is twice what it should be. The roulette table in question will be in use that evening. In that casino, European roulette is played. You go with 1,000 euros and intend to make 100 bets. Your betting strategy is as follows: each time you stake a multiple of five euros on the number 13 and you choose that multiple that is closest to 2.5% of your bankroll. You will receive a payoff of 36 times the amount staked if the ball lands on 13. Use computer simulation to determine the probability distribution of your bankroll at the end of the night. Specifically, determine the probability of your leaving the casino with more than 2,000 euros.

2.39 Sixteen teams remain in a soccer tournament. A drawing of lots will determine which eight matches will be played. Before the drawing takes place, it is possible to place bets with bookmakers over the outcome of the drawing. Use computer simulation to find the probability of correctly predicting i matches for $i = 0, 1, 2,$ and 3.

2.40 One hundred passengers line up to board an airplane with 100 seats. Each passenger is to board the plane individually, and must take his or her assigned seat before the next passenger may board. However, the passenger first in line has lost his boarding pass and takes a random seat instead. This passenger randomly selects another unoccupied seat each time it appears that he is not occupying his assigned seat. Use simulation to find the probability of the passenger changing seats five or more times before getting to his assigned seat. *Hint*: number the passengers in line as $1, 2, \ldots, 100$ and number their assigned seats accordingly.

2.41 Each of seven dwarfs has his own bed in a common dormitory. Every night, they retire to bed one at a time, always in the same sequential order. On a particular evening, the youngest dwarf, who always retires first, has had too much to drink. He randomly chooses one of the seven beds to fall asleep on. As each of the other dwarfs retires, he chooses his own bed if it is not occupied, and otherwise randomly chooses another unoccupied bed. Use computer simulation to find for $k = 1, 2, \ldots, 7$ the probability that the kth dwarf to retire can sleep in his own bed. This variant of the lost boarding pass puzzle is due to the Danish mathematician Henning Makholm.

2.42 A queue of 50 people is waiting at a box office in order to buy a ticket. The tickets cost five euros each. For any person, there is a probability of $\frac{1}{2}$ that she/he will pay with a five-euro note and a probability of $\frac{1}{2}$ that she/he will pay with a ten-euro note. When the box opens there is no money in the till. If each person just buys one ticket, what is the probability that none of them will have to wait for change? Use computer simulation.

2.43 Independently of each other, ten numbers are randomly drawn from the interval (0, 1). You may view the numbers one by one in the order in which they are drawn. After viewing each individual number, you are given the opportunity to take it or let it pass. You are not allowed to go back to numbers you have passed by. Your task is to pick out the highest number. Your strategy is as follows. If, after you have viewed a number, there are still k numbers left to view, you will take the number if it is the highest number to appear up to that point and if it is higher than a critical level a_k, where $a_0 = 0, a_1 = 0.500, a_2 = 0.690, a_3 = 0.776, a_4 = 0.825, a_5 = 0.856, a_6 = 0.878, a_7 = 0.894, a_8 = 0.906$, and $a_9 = 0.916$. Use simulation to determine the probability that you will pick out the highest number.

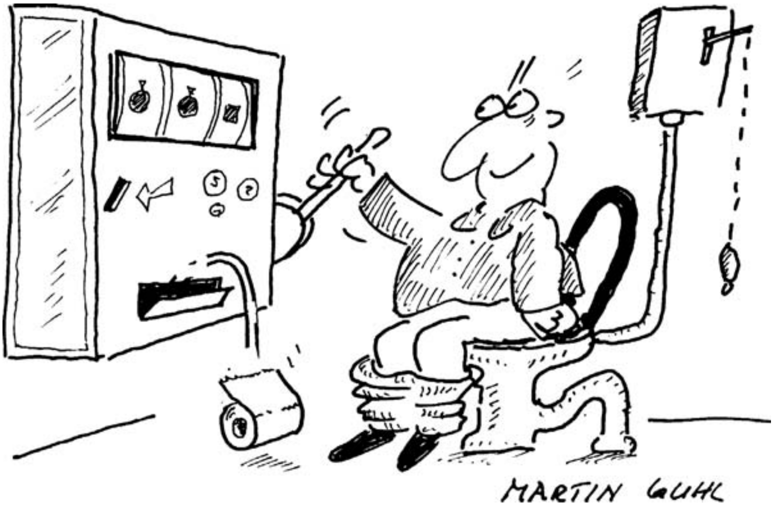
2.44 In a certain betting contest you may choose between two games A and B at the start of every turn. In game A you always toss the same coin, while in game B you toss either coin 1 or coin 2 depending on your bankroll. In game B you must toss coin 1 if your bankroll is a multiple of three; otherwise, you must toss coin 2. A toss of the coin from game A will land heads with a probability of $\frac{1}{2} - \epsilon$ and tails with a probability of $\frac{1}{2} + \epsilon$, where $\epsilon = 0.005$. Coin 1 in game B will land heads with probability $\frac{1}{10} - \epsilon$ and tails with probability $\frac{9}{10} + \epsilon$; coin 2 in game B will land heads with probability $\frac{3}{4} - \epsilon$ and tails with probability $\frac{1}{4} + \epsilon$. In each of the games A and B, you win one dollar if heads is thrown and you lose one dollar if tails is thrown. An unlimited sequence of bets is made in which you may continue to play even if your bankroll is negative (a negative bankroll corresponds to debt). Following the strategy $A, A, \ldots,$ you win an average of 49.5% of the bets over the long-term. Use computer simulation to verify that using strategy $B, B, \ldots,$ you will win an average of 49.6% of the bets over the long-term, but that using strategy $A, A, B, B, A, A, B, B, \ldots$ you will win 50.7% of the bets over the long-term. (The paradoxical phenomenon, that in special betting situations winning combinations can be made up of individually losing bets, is called *Parrondo's paradox* after the Spanish physicist Juan Parrondo, see also G.P. Harmer and D. Abbott, "Losing strategies can win by Parrondo's paradox," *Nature*, **402**, 23/30 December 1999. An explanation of the paradox lies in the dependency between the betting outcomes. Unfortunately, such a dependency is absent in casino games.)

2.45 Center court at Wimbledon is buzzing with excitement. The dream finale between Alassi and Bicker is about to begin. The weather is fine, and both players are in top condition. In the past, these two players have competed multiple times under similar conditions. On the basis of past outcomes, you know that 0.631 and 0.659 give the respective probabilities that Alassi and Bicker will win their own service points when playing against each other. Use computer simulation to determine the probability of Alassi winning the finale. Now assume that the first set has been played and won by Alassi, and that the second set is about to begin. Bookmakers are still accepting bets. What is now Alassi's probability of winning the finale?

3

Probabilities in everyday life

Computer simulation can be extremely useful to those who are trying to develop an understanding of the basic concepts of probability theory. The previous chapter recommended simulation as a means of explaining such phenomena as chance fluctuations and the law of large numbers. Fast computers allow us to simulate models swiftly and to achieve a graphic rendering of our outcomes. This naturally enhances our understanding of the laws of probability theory.



Monte Carlo simulation is the name given to the type of simulation used to solve problems that contain a random element. In such simulations, the computer's random-number generator functions as a sort of roulette wheel. Monte Carlo simulation is widely applicable. Often, it is the only possible method of solving probability problems. This is not to say, however, that it does not have its limitations. It is not a "quick fix" to be applied haphazardly. Before beginning, one must think carefully about the model to be programmed. The development of simulation models for complex problems can require a lot of valuable time. Monte Carlo simulation gives numerical results, but the vast amounts of numerical data resulting can make it difficult to draw insightful conclusions, and insight is often more important than the numbers themselves. In general, the mathematical solution of a model will render both numbers and insight.[†] In practice, then, a purely mathematical model that is limited to the essentials of a complex problem can be more useful than a detailed simulation model. Sometimes a combination of the two methods is the most useful, as in the use of simulation to test the practical usefulness of results gained from a simplified mathematical equation.

In this chapter, we will discuss a number of interesting probability problems. We will solve each of these problems twice: first by means of simulation and subsequently by means of a theoretical model that can be solved mathematically. The first problem we will tackle is the birthday problem, one of the most surprising problems in the field of probability theory. Thereafter, we will look at a few of the casino problems encountered in the games of craps and roulette, and a scratch-lottery problem. The common element in all of these problems is that they playfully demonstrate some of the important concepts and solution methods used in the field of probability theory.

3.1 The birthday problem

The birthday problem is very well known in the field of probability theory. It raises the following interesting questions: what is the probability that, in a group of randomly chosen people, at least two of them will have been born on the same day of the year? How many people are needed to ensure a probability greater than 0.5? Excluding February 29 from our calculations and assuming that the remaining 365 possible birthdays are all equally probable, we may be surprised to realize that, in a group of only 23 people, the probability of two people having the same birthday is greater than 0.5 (the exact probability is 0.5073). Then, again, perhaps this result is not so very surprising: think back to your school days and consider how often two or more classmates celebrated birthdays on the same day. In Section 4.2.3 of Chapter 4, further insight will be given to

[†] Simulation may even be inadequate in some situations. As an example, try to find by simulation the average value of the ratio of the length of the longer piece to that of the shorter piece of a broken stick when many sticks are broken at random into two pieces. The analytical solution to this problem can be found in Section 10.1.3.

the fact that a group of only 23 people is large enough to have about a 50–50 chance of at least one coincidental birthday. What about the assumption that birthdays are uniformly distributed throughout the year? In reality, birthdays are not uniformly distributed. The answer is that the probability of a match only becomes larger for any deviation from the uniform distribution. This result can be mathematically proved. Intuitively, you might better understand the result by thinking of a group of people coming from a planet on which people are always born on the same day.

3.1.1 Simulation approach

A simulation model is easily constructed. Imagine that you want to calculate the probability of two people out of a group of 23 randomly chosen people having their birthdays on the same day. In each simulation experiment, 23 random drawings will be made out of the numbers $1, \dots, 365$. A random drawing from these numbers is given by $1 + \lfloor 365u \rfloor$ when u is a random number between 0 and 1 (see Section 2.9). A simulation experiment is said to be successful when the same number is drawn at least twice. After a sufficiently large number of experiments the probability of at least two persons having the same birthday can be estimated by

$$\frac{\text{number of successful simulation experiments}}{\text{total number simulation experiments}}.$$

3.1.2 Theoretical approach

In order to calculate the probability of two people in a randomly chosen group of n people having birthdays on the same day, the following approach is applicable. First, calculate the *complementary probability*, i.e., the probability of no two birthdays falling on the same day. This probability is simpler to calculate.† Imagine that the n people are numbered in order from $1, \dots, n$. There are 365^n outcomes for the possible birth dates of the n ordered people. Each of these outcomes is equally probable. The number of outcomes showing no common birthdays is equal to $365 \times 364 \times \cdots \times (365 - n + 1)$. The probability then, of no two of the n people having a common birthday, is equal to $365 \times 364 \times \cdots \times (365 - n + 1)$ divided by 365^n. From this it follows that, in a group

† The simple technique of working with complementary probabilities is also handy in the solution of the De Méré problem described in the Introduction to this book: the probability of rolling a double six in n rolls of a fair pair of dice is equal to 1 minus the complementary probability of rolling no double sixes at all in n rolls ($= 1 - \frac{35^n}{36^n}$).

Table 3.1. *Probabilities for the birthday problem.*

n	15	20	23	25	30	40	50	75
p_n	0.2529	0.4114	0.5073	0.5687	0.7063	0.8912	0.9704	0.9997

of n people, the probability of two people having the same birthday can be given by

$$p_n = 1 - \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}.$$

In Table 3.1, the probability p_n is given for various values of n. It is surprising to see how quickly this probability approaches 1 as n grows larger. In a group of 75 people it is practically certain that at least two people will have the same birthday. An approximation formula for probability p_n shows how quickly p_n increases as n grows larger. In Problem 3.12 you are asked to derive the approximation formula

$$p_n \approx 1 - e^{-\frac{1}{2}n(n-1)/365}.$$

We come back to this approximation formula in Section 4.2.3 of Chapter 4.

John Allen Paulos' *Innumeracy* contains a wonderful example of the misinterpretation of probabilities in everyday life. On late-night television's The Tonight Show with Johnny Carson, Carson was discussing the birthday problem in one of his famous monologues. At a certain point, he remarked to his audience of approximately 100 people: "Great! There must be someone here who was born on my birthday!" He was off by a long shot. Carson had confused two distinctly different probability problems: (1) the probability of one person out of a group of 100 people having the same birth date as Carson himself, and (2) the probability of any two or more people out of a group of 101 people having birthdays on the same day. How can we calculate the first of these two probabilities? First we must recalculate the complementary probability of no one person in a group of 100 people having the same birth date as Carson. A random person in the group will have a probability of $\frac{364}{365}$ of having a different birth date than Carson. The probability of no one having the same birthday as Carson is equal to $(\frac{364}{365}) \times \cdots \times (\frac{364}{365}) = (\frac{364}{365})^{100}$. Now, we can calculate the probability of at least one audience member having the same birthday as Carson to be equal to $1 - (\frac{364}{365})^{100} = 0.240$ (and not 0.9999998). Verify for yourself that the audience would have had to consist of 253 people in order to get about a 50-50 chance of someone having the same birthday as Carson.

3.1.3 Another birthday surprise

On Wednesday, June 21, 1995, a remarkable thing occurred in the German Lotto 6/49, in which six different numbers are drawn from the numbers 1, . . . , 49. On the day in question, the mid-week drawing produced this six-number result: 15-25-27-30-42-48. These were the same numbers as had been drawn previously on Saturday, December 20, 1986, and it was for the first time in the 3,016 drawings of the German Lotto that the same sequence had been drawn twice. Is this an incredible occurrence, given that in German Lotto there are nearly 14 million possible combinations of the six numbers in question? Actually, no, and this is easily demonstrated if we set the problem up as a birthday problem. In this birthday problem, there are 3,016 people and 13,983,816 possible birthdays. The 3,016 people correspond with the 3,016 drawings, while the binomial coefficient $\binom{49}{6} = 13{,}983{,}816$ gives the total number of possible combinations of six numbers drawn from the numbers 1, . . . , 49. The same reasoning used in the classic birthday problem leads to the conclusion that there is a probability of

$$\frac{13{,}983{,}816 \times (13{,}983{,}816 - 1) \times \cdots (13{,}983{,}816 - 3{,}015)}{(13{,}983{,}816)^{3016}} = 0.7224$$

that no combination of the six numbers will be drawn multiple times in 3,016 drawings of the German Lotto. In other words, there is a probability of 0.2776 that a same combination of six numbers will be drawn two or more times in 3,016 drawings. And this probability is not negligibly small!

3.1.4 The almost-birthday problem

In the almost-birthday problem, we undertake the task of determining the probability of two or more people in a randomly assembled group of n people having their birthdays within r days of each other. Denoting this probability by $p_n(r)$, it is given by

$$p_n(r) = 1 - \frac{(365 - 1 - nr)!}{365^{n-1}\big(365 - (r+1)n\big)!}.$$

The proof of this formula is rather tricky and can be found in J.I. Nauss, “An Extension of the Birthday Problem,” *The American Statistician* **22** (1968): 27–29. Although the almost-birthday problem is far more complicated than the ordinary birthday problem when it comes to theoretical analysis, this is not the case when it comes to computer simulation. Just a slight adjustment to the simulation program for the birthday problem makes it suitable for the almost-birthday problem. This is one of the advantages of simulation. For several values of n, Table 3.2 gives the value of the probability $p_n(1)$ that in a randomly

Table 3.2. *Probabilities for the almost-birthday problem* (r = *1).*

n	10	14	20	25	30	35	40
$p_n(1)$	0.3147	0.5375	0.8045	0.9263	0.9782	0.9950	0.9991

assembled group of n people at least two people will have birthdays within one day of each other ($r = 1$). A group of 14 people is large enough to end up with a probability of more than 50% that at least two people will have birthdays within one day of each other. Taking $r = 7$, one calculates that if seven students are renting a house together, there is a probability of more than 50% that at least two of them will have birthdays within one week of each other.

3.1.5 Coincidences

The birthday and almost-birthday problems handsomely illustrate the fact that concurrent circumstances are often less coincidental than we tend to think. It pays to be aware of a world full of apparently coincidental events that, on closer examination, are less improbable than intuition alone might lead one to suppose.[†]

The following example represents another case of coincidence turning out to be something less than coincidental. You answer the telephone and find yourself in conversation with a certain friend whose name had come up earlier that day in a conversation with others. How coincidental is this? A few rough calculations on a piece of scrap paper will show that, over a period of time, this is less coincidental than you might think. Making a rough calculation on scrap paper means simplifying without detracting from the essence of the problem. Let's begin by roughly estimating that over the years, you have discussed your friend with others one hundred or so times and that the probability of this friend telephoning you on any given day is equal to $p = \frac{1}{100}$. Instead of calculating the probability of your friend calling on a day when you have previously mentioned his name, let's calculate the complementary probability of your friend not telephoning you on any of the $n = 100$ days when he has been the subject of a conversation. This complementary probability is equal to $(1 - p)^n$. The probability, then, of your being telephoned at least one time by your friend on a day when you had previously mentioned him is given by $1 - (1 - p)^n$. For every value of $p > 0$ this probability comes arbitrarily close to 1 if n is large enough.

[†] See also P. Diaconis and F. Mosteller, "Methods for Studying Coincidences," *Journal of the American Statistical Association* **84** (1989): 853–861.

In particular, the probability $1-(1-p)^n$ has a value of 0.634 for $n=100$ and $p=\frac{1}{100}$. Over a period of time then, it is not particularly exceptional to have been telephoned by someone whom you had spoken of earlier that same day. This argumentation is applicable to many comparable situations; for example, a newspaper story reporting the collision of two Mercedes at a particular intersection, and that both drivers were called John Smith. This seems like an exceptional occurrence, but, if you think about it, there must be quite a few men called John Smith who drive Mercedes and pass one another every day at intersections. Of course, the newspaper never mentions these noncollisions. We only receive the filtered information about the collision, and it therefore appears to be exceptional.

Remarkable event in Monte Carlo

On August 18, 1913, a memorable event occurred in a Monte Carlo casino: the roulette wheel stopped no less than 26 times in a row on black. How exceptional can we consider a streak of this kind to be? In 1913, the Monte Carlo casino had been in operation for approximately 50 years. We can roughly estimate that over all of those 50 years, the roulette table had completed between three and five million runs. The probability of the wheel stopping on either red or black 26 times in a row in n rounds can be computed to have the value 0.022 for $n=3{,}000{,}000$ and the value 0.037 for $n=5{,}000{,}000$. Thus, it can be said to be exceptional that, in the first 50 years of the existence of the world's first casino, the roulette wheel stopped 26 times in a row on one and the same color. Today, well-trafficked casinos are to be found far and wide, and each is likely to have quite a number of roulette tables. On these grounds, one could hardly call it risky to predict that somewhere in the world during the coming 25 years, a roulette ball will stop on either red or black 26 or more times in a row. Any event with a nonzero probability will eventually occur when it is given enough opportunity to occur. This principle can be seen most clearly in the Lotto. Each participant has a probability almost equal to zero of winning the jackpot. Nevertheless, there is a large probability of the jackpot being won when the number of participants is sufficiently large.

3.2 The coupon collector's problem

In order to introduce a new kind of chips, the producer has introduced a campaign offering a "flippo" in each bag of chips purchased. There are ten different flippos. How many bags of chips do you expect to buy in order to get all ten flippos? In probability theory, this problem is known as the coupon collector's problem. The problem comes in many variations.

3.2.1 Simulation approach

In the Monte Carlo simulation, each simulation experiment consists of generating random drawings from the numbers 1, . . . , 10 until each of the ten numbers has been drawn at least one time. The number of drawings necessary is seen as the result of this experiment. After a sufficiently large number of experiments, the expected value we are looking for can be estimated by

$$\frac{\text{the sum of the outcomes of the experiments}}{\text{the total number of experiments}}.$$

The Monte Carlo study has to be redone when the number of flippos involved changes. This is not the case for the theoretical approach. This approach gives a better qualitative insight than the simulation approach.

3.2.2 Theoretical approach

Let's assume that there are n different flippos. Define the random variable X as the number of bags of chips that must be purchased in order to get a complete set of flippos. The random variable X can, in principle, take on any of the values $1, 2, \ldots$ and has thus a discrete distribution with infinitely many possible values. The expected value of X is defined by

$$E(X) = 1 \times P(X = 1) + 2 \times P(X = 2) + 3 \times P(X = 3) + \cdots .$$

A straightforward calculation of $E(X)$ is far from simple. Nevertheless, $E(X)$ is fairly easy to find indirectly by defining the random variable Y_i as

$$Y_i = \text{the number of bags of chips needed in order to go from } i-1 \text{ to } i \text{ different flippos.}$$

Now we can write X as

$$X = Y_1 + Y_2 + \cdots + Y_n.$$

The trick of representing a random variable by a sum of simpler random variables is a very useful one in probability theory. The expected value of the original random variable follows by taking the sum of the expected values of the simpler random variables. In Chapter 9, it will be shown that the expected value of a finite sum of random variables is always equal to the sum of the expected values.

In order to calculate $E\,(Y_i)$, the so-called geometric probability model is used. Consider an experiment having two possible outcomes. Call these outcomes "success" and "failure" and notate the probability of a "success" as p. In the geometric probability model, independent trials of an experiment are done until the first "success" occurs. Since the outcomes of the trials are independent of each other, it is reasonable to assign the probability $(1-p)^{k-1}p$ to the event of the first $k-1$ trials of the experiment delivering no success, and the kth delivering a success. It is obvious that the geometric probability model is applicable in the case of the Y_i variables. Let p_i represent the probability that the next bag of chips purchased will contain a new flippo when as many as $i-1$ differing flippos have already been collected. The probability p_i is equal to $\frac{n-(i-1)}{n}$ and the distribution of Y_i is given by

$$P\,(Y_i = k) = (1-p_i)^{k-1}\,p_i \qquad \text{for } k = 1, 2, \ldots .$$

For each $i = 1, \ldots, n$, the expected value of Y_i is given by

$$\begin{aligned} E\,(Y_i) &= p_i + 2\,(1-p_i)\,p_i + 3\,(1-p_i)^2\,p_i + \cdots \\ &= p_i[1 + 2\,(1-p_i) + 3\,(1-p_i)^2 + \cdots] \\ &= \frac{p_i}{[1-(1-p_i)]^2} = \frac{n}{n-i+1}, \end{aligned}$$

using the fact that the infinite series $1 + 2a + 3a^2 + \cdots$ has the sum $\frac{1}{(1-a)^2}$ for each a with $0 < a < 1$ (see the Appendix). The sought-after value of $E(X)$ now follows from

$$E(X) = E\,(Y_1) + E\,(Y_2) + \cdots + E\,(Y_n)\,.$$

Filling in the expression for $E(Y_i)$ leads to

$$E(X) = n\left[\frac{1}{n} + \frac{1}{n-1} + \cdots + 1\right].$$

For $n = 10$, we find then that the expected number of bags of chips needed in order to get a complete set of flippos is equal to 29.3.

The formula given above for $E(X)$ can be rewritten in a form that gives more insight into the way that $E(X)$ increases as a function of n. A well-known mathematical approximation formula is

$$1 + \frac{1}{2} + \cdots + \frac{1}{n} \approx \ln(n) + \gamma + \frac{1}{2n},$$

where $\gamma = 0.57722\ldots$ is the Euler constant. This leads to the insightful approximation

$$E(X) \approx n\ln(n) + \gamma n + \tfrac{1}{2}.$$

The coupon's collector problem appears in many forms. For example, how many rolls of a fair die are needed on average before each of the point surfaces has turned up at least one time? This problem is identical to the flippo problem with $n = 6$ flippos. Taking $n = 365$ flippos, the flippo problem also gives us the expected value of the number of people needed before we can assemble a random group of people in which all of the possible 365 birthdays are represented.

3.3 Craps

The wildly popular game of craps, first played in the United States in the twentieth century, is based on the old English game of Hazard. Craps is an extremely simple game in its most basic form; however, casinos have added on twists and turns enough to make most players' heads spin. The basic rules are as follows. A player rolls a pair of dice and the sum of the points is tallied. The player has won if the sum of the points is equal to seven or eleven, and has lost if the sum is equal to two, three, or twelve. In the case of all other point combinations, the player continues to roll until the sum of the first roll is repeated, in which case the player wins, or until rolling a total of seven, in which case the player loses. What is the probability of the player winning in craps?

3.3.1 Simulation approach

In a simulated craps experiment the rolls of a pair of dice are perpetuated until the game is ended. We simulate a roll of the dice by drawing a random number twice out of the numbers $1, \ldots, 6$, and adding up the sum of the two numbers. A key variable in the simulation is the total obtained in the first roll. Let's call this number the chance point. The experiment ends immediately if the chance point turns out to be seven or eleven (a win), or if it turns out to be a two, three or twelve (a loss). If none of these totals occurs, the simulation continues to "roll" until the chance point turns up again (a win), or until a total of seven appears (a loss). The probability of the player winning is estimated by dividing the number of simulated experiments leading to wins by the total number of experiments.

3.3.2 Theoretical approach

A simulation approach first requires looking at the number of points received in the first roll of the game. Depending on that, the next step of the simulation program is determined. In the theoretical model, we work along the same lines. In this case, we make use of the concept of *conditional probability*. Conditional probabilities have a bearing on a situation in which partial information over the outcome of the experiment is available. Probabilities alter when the available information alters. The notation $P(A|B)$ refers to the conditional probability that event A will occur *given* that event B has occurred.[†] In most concrete situations, the meaning of conditional probability and how it is calculated are obvious.[‡]

The law of conditional probabilities is an extremely useful result of applied probability theory. Let A be an event that can only occur after one of the events $B_1, \ldots, B_n$ has occurred. It is essential that the events $B_1, \ldots, B_n$ are disjoint, that is, only one of the $B_1, \ldots, B_n$ events can occur at a time. Under these conditions, the *law of conditional probabilities* says that

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n)$$

or, in abbreviated form,

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i).$$

We find the (unconditional) probability $P(A)$, then, by averaging the conditional probabilities $P(A \mid B_i)$ over the probabilities $P(B_i)$ for $i = 1, \ldots, n$. It is insightful to represent schematically the law of conditional probabilities by the tree diagram shown in Figure 3.1. A mathematical proof of this law will be given in Chapter 8.

Usually conditional probabilities are easy to calculate when the disjoint events $B_1, \ldots, B_n$ are suitably chosen. In determining the choice of these events, it may helpful to think of what you would do when writing a simulation program. In the craps example, we choose B_i as the event in which the first roll of the dice delivers i points for $i = 2, \ldots, 12$. Denote by $P(win)$ the probability of

[†] The precise definition of $P(A \mid B)$ will be given in Chapter 8. It boils down to the formula $P(AB) = P(A \mid B)P(B)$, where $P(AB)$ represents the probability that both event A and event B will occur. In words, the probability of the occurrence of both event A and event B equals the probability of the occurrence of event A given that event B has occurred multiplied by the probability of the occurrence of event B.

[‡] An illustrative example is as follows. Someone first draws at random a number from the integers $1, \ldots, 10$ and next draws at random a number from the remaining nine integers. Denote by E_i the event that the ith number drawn is even for $i = 1, 2$. The conditional probability $P(E_2 \mid E_1)$ is nothing else than the probability of getting an even number when drawing at random a number from four even numbers and five odd numbers, and so $P(E_2 \mid E_1) = \frac{4}{9}$. This gives $P(E_1E_2) = P(E_2 \mid E_1)P(E_1) = \frac{4}{9} \times \frac{5}{10} = \frac{2}{9}$.

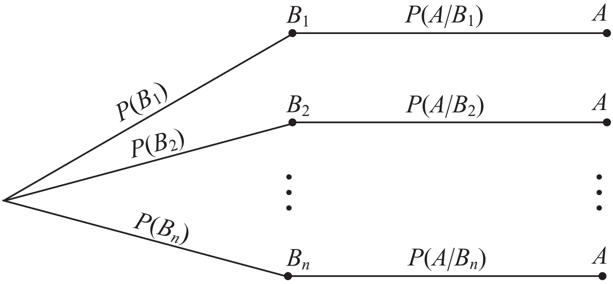


Fig. 3.1. Tree diagram for the law of conditional probabilities.

the player winning in craps and let $P(win \mid B_i)$ denote the revised value of this probability given the information of the occurrence of the event B_i. Then

$$P(win) = \sum_{i=2}^{12} P(win \mid B_i)P(B_i).$$

The conditional win probabilities are easy to calculate. Naturally,

$$P(win \mid B_i) = \begin{cases} 1 & \text{for } i = 7, 11, \\ 0 & \text{for } i = 2, 3, 12. \end{cases}$$

Prior to calculating $P(win \mid B_i)$ for the other values of i, we first determine the probabilities $P(B_i)$. The sample space for the experiment of rolling a pair of dice consists of the 36 outcomes (j, k), where $j, k = 1, 2, \ldots, 6$. The outcome (j, k) occurs if j points turn up on the first (red) die and k points turn up on the second (blue) die. The dice are fair, so the same probability $\frac{1}{36}$ is assigned to each of the 36 possible outcomes. The outcome (j, k) results in the value $i = j + k$ for the total of the points. Using the shorthand p_i for the probability $P(B_i)$, it is readily verified that

$$p_2 = \frac{1}{36},\ p_3 = \frac{2}{36},\ p_4 = \frac{3}{36},\ p_5 = \frac{4}{36},\ p_6 = \frac{5}{36},\ p_7 = \frac{6}{36},$$
$$p_8 = \frac{5}{36},\ p_9 = \frac{4}{36},\ p_{10} = \frac{3}{36},\ p_{11} = \frac{2}{36},\ p_{12} = \frac{1}{36}.$$

Then, we calculate the conditional probabilities $P(win \mid B_i)$. In order to do this, we first give the meaning of these probabilities in the concrete situation of the craps game. For example, the conditional probability $P(win \mid B_4)$ is no other than the unconditional probability that the total of 4 will appear before the total of 7 does in the (compound) experiment of repetitive dice rolling. The total of 4 will appear before the total of 7 only if one of the disjoint events $A_1, A_2, \ldots$ occurs, where A_k is the event that the first consecutive $k - 1$ rolls give neither the total of 4 nor the total of 7 and the kth consecutive roll gives a total of 4. Since the events $A_1, A_2, \ldots$ are mutually disjoint, $P(A_1 \cup A_2 \cup \cdots)$ is obtained by adding the probabilities $P(A_k)$ for $k = 1, 2, \ldots$. This gives

$$P(4 \text{ before } 7) = P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots .$$

The event A_k is generated by physically independent subexperiments and thus the probabilities of the individual outcomes in the subexperiments are multiplied by each other in order to obtain

$$P(A_k) = (1 - p_4 - p_7)^{k-1}\, p_4 \qquad \text{for } k = 1, 2, \ldots .$$

This leads to the formula

$$\begin{aligned} P(4 \text{ before } 7) &= p_4 + (1 - p_4 - p_7)p_4 + (1 - p_4 - p_7)^2 p_4 + \cdots \\ &= \frac{p_4}{p_4 + p_7}, \end{aligned}$$

using the fact that the geometric series $1 + a + a^2 + \cdots$ has a sum of $\frac{1}{1-a}$ for a with $0 < a < 1$ (see the Appendix). In this way, we find that

$$P(win \mid B_i) = \frac{p_i}{p_i + p_7} \qquad \text{for } i = 4, 5, 6, 8, 9, 10.$$

If we fill in the p_i values we get

$$\begin{aligned} &P(win \mid B_4) = \frac{3}{9}, \quad P(win \mid B_5) = \frac{4}{10}, \quad P(win \mid B_6) = \frac{5}{11}, \\ &P(win \mid B_8) = \frac{5}{11}, \quad P(win \mid B_9) = \frac{4}{10}, \quad P(win \mid B_{10}) = \frac{3}{9}. \end{aligned}$$

Putting it all together, we get

$$\begin{aligned} P(win) = {}& 0 \times \frac{1}{36} + 0 \times \frac{2}{36} + \frac{3}{9} \times \frac{3}{36} + \frac{4}{10} \times \frac{4}{36} + \frac{5}{11} \times \frac{5}{36} \\ & + 1 \times \frac{6}{36} + \frac{5}{11} \times \frac{5}{36} + \frac{4}{10} \times \frac{4}{36} + \frac{3}{9} \times \frac{3}{36} + 1 \times \frac{2}{36} \\ & + 0 \times \frac{1}{36} = 0.4929. \end{aligned}$$

In other words, the probability of the player losing is 0.5071. The casino payout is 1:1, so that you would lose on average $(0.5071 - 0.4929) \times 100 = 1.42$ cents per dollar staked. The fact that the house percentage is lower with the game of craps than with other casino games partly explains the popularity of the game. In the most basic version of craps, the players are passive during follow-up rolls of the dice, when the first roll has not been decisive. Players like action, and casinos like to keep players active. For this reason, casinos have added quite a few options onto the basic formula such that during the passive rounds, players can make seemingly attractive bets (which actually only raise the house advantage).

3.4 Gambling systems for roulette

The origins of probability theory lie in the gambling world. The best-known casino game is roulette. The oldest form of roulette is European roulette, which was developed in France around the year 1800. Players bet on the outcome of a turning wheel, which is outfitted with 37 spokes numbering from 0 to 36. Of the spokes numbered from 1 to 36, 18 are red and 18 are black. The 0, neither red nor black, represents a win for the casino. Players can bet on individual numbers, on combinations of numbers or on colors. The casino payout depends on the type of bet made. For example, a bet on the color red has a 1 to 1 payout. This means that if you stake one dollar on red and the wheel falls on a red number, you win back your dollar plus one more dollar. If the wheel does not stop on a red number, you lose your bet and forfeit the dollar you staked. For a bet on red, the probability of the player winning is $\frac{18}{37}$ and so the expected value of the casino payout is $2 \times \frac{18}{37} = 0.973$ dollars. In the long run, the casino keeps \$0.0270 of a one-dollar bet on red, or rather a house percentage of 2.70% for a bet on red. This house percentage of 2.70% remains constant for every type of bet in European roulette, as shown at the end of Section 2.6. The term *house percentage* (or *house advantage*) is much used by casinos and lotteries. The house percentage is defined as 100% times the casino's long-run average win per dollar staked.

3.4.1 Doubling strategy

A seemingly attractive strategy is known as the doubling strategy for a bet on red. This system works as follows. The player begins by staking one dollar on red. If he loses, he doubles his stake, and continues doubling until red wins. Theoretically, this system guarantees the player of an eventual one-dollar win. But, in practice, a player cannot continue to double unlimitedly. At a certain point he will either cross over the high stake limit or simply run out of money. Whatever the high stake limit is, over the long run a player loses 2.70 cents on every dollar staked. We will illustrate this by manner of a stake limit of \$1,000. Players reach this limit after losing ten times in a row. And by the tenth bet, players are staking an amount of $2^9 = 512$ dollars. A doubling round then consists of 11 bets at the most, of which the maximum stake of \$1,000 is made in the eleventh bet. We will assume that the starting capital is sufficiently large to play out the entire round.

3.4.2 Simulation approach

A simulation model for this problem is simply constructed. In each simulation experiment, a doubling round is replicated. This consists of taking random drawings from the numbers 0, 1, . . . , 37. The experiment ends as soon as a number corresponding to red is drawn, or when the eleventh drawing has been completed. The outcome of the experiment is 1 if you end a winner; otherwise, it is the negative of the total amount staked in the experiment. After a sufficiently large number of experiments, you can estimate the percentage of your win or loss per dollar staked by

$$\frac{\text{the sum of the outcomes in the experiments}}{\text{the sum of the total amounts staked in the experiments}} \times 100\%.$$

In running the simulation study, you will end up with a loss percentage estimated somewhere in the neighborhood of 2.7%.

3.4.3 Theoretical approach

Using a theoretical approach to calculate the average loss per doubling round, we must first determine the distribution of the random variable X that represents the number of bets in a single doubling round. The probability of the wheel stopping on red in the first bet is $\frac{18}{37}$, so we can say that $P(X = 1) = \frac{18}{37}$. The random variable X takes on the value k with $2 \leq k \leq 10$ if red does not result in the first $k - 1$ bets and then does result in the kth bet. The random variable X takes on the value 11 when red has not resulted in the first ten bets. This leads to

$$P(X = k) = \begin{cases} \left(\frac{19}{37}\right)^{k-1} \frac{18}{37} & \text{for } k = 1, \ldots, 10, \\ \left(\frac{19}{37}\right)^{10} & \text{for } k = 11. \end{cases}$$

Denote by a_k the total amount staked when the doubling round ends after k bets. Then

$$a_k = \begin{cases} 1 & \text{for } k = 1, \\ 1 + 2 + \ldots + 2^{k-1} & \text{for } k = 2, \ldots, 10, \\ 1 + 2 + \ldots + 2^9 + 1{,}000 & \text{for } k = 11. \end{cases}$$

If we fill in the values for $P(X = k)$ and a_k, we find that

$$E(\text{amount staked in a doubling round}) = \sum_{k=1}^{11} a_k P(X = k) = \$12.583.$$

In a doubling round, a player's win is equal to one dollar if the round lasts for fewer than 11 bets. The player's loss is \$23 if the round goes to 11 bets and the eleventh bet is won; the loss is \$2,023 if the round goes to 11 bets and the eleventh bet is lost. A doubling round goes to 11 bets with a probability of $\left(\frac{19}{37}\right)^{10}$, this being the probability of losing ten bets in a row. This gives

$$\begin{aligned}&E(\text{win in a doubling round})\\&\quad = 1 \times \left[1 - \left(\frac{19}{37}\right)^{10}\right] - 23 \times \left(\frac{19}{37}\right)^{10} \times \frac{18}{37} - 2{,}023 \times \left(\frac{19}{37}\right)^{10} \times \frac{19}{37}\\&\quad = -0.3401 \text{ dollars.}\end{aligned}$$

The stake amounts and losses (wins) will vary from round to round. The law of large numbers guarantees nevertheless that, over the long run, the fraction of your loss per dollar staked will come arbitrarily close to

$$\frac{E(\text{loss in a doubling round})}{E(\text{amount staked in a doubling round})} = \frac{0.3401}{12.583} = 0.027.$$

This is the same house advantage of 2.7% that we saw earlier! You simply cannot beat the casino over the long run using the doubling strategy. The doubling strategy does rearrange your losses, but over the long run you would get the self-same result if you simply gave away 2.7 cents of every dollar you planned to stake.

3.4.4 The Labouchère system

The Labouchère system is often used in the game of roulette. According to this system, you must decide beforehand how much you want to win, and you make a list of positive numbers whose sum add up to this amount. You bet on red each time. For each bet, you stake an amount equal to the sum of the first and last numbers on your list (if the list consists of just one number, then that number is the amount of the stake). If you win the amount staked, you cross off the amounts you used from your list. If you lose, you add the amount lost to the bottom of your list. You continue in this manner until your list is used up (your target amount has been achieved) or until you have lost all of your money. The Labouchère system is exciting, but it must also be understood that with repeated play, this system will deliver an unavoidable average loss of 2.7 cents per dollar staked. In general, this is not easily calculated, mathematically. For each specific situation, however, it is easy to make a simulation study. Let's say that your goal is to win \$250 and that you have a starting capital of \$2,500. Your list consists of the numbers 50, 25, 75, 50, 25, 25. The first time then, you will stake \$50 + \$25 = \$75. If you win, your list will be narrowed down to contain the numbers 25, 75, 50, 25, whereas if you lose, the list will be extended to contain the numbers 50, 25, 75, 50, 25, 25, 75. It is worth stating that, at a given moment, the sum of the first and last numbers could be larger than the amount of money you have at that moment. For example, let's say the first and last numbers are 50 and 125, and you only have \$150. Naturally, then, you would stake the \$150. If you lose, your store of money is depleted and the game is over. If you win, you cross off the last amount on your list and bring the first amount back to 25. In every simulation experiment, you begin with a capital of \$2,500 and play the Labouchère system until you have either added \$250 to your starting capital or until you have lost all of your money. The outcome of one round is 250 for a win and −2,500 for a loss. The total amount staked in a round consists of the individual stakes from the starting capital of \$2,500 together with the amounts that are won and subsequently staked anew. If you run a sufficiently large number of experiments and divide the sum of the outcomes by the sum of the amounts staked in the experiments, you will arrive at an estimate for the average loss per dollar staked when the Labouchère system is played repeatedly. In simulation runs of 10,000 and 100,000 experiments, we found the estimates to be 0.0294 and 0.0276, respectively, for the average loss per dollar staked. Indeed, the estimates are very close to the theoretical value of 0.027. For the probability of winning a given round, we found in these two simulation runs the simulated values 0.890 and 0.886. Despite the high probability of success for each round, the gambler who uses the Labouchère system repeatedly will lose in the long run. The expected value of your net gain in each round is negative.

The conclusion is that the Labouchère system will not be of help to you in assuring a win. A nonmathematical but nonetheless convincing proof of the fact that a winning betting system does not exist for the game of roulette is evident from the fact that casinos have never shown any resistance to the use of any such system at the roulette table. In fact, the only sure way to get rich through roulette is to open a casino!

3.5 The 1970 draft lottery

In 1970, during the Vietnam War, the American army used a lottery system based on birth dates to determine who would be called up for service in the military forces. The lottery worked like this: each of the 366 days of the year (including February 29) was printed on a slip of paper. These slips of paper were placed into individual capsules. The capsules were then placed into a large receptacle, which was rotated in order to mix them. Then, the capsules were drawn one by one out of the receptacle. The first date drawn was assigned a draft number of “one,” the second date drawn was assigned a draft number of “two,” and so on, until each day of the year had been drawn out of the receptacle and assigned a draft number. Draftees were called up for service based on the draft number assigned to their dates of birth, with those receiving low draft numbers being called up first. Table 3.3 gives the numbers assigned to the days of the various months. Directly after the lottery drawing, doubts were raised as to its fairness. In Chapter 2, we discussed the errors made in the randomization procedure used in this lottery. But, for the sake of argument, let’s say we are unaware of these errors. Now, based on the results shown in Table 3.3, we must decide whether the lottery can reasonably be said to have been random. How can we do this? We can use a Monte Carlo simulation to test whether the order of the lottery numbers in Table 3.3 can be described as random. First, we aggregate the data in a suitable and insightful way. Table 3.4 provides, for each month, the average value of the numbers representing the days of that month that were chosen. The monthly averages should fluctuate around 183.5 (why?). One glance at Table 3.4 will be enough to raise serious doubts about the fairness of the draft lottery. After May, the monthly averages show an obvious decline. What we now must determine is whether the deviations in Table 3.4 can more reasonably be described as an example of how fate can be fickle or as hard evidence of an unfair lottery. In order to make this determination, let’s start out with the hypothesis that the lottery was fair. If we can show that the outcomes in Table 3.4 are extremely improbable under the hypothesis, we can reject our hypothesis and conclude that the lottery was most probably unfair. Many test criteria are possible. One generally applicable test criterion is to consider the sum of the absolute deviations of the outcomes from their expected values. The expected value of the average draft number for a given month is 183.5 for each month. For convenience of notation, denote by $g_1 = 201.2, \ldots, g_{12} = 121.5$ the observed values for the average draft numbers for the months $1, \ldots, 12$ (see Table 3.4). The sum of the absolute deviations of the outcomes $g_1, \ldots, g_{12}$ from their expected values is

$$\sum_{i=1}^{12} |g_i - 183.5| = 272.4.$$

Is this large? We can answer this by means of a simple model. Determine a random permutation $(n_1, \ldots, n_{366})$ of the days $1, \ldots, 366$. Assign lottery number n_1 to January 1, number n_2 to January 2, etc., ending with lottery number n_{366} for December 31. For this assignment, define the random variable

Table 3.3. *Draft numbers assigned by lottery.*

day	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sep.	Oct.	Nov.	Dec.
1	305	086	108	032	330	249	093	111	225	359	019	129
2	159	144	029	271	298	228	350	045	161	125	034	328
3	251	297	267	083	040	301	115	261	049	244	348	157
4	215	210	275	081	276	020	279	145	232	202	266	165
5	101	214	293	269	364	028	188	054	082	024	310	056
6	224	347	139	253	155	110	327	114	006	087	076	010
7	306	091	122	147	035	085	050	168	008	234	051	012
8	199	181	213	312	321	366	013	048	184	283	097	105
9	194	338	317	219	197	335	277	106	263	342	080	043
10	325	216	323	218	065	206	284	021	071	220	282	041
11	329	150	136	014	037	134	248	324	158	237	046	039
12	221	068	300	346	133	272	015	142	242	072	066	314
13	318	152	259	124	295	069	042	307	175	138	126	163
14	238	004	354	231	178	356	331	198	001	294	127	026
15	017	089	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	055	274	120	044	207	254	107	096
17	235	189	033	260	112	073	098	154	255	288	143	304
18	140	292	332	090	278	341	190	141	246	005	146	128
19	058	025	200	336	075	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	063	192	185	135
21	186	363	334	062	250	060	027	291	204	243	156	070
22	337	290	265	316	326	247	153	339	160	117	009	053
23	118	057	256	252	319	109	172	116	119	201	182	162
24	059	236	258	002	031	358	023	036	195	196	230	095
25	052	179	343	351	361	137	067	286	149	176	132	084
26	092	365	170	340	357	022	303	245	018	007	309	173
27	355	205	268	074	296	064	289	352	233	264	047	078
28	077	299	223	262	308	222	088	167	257	094	281	123
29	349	285	362	191	226	353	270	061	151	229	099	016
30	164		217	208	103	209	287	333	315	038	174	003
31	211		030		313		193	011		079		100

Table 3.4. *Average draft number per month.*

January	201.2	July	181.5
February	203.0	August	173.5
March	225.8	September	157.3
April	203.7	October	182.5
May	208.0	November	148.7
June	195.7	December	121.5

Table 3.5. *Index numbers for the 1970 draft lottery.*

month	1	2	3	4	5	6	7	8	9	10	11	12
index	5	4	1	3	2	6	8	9	10	7	11	12

G_i as the average value of the lottery numbers assigned to the days of month i for $i = 1, \ldots, 12$. In order to answer the above question, we need

$$P\left(\sum_{i=1}^{12} |G_i - 183.5| \geq 272.4\right).$$

Deriving a versatile mathematical formula for this probability seems like an endless task. The value for this probability, however, is easily determined with the help of a Monte Carlo simulation. You conduct a large number of independent simulation trials, and in each trial a random permutation of the whole numbers $1, \ldots, 366$ is determined in order to assign lottery numbers to the days of the various months. A procedure for the determination of a random permutation is given in Section 2.9. A simulation trial is considered a "success" when the resulting monthly averages G_i measure up to $\sum_{i=1}^{12} |G_i - 183.5| \geq 272.4$. If you divide the number of successes by the total number of trials, you will come out with an estimate for the probability you are seeking. In a Monte Carlo study with 100,000 simulation runs, we came out with a simulated value of 0.012 for the probability in question.

Still another, yet stronger, indication that the lottery was not fair can be found in a test criterion that bears in mind the established trend of the monthly averages in Table 3.4. You would assign the index number 1 to the month with the highest monthly average, index number 2 to the month with the second highest monthly average, etc. For the 1970 draft lottery, these index numbers are shown in Table 3.5. They result in the permutation $(5, 4, \ldots, 12)$ of the numbers $1, 2, \ldots, 12$. Under the hypothesis that the lottery is fair, this permutation would have to be a "random" permutation. How can we test this? First, for a random permutation $\sigma = (\sigma_1, \ldots, \sigma_{12})$ of the numbers $1, \ldots, 12$, we define the distance measure $d(\sigma)$ by $d(\sigma) = \sum_{i=1}^{12} |\sigma_i - i|$. You can immediately verify that for each permutation σ, it holds that $0 \leq d(\sigma) \leq 72$. For the permutation $\sigma^* = (5, 4, \ldots, 12)$ from Table 3.5 it holds that $d(\sigma^*) = 18$. In order to judge whether the value 18 is "small" you must know, for a randomly chosen permutation σ, how likely the distance measure $d(\sigma)$ is less than or equal to 18. Again, you can apply a Monte Carlo simulation in order to find the value for this probability. You generate a large number of random permutations of the numbers $1, \ldots, 12$ and determine the proportion of permutations in which the distance measure $d(\sigma)$ is less than or equal to 18. A Monte Carlo study with 100,000 generated random permutations led us to an estimate of 0.0009 for our sought-after probability. This is strong evidence that the 1970 draft lottery did not proceed fairly.

3.6 Bootstrap method

In the statistical analysis of the 1970 draft lottery from Section 3.5, we used a powerful, generally applicable form of statistical methodology, namely the *bootstrap method.* This new method, developed in 1977 by American statistician Bradley Efron, has modern computer technology to thank for its efficacious calculating power. Conventional statistical methods were, for the most part, developed before we had computers at our disposal. The standard methods, therefore, necessarily relied on simplifying assumptions and relatively simple statistical measures that could be calculated from mathematical formulas. In contrast to these methods, the bootstrap method is letting the data speak for themselves by making use of the number-breaking power of modern-day computers, through the use of which calculation-intensive simulations can be made in virtually no time. A typical application can be described by the following situation: in order to test a new skin infection remedy, 20 healthy volunteers are infected with the corresponding ailment. They are then split up into two groups of equal size: a remedy group and a placebo group. The study being a double-blind study, the volunteers are not aware of which group they are assigned to, and the doctors do not have this information either. Each volunteer undergoes daily examinations until the malady is cured. In the remedy group the values for the number of days required until all patients are cured are 7, 9, 9, 11, 12, 14, 15, 15, 15, and 17. In the placebo group, the values for the number of days until all patients are cured are 9, 11, 11, 11, 12, 15, 17, 18, 18, and 20. In order to test whether the remedy helps or not, we take the difference in the total number of days until cured between the placebo group and the remedy group as test statistic T. For the sample data, the one-sided test statistic T takes on the value $142 - 124 = 18$. In order to make a statistical statement of whether the remedy works or not, we assume that it does not matter whether or not the remedy is used. Under this so-called null hypothesis, our 20 case studies can be seen as 20 independent drawings from the distribution of the time elapsed until cure is effected. Data from the experiment can be used for the empirical distribution of the time until cure. One of the 20 case studies reports a time value of 7, three of the 20 cases report a value of 9, and so on. Thus, for the time required until cure is effected, the respective probabilities $1/20$, $3/20$, $4/20$, $2/20$, $1/20$, $4/20$, $2/20$, $2/20$, and $1/20$ are assigned to the possible values

7, 9, 11, 12, 14, 15, 17, 18, and 20. According to the bootstrap method, you would now instruct your computer to make a large number of drawings, say 10,000, from this distribution (the array method from Section 2.9 can be used for this purpose). For each of the 10,000 simulation runs you determine the difference between the sum of the last ten values drawn and the sum of the first ten values drawn. The proportion of the number of simulation runs in which this difference is greater than or equal to 18 gives the bootstrap estimate for the probability that the test statistic T will take on a value greater than or equal to 18 under the null hypothesis. If this probability is smaller than a previously chosen threshold value, say 0.01, then the null hypothesis is discarded. Using the original data and performing 10,000 simulation runs, we found a value of 0.135 for the probability $P(T \geq 18)$. This probability is not small enough to reject the null hypothesis. The conclusion seems to be that the experiment must be redone using larger groups of people before any definitive conclusion can be reached about the remedy's effectiveness.

Another example of the bootstrap method is the prediction of election results based on probability statements made by polled voters. Consider the polling method in which respondents are asked to indicate not which candidate is their favorite, but rather what the various probabilities might be of their voting for each of the candidates in the running. Let's assume that a representative group of 1,000 voters is polled in this way. We then have 1,000 probability distributions over the various political candidates. Next, the computer allows us to draw from these 1,000 probability distributions a large number of times. In this way, we can simulate the probability that a given candidate will receive the most votes or the probability that, in the parliamentary system, a given two parties will receive more than half of the number of votes cast. In Section 12.3 of Chapter 12, we come back to this application.

3.6.1 A statistical test problem

The bootstrap method can also be used to solve Question 9 from Chapter 1. The question was whether someone can credibly claim to have rolled a one 196 times, a two 202 times, a three 199 times, a four 198 times, a five 202 times, and a six 203 times in 1,200 rolls of one fair die. In order to test whether something is credible or not, you must choose a suitable test statistic. A claim of having tossed 100 heads in a row in 100 tosses of a fair coin cannot be said to be incredible based simply on the grounds that the sequence $HH \ldots H$, consisting of 100 heads, has an inconceivably small probability of $\left(\frac{1}{2}\right)^{100}$ of occurring. Actually, each *specific* sequence of heads and tails of length 100 has a probability of $\left(\frac{1}{2}\right)^{100}$. No, the claim is incredible on the basis of the test statistic, which counts the total number of heads in 100 tosses of a fair coin. A value of 100 for this test statistic is improbably far away from the expected value of 50. Similarly, in the situation described in Question 9 of Chapter 1, you may observe that each of the outcomes is suspiciously close to its expected value of 200. The sum of the absolute deviations of the outcomes from their expected values of 200 is $4 + 2 + 1 + 2 + 2 + 3 = 14$. This sum is a natural touchstone for the question of whether the outcomes are invented or not. For the case that a single die actually is rolled 1,200 times, define the random variable X as

$$X = \sum_{i=1}^{6} |N_i - 200|,$$

where N_i represents the number of times that i points are rolled. The distribution of this test statistic cannot be calculated by mathematical equations (a related test statistic whose distribution can be approximated by a mathematically tractable distribution will be discussed in Section 12.4). However, the distribution of X can easily be found using Monte Carlo simulation. We can find an estimate for the probability $P(X \leq 14)$ by simulating 1,200 rolls of one die many times (you simulate the outcome of a given roll by drawing a random integer from $1, \ldots, 6$). If we divide the number of times that the simulated value of the random variable X is less than or equal to 14 by the total number of simulation runs, we arrive at an estimate for $P(X \leq 14)$. A simulation study with 100,000 runs leads us to an estimate of 0.0020 for our sought-after probability. This small probability means that the reported outcomes of 1,200 rolls of the die are difficult to explain as a chance variation. In other words, this is a strong indication that the outcomes claimed above are invented. Statistics can never definitely prove that data are fabricated. In statistics, there are no absolute certainties such as "water boils at a temperature of 100 degrees Celsius," but statistics does provide answers such as "there is clear evidence against the null hypothesis."

3.7 Problems

3.1 Is it credible if a local newspaper somewhere in the world reports on a given day that a member of the local bridge club was dealt a hand containing a full suit of 13 clubs?

3.2 Is the probability of a randomly chosen person having his/her birthday fall on a Monday equal to the probability of two randomly chosen people having their birthdays fall on the same day of the week?

3.3 In both the Massachusetts Numbers Game and the New Hampshire Lottery, a four-digit number is drawn each evening from the sequence 0000, 0001, . . . , 9999. On

Tuesday evening, September 9, 1981, the number 8092 was drawn in both lottery games. Lottery officials declared that the probability of both lotteries drawing the same number on that particular Tuesday evening was inconceivably small and was equal to one in one hundred million. Do you agree with this?

3.4 The national lottery is promoting a special, introductory offer for the upcoming summer season. Advertisements claim that, during the four scheduled summer drawings, it will hardly be possible not to win a prize, because four of every ten tickets will win at each drawing. What do you think of this claim?

3.5 What is the probability of a randomly chosen five-digit number lining up in the same order from right to left as it does from left to right?

3.6 The Yankees and the Mets are playing a best-four-of-seven series. The winner takes all of the prize money of one million dollars. Unexpectedly, the competition must be suspended when the Yankees lead two games to one. How should the prize money be divided between the two teams if the remaining games cannot be played? Assume that the Yankees and the Mets are evenly matched and that the outcomes of the games are independent of each other. (This problem is a variant of the famous "problem of points" that, in 1654, initiated the correspondence between the great French mathematicians Pascal and Fermat).

3.7 Five friends go out to a pub together. They agree to let a roll of the dice determine who pays for each round. Each friend rolls one die, and the one getting the lowest number of points picks up the tab for that round. If the low number is rolled by more than one friend in any given round, then the tab will be divided among them. At a certain point in the evening, one of the friends decides to go home; however, rather than withdraw from the game he proposes to participate in absentia, and he is assigned a point value of $2\frac{1}{2}$. Afterward, he will be responsible for paying up on the rounds he lost, calculating in an amount for the rounds he won. Is this a fair deal?

3.8 Suppose that a large group of people are undergoing a blood test for a particular illness. The probability that a random person has the illness in question is equal to 0.001. In order to save on the work, it is decided to split the group into smaller groups each consisting of r people. The blood samples of the r people are then mixed and tested all at once. If the test results are favorable, then one test will have been sufficient for that whole group. Otherwise, r extra tests will be necessary in order to test each of the r people individually. What is the expected value of the number of tests that will have to be done for a group of r people? Verify that $r = 32$ is the optimal group size.

3.9 You bet your friend that, of the next 15 automobiles to appear, at least two will have license plates beginning and ending with the same number. What is your probability of winning?

3.10 What is the probability that the same number will come up at least twice in the next ten spins of a roulette wheel?

3.11 A group of seven people in a hotel lobby are waiting for the elevator to take them up to their rooms. The hotel has 25 floors, each floor containing the same number of rooms. Suppose that the rooms of the seven waiting people are randomly distributed around the hotel.

(a) What is the probability of at least two people having rooms on the same floor?

(b) Suppose that you, yourself, are one of the seven people. What is the probability of at least one of the other six people having a room on the same floor as you?

3.12 The birthday problem and those cited in Problems 3.9–3.11 can be described as a special case of the following model. Randomly, you drop n balls in c compartments such that each ball is dropped independently of the others. It is assumed that $c > n$. What is the probability p_n that at least two balls will drop into the same compartment?

(a) Verify that the probability p_n is given by

$$p_n = 1 - \frac{c \times (c-1) \times \cdots \times (c-n+1)}{c^n}.$$

(b) Prove the approximation formula

$$p_n \approx 1 - e^{-\frac{1}{2}n(n-1)/c}$$

for c sufficiently large in comparison with n (use the fact that $e^{-x} \approx 1 - x$ for x close to 0).

(c) Verify that with a fixed c the value n must be chosen as

$$n \approx 1.18\sqrt{c}$$

in order to get a "50:50" chance of at least two balls dropping into the same compartment.

3.13 Suppose that someone has played bridge 30 times a week on average over a period of 50 years. Apply the result from Problem 3.12(b) to calculate the probability that this person has played exactly the same hand at least twice during the span of the 50 years.

3.14 In the Massachusetts Numbers Game, a four-digit number is drawn from the numbers $0000, 0001, \ldots, 9999$ every evening (except Sundays). Let's assume that the same lottery takes place in ten other states each evening.

(a) What is the probability that the same number will be drawn in two or more states next Tuesday evening?

(b) What is the probability that on some evening in the coming 300 drawings, the same number will be drawn in two or more states?

3.15 Of the unclaimed prize monies from the previous year, a lottery has purchased 500 automobiles to raffle off as bonus prizes among its 2.4 million subscribing members. Bonus winners are chosen by a computer programmed to choose 500 random numbers from among the 2.4 million registration numbers belonging to the subscribers. The computer is not programmed to avoid choosing the same number more than one time. What is the probability that someone will win two or more automobiles?

3.16 You received a tip that the management of a theater will give a free ticket to the first person in line having the same birthday as someone before him/her in line. Assuming that people enter the line one at a time and you do not know those people, what is the best position to take in the line if you can join it at any time?

3.17 A company has 110 employees in service. Use computer simulation to find the probability of there being 12 or more separate occasions when two or more employees have the same birthday. Also, determine the probability that, in each of the 12 months, two or more employees have the same birthday.

3.18 A commercial radio station is advertising a particular call-in game that will be played in conjunction with the introduction of a new product. The game is to be played every day for a period of 30 days. The game is only open to listeners between the ages of 15 and 30. Each caller will be the possible winner of one million dollars. The game runs as follows. At the beginning of each day the radio station randomly selects one date (day/month/year) from within a 15-year, span, that span consisting of the period from 15 to 30 years ago. Listeners whose birthday fall on the current day will be invited to call in to the station. At the end of the day, one listener will be chosen at random from among all of the listeners that called in that day. If that person's birth date matches the predetermined date picked by the radio station exactly, he/she will win one million dollars. What is the probability of someone winning the prize money during the 30-day run of the game?

3.19 In a television game show, the contestant can win a small prize, a medium prize, and a large prize. The large prize is a sports car. Each of the three prizes is "locked up" in a separate box. There are five keys randomly arranged in front of the contestant. One opens the lock to the small prize, another to the medium prize, another to the large prize. Another key is a dud that does not open any of the locks. The final key is the "master key" that opens all three locks. The contestant has a chance to choose up to two keys. For that purpose, the contestant is asked two quiz questions. For each correct answer, he/she can select one key. The probability of correctly answering any given quiz question is 0.5. The contestant tries the keys he/she has gained (if any) on all three doors. What is the probability that the contestant wins the sports car?

3.20 In a particular game, you begin by tossing a die. If the toss results in i points, then you go on to toss i dice together. If the sum of the points resulting from the toss of the i dice is greater than (less than) 12, you win (lose) one dollar, and if the sum of those points is equal to 12, you neither win nor lose anything. Use either simulation or a theoretical approach to determine the expected value of your net win in one round of this game.

3.21 In the popular English game of Hazard, a player must first determine which of the five numbers from $5, \ldots, 9$ will be the "main" point. The player does this by rolling two dice until such time as the point sum equals one of these five numbers. The player then rolls again. He/she wins if the point sum of this roll corresponds with the "main" point as follows: main 5 corresponds with a point sum of 5, main 6 corresponds with a point sum of 6 or 7, main 7 corresponds with sum 7 or 11, main 8 corresponds with sum 8 or 12, and main 9 corresponds with sum 9. The player loses if, having taken on a main point of 5 or 9, he/she then rolls a sum of 11 or 12, or by rolling a sum of 11 against a main of 6 or 8, or by rolling a sum of 12 against a main of 7. In every other situation the sum thrown becomes the player's "chance" point. From here on, the player rolls two dice until either the "chance" point (player wins) or the "main" point (player loses) reappears. Verify that the probability of the player winning is equal to 0.5228, where the main and the chance points contribute 0.1910 and 0.3318, respectively, to the probability of winning. What is the house percentage if the house pays the player $1\frac{1}{2}$ and 2 dollars per dollar staked for a main point win and a chance point win, respectively?

3.22 Go back and take another look at Problem 2.29 from Chapter 2. For ease of notation, let us rename the numbers 5, 10, ..., 100 on the wheel as 1, 2, ..., 20. For any $a = 1, 2, \ldots, 20$, let $S(a)$ denote the probability of candidate A winning if candidate A stops after the first spin giving a score of a points and let $C(a)$ denote the probability of candidate A winning if candidate A continues after the first spin giving a score of a points. Use conditional probabilities to find first an expression for $S(a)$ and next an expression for $C(a)$. Derive from these expressions the optimal stopping rule for candidate A and the maximal probability of candidate A winning. Repeat the calculations for the case where the numbers 1, 2, ..., 100 are on the wheel rather than the numbers 1, 2, ..., 20.

3.23 The game "Casino War" is played with a deck of cards compiled of six ordinary decks of 52 playing cards. Each of the cards is worth the face value shown (color is irrelevant). The player and the dealer each receive one card. If the player's card has a higher value than the dealer's, he wins double the amount he staked. If the dealer's card is of a higher value, then the player loses the amount staked. If the cards are of an equal value, then there is a clash and the player doubles his original bet. The dealer then deals one card to the player, one card to himself. If the value of the player's card is higher than the dealer's, he wins twice his original stake, otherwise he loses his original stake and the amount of the added raise. Using either simulation or a theoretical approach, determine the house percentage on this game.

3.24 A gang of thieves has gathered at their secret hideaway. Just outside, a beat-cop lurking about realizes that he has happened upon the notorious hideaway and takes it upon himself to arrest the gang leader. He knows that the villains, for reasons of security, will exit the premises one by one in a random order, and that as soon as he were to arrest one of them, the others would be alerted and would flee. For this reason, the agent plans only to make an arrest if he can be reasonably sure of arresting the top man himself. Fortunately, the cop knows that the gang leader is the tallest member of the gang, and he also knows that the gang consists of ten members. How can he maximize his probability of arresting the gang leader?

(a) Suppose a strategy whereby the cop always passes over the first $s - 1$ gang members that exit the hideaway, and then arrests the first gang member that is taller than the members who have previously exited the premises. Argue that this strategy will allow the cop to arrest the gang leader with a probability of

$$p(s, n) = \sum_{k=s}^{n} \left(\frac{s-1}{k-1} \times \frac{1}{n} \right),$$

where $n(= 10)$ is the number of gang members.

(b) For fixed n, analyze the difference function $p(s + 1, n) - p(s, n)$ and demonstrate that the probability $p(s, n)$ is maximal for the unique value of s, which satisfies

$$\frac{1}{s} + \frac{1}{s+1} + \cdots + \frac{1}{n-1} < 1 \leq \frac{1}{s-1} + \frac{1}{s} + \cdots + \frac{1}{n-1}.$$

Using the approximate expression for $\sum_{i=1}^{n} \frac{1}{i}$ stated in Section 3.2, verify that the optimal value of s satisfies $\ln(\frac{n}{s}) \approx 1$ when n is sufficiently large. Next, prove that the optimal value of s and the corresponding probability $p(s, n)$

are given by $s^* \approx \frac{n}{e}$ and $p(s^*, n) \approx \frac{1}{e}$ for n large, where $e = 2.7183\ldots$. The maximal probability of arresting the gang leader is, then, by good approximation, equal to 36.8% regardless of the magnitude of the gang (for $n = 10$ gang members the precise value of the probability is equal to 0.3987 and is achieved for $s^* = 4$).

3.25 Red Dog is a casino game played with a deck of 52 cards. Suit plays no role in determining the value of each card. An ace is worth 14, king 13, queen 12, jack 11, and numbered cards are worth the number indicated on the card. After staking a bet a player is dealt two cards. If these two cards have a "spread" of one or more, a third card is dealt. The spread is defined as the number of points between the values of the two cards dealt (e.g., if a player is dealt a 5 and a 9, he has a spread of three). When a player has a spread of at least one, he may choose to double his initial stake before the third card is dealt. At this point, the third card is dealt. If the value of the third card lies between the two cards dealt earlier, the player gets a payoff of s times his final stake plus the final stake itself, where $s = 5$ for a spread of 1, $s = 4$ for a spread of 2, $s = 2$ for a spread of 3, and $s = 1$ for a spread of 4 or more. In cases where the value of the two cards dealt is sequential (e.g., 7 and 8), no third card is dealt and the player gets his initial stake back. If the values of the two cards dealt are equal, the player immediately gets a third card. If this third card has the same value as the other two, the player gets a payoff of 11 times his initial stake plus the stake itself. The player applies the following simple strategy. The initial stake is only doubled if the spread equals 7 or more. Can you explain why it is not rational to double the stake if the spread is less than 7? Using computer simulation, determine the house percentage for Red Dog.

3.26 You are playing rounds of a certain game against an opponent until one of you has won all of the other one's betting money. At the start of each round, each of you stakes one dollar. The probability of winning any given round is equal to p, and the winner of a round gets the other player's dollar. Your starting capital is a dollars, and your opponent's starting capital is equal to b dollars. What is the probability of your winning all of the money? The renowned *gambler's formula* is

$$P(\text{you win all the money}) = \frac{1 - [(1-p)/p]^a}{1 - [(1-p)/p]^{a+b}},$$

with $p \neq \frac{1}{2}$ (otherwise your probability of winning is equal to $a/(a+b)$). In order to prove this formula, argue first the recursion relation

$$P_i = pP_{i+1} + (1-p)P_{i-1} \qquad \text{for } i = 1, \ldots, a+b-1,$$

in which P_k is defined as the probability of your eventually winning all of the money, when your capital is k dollars and your opponent's capital is $a + b - k$ dollars ($P_0 = 0$ and $P_{a+b} = 1$). Next, verify through substitution that the above formula is correct.

3.27 Suppose you go to the local casino with \$50 in your pocket, and it is your goal to multiply your capital to \$250. You are playing (European) roulette, and you stake a fixed amount on red for each spin of the wheel. What is the probability of your reaching your goal when you stake fixed amounts of \$5, \$10, \$25, and \$50, respectively, on each spin of the wheel? What do you think happens to the expected value of the number of bets it takes for you to either reach your goal or lose everything if the size of your stake increases? Can you intuitively explain why the probability of reaching your goal is higher for bold play than for cautious play?

3.28 A drunkard is wandering back and forth on a road. At each step he moves two units distance to the north with a probability of $\frac{1}{2}$, or one unit to the south with a probability $\frac{1}{2}$. Let a_k denote the probability of the drunkard ever returning to his point of origin if the drunkard is k units distance away in the northwards direction. Use the law of conditional probabilities to argue that $a_k = \frac{1}{2}a_{k+2} + \frac{1}{2}a_{k-1}$ for $k \geq 1$. Next, show that $a_k = q^k$ for all k, where $q = \frac{1}{2}(\sqrt{5} - 1)$. Could you give a probabilistic explanation of why a_k must be of the form q^k for some $0 < q < 1$? Use the result for the drunkard's walk to prove that the probability of the number of heads ever exceeding twice the number of tails is $\frac{1}{2}(\sqrt{5} - 1)$ if a fair coin is tossed over and over.

3.29 You have \$800 but you desperately needs \$1,000 before midnight. The casino must bring help. You decide for bold play at European roulette. You bet on red each time. The stake is \$200 if your bankroll is \$200 or \$800 and is \$400 if your bankroll is \$400 or \$600. You quit as soon as you have either reached your goal or lost everything. Use simulation to find the probability of reaching your goal. What is the expected value of your loss and what is the expected value of the total amount staked during your visit to the gambling table?

3.30 Twenty-five persons attended a "reverse raffle," in which everyone bought a number. Numbered balls were then drawn out of a bin one at a time at random. The last ball in the bin would be the winner. But when the organizers got down to the last ball, they discovered that three numbered balls had been unintentionally overlooked. They added those balls to the bin and continued the drawing. Was the raffle still fair? Use conditional probabilities to motivate your answer.

3.31 In the last 250 drawings of Lotto 6/45, the numbers $1, \ldots, 45$ were drawn

$$46, 31, 27, 32, 35, 44, 34, 33, 37, 42, 35, 26, 41, 38, 40,$$
$$38, 23, 27, 31, 37, 28, 25, 37, 33, 36, 32, 32, 36, 33, 36,$$
$$22, 31, 29, 28, 32, 40, 31, 30, 28, 31, 37, 40, 38, 34, 24$$

times, respectively. Using simulation, determine whether these results are suspicious, statistically speaking.

3.32 Jeu de Treize was a popular card game in seventeenth century France. This game was played as follows. One person is chosen as dealer and the others are players. Each player puts up a stake. The dealer takes a full deck of 52 cards and shuffles them thoroughly. Then the dealer turns over the cards one at a time, calling out "one" as he turns over the first card, "two" as he turns over the second, "three" as he turns over the third , and so on up to the thirteenth. A match occurs if the number the dealer is calling corresponds to the card he turns over, where "one" corresponds to an ace of any suit, "two" to a two of any suit, "three" to a three of any suit, ..., "13" to a king of any suit. If the dealer goes through a sequence of 13 cards without a match, the dealer pays the players an amount equal to their stakes, and the deal passes to the player sitting to his right. If there is a match, the dealer collects the player's stakes and the players put up new stakes for the next round. Then the dealer continues through the deck and begins over as before, calling out "one," and then "two," and so on. If the dealer runs out of cards, he reshuffles the full deck and continues the count where he left off. Use computer simulation to find the probability that the dealer wins k or more consecutive rounds for $k = 0, 1, \ldots, 8$. Also, verify by computer simulation that the expected number of rounds won by the dealer is equal to 1.803.

4
Rare events and lotteries

How does one calculate the probability of throwing heads more than 15 times in 25 tosses of a fair coin? What is the probability of winning a lottery prize? Is it exceptional for a city that averages eight serious fires per year to experience 12 serious fires in one particular year? These kinds of questions can be answered by the probability distributions that we will be looking at in this chapter. These are the binomial distribution, the Poisson distribution, and the hypergeometric distribution. A basic knowledge of these distributions is essential in the study of probability theory. This chapter gives insight into the different types of problems to which these probability distributions can be applied. The binomial model refers to a series of independent trials of an experiment that has *two* possible outcomes. Such an elementary experiment is also known as a *Bernoulli experiment*, after the famous Swiss mathematician Jakob Bernoulli (1654–1705). In most cases, the two possible outcomes of a Bernoulli experiment will be specified as "success" or "failure." Many probability problems boil down to determining the probability distribution of the total number of successes in a series of independent trials of a Bernoulli experiment. The Poisson distribution is another important distribution and is used, in particular, to model the occurrence of *rare* events. When you know the expected value of a Poisson distribution, you know enough to calculate all of the probabilities of that distribution. You will see that this characteristic of the Poisson distribution is exceptionally useful in practice. The hypergeometric distribution goes hand in hand with a model known as the "urn model." In this model, a number of red and white balls are selected out of an urn without any being replaced. The hypergeometric probability distribution enables you to calculate your chances of winning in lotteries.

SIXTH SET OF WINNING NUMBERS THAT BLOKE'S HAD IN AS MANY WEEKS...
LOTTERY PAYOUT
NOSTRADAMUS
£
JOHN BYRNE

4.1 The binomial distribution

The binomial probability distribution is the most important of all the discrete probability distributions. The following simple probability model underlies the binomial distribution: a certain chance experiment has two possible outcomes ("success" and "failure"), the outcome "success" having a given probability of p and the outcome "failure" a given probability of $1 - p$. An experiment of this type is called a Bernoulli experiment. Consider now the compound experiment that consists of n independent trials of the Bernoulli experiment. Define the random variable X by

$X =$ the total number of successes in n independent trials of the Bernoulli experiment.

The distribution of X is then calculated thus:

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

This discrete distribution is called the *binomial distribution* and is derived as follows. Let's say that a success will be recorded as a "one" and a failure as a "zero." The sample space of the compound experiment is made up of all the possible sequences of zeros and ones to a length of n. The n trials of the Bernoulli experiment are physically independent and thus the probability assigned to an element of the sample space is the product of the probabilities of the individual outcomes of the trials. A specific sequence with k ones and $n-k$ zeros gets assigned a probability of $p^k(1-p)^{n-k}$. The total number of ways by which k positions can be chosen for a one and $n-k$ positions can be chosen for a zero is $\binom{n}{k}$ (see the Appendix). Using the addition rule, the formula for $P(X = k)$ follows.

The expected value of the binomial variable X is given by

$$E(X) = np.$$

The proof is simple. Write $X = Y_1 + \cdots + Y_n$, where Y_i is equal to 1 if the ith trial is a success and 0 otherwise. Noting that $E(Y_i) = 0 \times (1-p) + 1 \times p = p$ and using the fact that the expected value of a sum of random variables is the sum of the expected values, the desired result follows.

The binomial probability model has many applications, in illustration of which we offer four examples.

Example 4.1 Daily Airlines flies from Amsterdam to London every day. The price of a ticket for this extremely popular flight route is \$75. The aircraft has a passenger capacity of 150. The airline management has made it a policy to sell 160 tickets for this flight in order to protect themselves against no-show passengers. Experience has shown that the probability of a passenger being a no-show is equal to 0.1. The booked passengers act independently of each other. Given this overbooking strategy, what is the probability that some passengers will have to be bumped from the flight?

Solution. This problem can be treated as 160 independent trials of a Bernoulli experiment with a success rate of $\frac{9}{10}$, where a passenger who shows up for the flight is counted as a success. Use the random variable X to denote number of passengers that show up for a given flight. The random variable X is binomially distributed with the parameters $n = 160$ and $p = \frac{9}{10}$. The probability in question is given by $P(X > 150)$. If you feed the parameter values $n = 160$ and $p = \frac{9}{10}$ into a software module for a binomial distribution, you get the numerical value $P(X > 150) = 0.0359$. Thus, the probability that some passengers will be bumped from any given flight is 3.6%.

Example 4.2 In a desperate attempt to breathe new life into the commercial television network "Gamble 7" and to acquire wider cable access, network management has decided to broadcast a lottery called "Choose your favorite spot." Here is how the lottery works: individual participants purchase lottery tickets that show a map of the Netherlands split into four regions, each region listing 25 cities. They choose and place a cross next to the name of one city in each of the four regions. In the weekly television broadcast of the lottery show, one city is randomly chosen for each region. If Gamble 7 has cable access in the cities whose names were drawn, it will make a donation to the cultural coffers of the local government of those cities. In order to determine the prize amount for individual participants in the lottery, Gamble 7 wants to know the probability of one participant correctly guessing the names of four, three, or two of the cities drawn. What are these probabilities?

Solution. What we have here is four trials of a Bernoulli experiment (four times a selection of a city), where the probability of success on each trial is $\frac{1}{25}$. This means that the binomial probability model, with $n = 4$ and $p = \frac{1}{25}$, is applicable. In other terms

$$P(\text{you have } k \text{ cities correct}) = \binom{4}{k}\left(\frac{1}{25}\right)^k\left(\frac{24}{25}\right)^{4-k}, \qquad k = 0, \ldots, 4.$$

This leads to the numerical values

$$P(\text{you have 4 cities correct}) = 2.56 \times 10^{-6}$$
$$P(\text{you have 3 cities correct}) = 2.46 \times 10^{-4}$$
$$P(\text{you have 2 cities correct}) = 8.85 \times 10^{-3}.$$

Example 4.3 Gordie the Gambler, a familiar figure in the cafés of central Amsterdam, offers café customers a game of chance called Chuck-a-Luck. To play this game, a customer chooses one number from the numbers $1, \ldots, 6$. A die is then rolled three times. If the customer's number does not come up at all in the three rolls, the customer pays Gordie 100 dollars. If the chosen number comes up one, two, or three times, Gordie pays the customer \$100, \$200, or \$300 respectively. How remunerative is this game for Gordie?

Solution. This game seems at first glance to be more favorable for the customer. Many people think that the chosen number will come up with a probability of $\frac{1}{2}$. This is actually not the case, even if the expected value of the number of times the chosen number comes up is equal to $\frac{1}{2}$. The number of times the customer's number comes up is seen as the number of successes in $n = 3$ independent trials of a Bernoulli experiment with a probability of success of $p = \frac{1}{6}$. This gives

$$P(\text{the chosen number comes up } k \text{ times}) = \binom{3}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{3-k}$$

for $k = 0, 1, 2, 3$. Hence the average win for Gordie per wager is given by

$$\begin{aligned}&100 \times \tfrac{125}{216} - 100 \times 3 \times \tfrac{25}{216} - 200 \times 3 \times \tfrac{5}{216} - 300 \times \tfrac{1}{216}\\ &\quad = 100 \times \tfrac{17}{216} = 7.87 \text{ dollars.}\end{aligned}$$

Not a bad profit return for a small businessman!

Example 4.4 Joe and his friend make a guess every week about the price trend of ten mutual funds. Independently of each other, they predict for each of the ten funds whether the price of the fund will be higher or not at the end of the week. They make their predictions by tossing a fair coin. Both put \$10 in the pot. Joe asks his friend if he could contribute \$20 to the pot and submit his guesses together with those of his brother. The friend agrees. For each fund, however, Joe's brother submits a prediction opposite to that of Joe. The person with the highest number of correct predictions wins the entire pot. If more than one person has the highest score, the winning persons split the pot evenly. How favorable is the game to Joe and his brother?

Solution. Let's denote by $p_{\mathrm{J}}(j)$ the probability that Joe will have j correct predictions and by $p_{\mathrm{F}}(f)$ the probability that his friend will have f correct predictions. Both the number of correct predictions of Joe and that of his friend have a binomial distribution with parameters $n = 10$ and $p = 0.5$. Hence

$$p_{\mathrm{J}}(i) = p_{\mathrm{F}}(i) = \binom{10}{i}\left(\frac{1}{2}\right)^i\left(\frac{1}{2}\right)^{10-i} \qquad \text{for } i = 0, 1, \ldots, 10.$$

For ease of notation, we denote by $P_{\mathrm{F}}(f) = \sum_{i=0}^{f} p_{\mathrm{F}}(i)$ the probability that Joe's friend will have no more than f correct predictions. Let the random variable W be defined as the winnings of Joe and his brother. The random variable W takes on the values 30, 20, 15, and 0. To calculate $P(W = 30)$, let A_j be the event that Joe has j correct predictions and wins together with his brother the entire pot. Then, $P(W = 30) = \sum_{j=0}^{10} P(A_j)$. Using the fact that Joe's brother has $10 - j$ correct predictions if Joe has j correct predictions, we have $P(A_j) = p_{\mathrm{J}}(j)P_{\mathrm{F}}(10 - j - 1)$ for $0 \leq j \leq 5$ and $P(A_j) = p_{\mathrm{J}}(j)P_{\mathrm{F}}(j - 1)$ for $6 \leq j \leq 10$. This gives

$$P(W = 30) = \sum_{j=0}^{5} p_{\mathrm{J}}(j)P_{\mathrm{F}}(10 - j - 1) + \sum_{j=6}^{10} p_{\mathrm{J}}(j)P_{\mathrm{F}}(j - 1) = 0.6468.$$

Table 4.1. *Expected profit for Joe and his brother.*

m	1	2	3	4	5	6	7	8	9	10
$E(m)$	2.50	1.88	2.50	2.15	2.50	2.26	2.50	2.31	2.50	2.35

The random variable W equals 20 if both Joe and his friend have five correct predictions (in which case Joe's brother also has five correct). Hence,

$$P(W = 20) = p_{\mathrm{J}}(5)p_{\mathrm{F}}(5) = 0.0606.$$

The random variable W equals 15 if Joe does not have five correct and the number of correct predictions of Joe's friend is the same as the largest of the number of correct predictions of Joe and that of his brother. Hence,

$$P(W = 15) = \sum_{j=0}^{4} p_{\mathrm{J}}(j)p_{\mathrm{F}}(10 - j) + \sum_{j=6}^{10} p_{\mathrm{J}}(j)p_{\mathrm{F}}(j) = 0.1156.$$

We are now in a position to calculate

$$E(W) = \$30 \times 0.6468 + \$20 \times 0.0606 + \$15 \times 0.1156 = \$22.35.$$

Joe and his brother have an expected profit of \$2.35. This expected profit applies to the situation of predicting the price trend of ten mutual funds. In the same way you can calculate the expected profit for Joe and his brother for the situation of predicting the price trend of m mutual funds. In Table 4.1 we give the expected profit $E(m)$ for several values of m. A surprising finding is that, for an odd number of funds, the expected profit is independent of the number of funds.

4.2 The Poisson distribution

In 1837, the famous French mathematician Siméon-Denis Poisson (1781–1840) published his *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile*. Indirectly, this work introduced a probability distribution that would later come to be known as the Poisson distribution, and this would develop into one of the most important distributions in probability theory. In this section, the Poisson distribution will be revealed in all its glory. The first issue at hand will be to show how this distribution is realized, namely as a limiting distribution of the binomial distribution. In case of a very large number of independent trials of a Bernoulli experiment with a very small probability of success, the binomial distribution gives way to the Poisson distribution. This insight is essential in order to apply the Poisson distribution in practical situations. In the course of this account, we will offer illustrative applications of the Poisson distribution. Finally, we will delve into the Poisson process. This random process is closely allied with the Poisson distribution and describes the occurrence of events at random points in time.

4.2.1 The origin of the Poisson distribution

A random variable X is *Poisson distributed* with parameter λ if

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad \text{for } k = 0, 1, \ldots,$$

where $e = 2.7182\ldots$ is the base of the natural logarithm. The Poisson distribution is characterized by just a single parameter λ, where λ is a positive real number. The expected value of the Poisson distribution is equal to this parameter λ. This follows from

$$\begin{aligned} E(X) &= 0 \times P(X = 0) + 1 \times P(X = 1) + 2 \times P(X = 2) + \cdots \\ &= \lambda e^{-\lambda} + 2\frac{\lambda^2}{2!}e^{-\lambda} + 3\frac{\lambda^3}{3!}e^{-\lambda} + \cdots \\ &= \lambda e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots\right) = \lambda e^{-\lambda}e^{\lambda} = \lambda, \end{aligned}$$

where we make use of the well-known power series $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots$ for every real number x (see the Appendix).

Many practical phenomena can be described according to the Poisson distribution. Evidence of this lies in the following important result:

> **in a *very large* number of independent repetitions of a Bernoulli experiment having a *very small* probability of success, the total number of successes is approximately Poisson distributed with the expected value $\lambda = np$, where $n =$ the number of trials and $p =$ the probability of success.**

To give a precise mathematical formulation of this result, let Z represent a binomially distributed random variable with the parameters n and p. In other words, Z represents the number of successes in n independent repetitions of a Bernoulli experiment with a success probability of p. Assume now that n becomes *very large* and p becomes *very small* so that np remains equal to the constant λ. The following is then true

$$\lim_{n\to\infty,\, p\to 0} P(Z = k) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad \text{for } k = 0, 1, \ldots.$$

Table 4.2. *Binomial probabilities and Poisson probabilities.*

k	$n = 25$	$n = 100$	$n = 500$	$n = 1{,}000$	Pois(1)
0	0.3604	0.3660	0.3675	0.3677	0.3679
1	0.3754	0.3697	0.3682	0.3681	0.3679
2	0.1877	0.1849	0.1841	0.1840	0.1839
3	0.0600	0.0610	0.0613	0.0613	0.0613
4	0.0137	0.0149	0.0153	0.0153	0.0153
5	0.0024	0.0029	0.0030	0.0030	0.0031

Proving this is not difficult. Since $p = \frac{\lambda}{n}$,

$$\begin{aligned} P(Z = k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{n!}{k!(n-k)!}\frac{\lambda^k}{n^k}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^k} \\ &= \frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^n\left[\frac{n!}{n^k(n-k)!}\right]\left(1 - \frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

Now let's look at the different terms separately. Assign a fixed value to k of $0 \leq k \leq n$. The term $\frac{n!}{n^k(n-k)!}$ is equal to

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} = \left(1 - \frac{1}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right).$$

With a *fixed* k, this term approaches 1 as $n \to \infty$, as does the term $(1 - \lambda/n)^{-k}$. The function e^x has the property that $(1 + b/n)^n$ tends to e^b as $n \to \infty$ for every real number b (see the Appendix). This results in $\lim_{n\to\infty}(1 - \lambda/n)^n = e^{-\lambda}$, which proves the result for the limit of $P(Z = k)$.

To give you an idea of how quickly the binomial distribution approaches the Poisson distribution, refer to Table 4.2, where the probabilities $P(X = k)$ are given for $k = 0, 1, \ldots, 5$ for a Poisson-distributed random variable X with expected value $\lambda = 1$ and for a binomially distributed random variable X with expected value $np = 1$, where n runs through the values 25, 100, 500, and 1,000.

The Poisson approximation is characterized by the pleasant fact that one does not need to know the precise number of trials and the precise value of the probability of success; it is enough to know what the product of these two values is. This product is the expected value of the total number of successes. The Poisson distribution is uniquely determined by its expected value. This fact is extremely useful for practical purposes.

The importance of the Poisson distribution cannot be emphasized enough. As is often remarked, the French mathematician Poisson did not recognize the huge practical importance of the distribution that would later be named after him. In his book, he dedicates just one page to this distribution. It was L. von Bortkiewicz in 1898, who first discerned and explained the importance of the Poisson distribution in his book *Das Gesetz der Kleinen Zahlen* (*The Law of Small Numbers*). One unforgettable example from this book applies the Poisson model to the number of Prussian cavalry deaths attributed to fatal horse kicks (in each of the years between 1875 and 1894). Here, indeed, one encounters a very large number of trials (the Prussian cavalrymen), each with a very small probability of "success" (fatal horse kick). The Poisson distribution is applicable to many other situations from daily life, such as the number of serious traffic accidents that occur yearly in a certain area, the weekly number of winners in a football pool, the number of serious earthquakes occurring in one year, the number of damage claims filed yearly with an insurance company, and the yearly number of mail carriers that are bitten, and so on.

4.2.2 Applications of the Poisson model

In this section, we will discuss a number of applications of the Poisson model. The examples are taken from everyday life.

Example 4.5 The Pegasus Insurance Company has introduced a policy that covers certain forms of personal injury with a standard payment of $100,000. The yearly premium for the policy is $25. On average, 100 claims per year lead to payment. There are more than one million policyholders. What is the probability that more than 15 million dollars will have to be paid out in the space of a year?

Solution. In fact, every policyholder conducts a personal experiment in probability after purchasing this policy, which can be considered to be "successful" if the policyholder files a rightful claim during the ensuing year. This example is characterized by an extremely large number of independent probability experiments each having an extremely small probability of success. This means that a Poisson distribution with an expected value of $\lambda = 100$ can be supposed for the random variable X, which is defined as the total number of claims that will be approved for payment during the year of coverage. The probability of having to pay out more than 15 million dollars within that year is equal to $P(X > 150)$. Entering $\lambda = 100$ into a software module for the Poisson distribution gives $P(X > 150) = 1.23 \times 10^{-6}$. Not a probability the insurance executives need worry about.

Example 4.6 During the last few years in Gotham City, a provincial city with more than 100,000 inhabitants, there have been eight serious fires per year, on average. Last year, by contrast, 12 serious fires blazed, leading to great consternation among the populace of the ordinarily tranquil city. The newspaper serving Greater Gotham, the *Gotham Echo*, went wild, carrying inflammatory headlines declaring "50% more fires" and demanding the resignation of the local fire chief. Is all this uproar warranted?

Solution. In a city as large as Gotham City, it is reasonable to assume that the number of fires occurring within one year has a Poisson distribution (why?). In order to determine whether 12 fires occurring in the past year is exceptional, one must know the probability of a Poisson-distributed random variable X with expected value $\lambda = 8$ taking on a value greater than 11. Entering $\lambda = 8$ in a software module for the Poisson distribution gives a result of

$$P(X > 11) = 0.112.$$

The question that follows is whether this probability of 11.2% is, in fact, so small that the occurrence of 12 or more fires must be qualified as exceptional. The answer to this question is subjective: some would say yes, some no. It is common practice, in statistics, to limit oneself to probabilities of less than 5% when speaking of exceptional outcomes. A statistician would, in this case, give the benefit of the doubt to the local fire brigade.

Example 4.7[†] The following item was reported in the February 14, 1986 edition of *The New York Times*: "A New Jersey woman wins the New Jersey State Lottery twice within a span of four months." She won the jackpot for the first time on October 23, 1985 in the Lotto 6/39. Then she won the jackpot in the new Lotto 6/42 on February 13, 1986. Lottery officials declare that the probability of winning the jackpot twice in one lifetime is approximately one in 17.1 trillion. What do you think of this statement?

Solution. The claim made in this statement is easily challenged. The officials' calculation proves correct only in the extremely farfetched case scenario of a given person entering a six-number sequence for Lotto 6/39 and a six-number sequence for Lotto 6/42 just one time in his/her life. In this case, the probability of getting all six numbers right, both times, is equal to

$$\frac{1}{\binom{39}{6}} \times \frac{1}{\binom{42}{6}} = \frac{1}{1.71 \times 10^{13}}.$$

But this result is far from miraculous when you begin with an extremely large number of people who have been playing the lottery for a long period of time,

[†] This example is based on the article "Jumping to coincidences: defying odds in the realm of the preposterous," by J.A. Hanley, in *American Statistician* **46** (1992): 197–202.

each of whom submit more than one entry for each weekly draw. For example, if every week 50 million people randomly submit five six-number sequences to one of the (many) Lottos 6/42, then the probability of one of them winning the jackpot twice in the coming four years is approximately equal to 63%. The calculation of this probability is based on the Poisson distribution, and goes as follows. The probability of your winning the jackpot in any given week by submitting five six-number sequences is

$$\frac{5}{\binom{42}{6}} = 9.531 \times 10^{-7}.$$

The number of times that a given player will win a jackpot in the next 200 drawings of a Lotto 6/42, then, is Poisson distributed with expected value

$$\lambda_0 = 200 \times \frac{5}{\binom{42}{6}} = 1.983 \times 10^{-4}.$$

For the next 200 drawings, this means that

$$\begin{aligned}&P(\text{any given player wins the jackpot two or more times})\\&\quad = 1 - e^{-\lambda_0} - e^{-\lambda_0}\lambda_0 = 1.965 \times 10^{-8}.\end{aligned}$$

Subsequently, we can conclude that the number of people under the 50 million mark, who win the jackpot two or more times in the coming four years, is Poisson distributed with expected value

$$\lambda = 50{,}000{,}000 \times (1.965 \times 10^{-8}) = 0.9825.$$

The probability in question, that at some point in the coming four years at least one of the 50 million players will win the jackpot two or more times, can be given as $1 - e^{-\lambda} = 0.626$. A few simplifying assumptions are used to make this calculation, such as the players choose their six-number sequences randomly. This does not influence the conclusion that it may be expected once in a while, within a relatively short period of time, that *someone* will win the jackpot two times.

4.2.3 Poisson model for weakly dependent trials

The Poisson distribution is derived for the situation of many independent trials each having a small probability of success. In case the independence assumption is not satisfied, but there is a "weak" dependence between the trial outcomes, the Poisson model may still be useful as an approximation. In surprisingly many probability problems, the Poisson approximation method enables us to obtain quick estimates for probabilities that are otherwise difficult to calculate. This approach requires that the problem is reformulated in the framework of a series of (weakly dependent) trials. The idea of the method is first illustrated by the birthday problem.

The birthday problem revisited

The birthday problem deals with the question of determining the probability of at least two people in a randomly formed group of m people having their birthdays on the same day. This probability can be approximated with the help of the Poisson model. To place the birthday problem in the context of a series of trials, some creativity is called for. The idea is to consider all of the possible combinations of two people and to trace whether in any of those combinations both people have birthdays on the same day. Only when such a combination exists can it be said that two or more people out of the whole group have birthdays on the same day. What you are doing, in fact, is conducting $n = \binom{m}{2}$ trials. Every trial has the same probability of success $p = \frac{1}{365}$ in showing the probability that two given people will have birthdays on the same day (this probability is the same as the probability that a person chosen at random matches your birthday). Assume that the random variable X indicates the number of trials where both people have birthdays on the same day. The probability that, in a group of m people, two or more people will have birthdays on the same day is then equal to $P(X \geq 1)$. Although the outcomes of the trials are dependent on one another, this dependence is considered to be weak because of the vast number (365) of possible birth dates. It is therefore reasonable to approximate the distribution of X using a Poisson distribution with expected value $\lambda = np$. In particular, $P(X \geq 1) \approx 1 - e^{-\lambda}$. In other words, the probability that, within a randomly formed group of m people, two or more people will have birthdays on the same day is approximately equal to

$$1 - e^{-\frac{1}{2}m(m-1)/365}.$$

This results in an approximate value of $1 - e^{-0.69315} = 0.5000$ for the probability that, in a group of 23 people, two or more people will have their birthdays on the same day. This is an excellent approximation for the exact value 0.5073 of this probability. The approximation approach with $\binom{23}{2} = 253$ trials and a success probability of $\frac{1}{365}$ on each trial explains why a relatively small group of 23 people is sufficient to give approximately a 50% probability of encountering two people with birthdays on the same day. The exact solution for the birthday problem does not provide this insight. The birthday problem is not the only problem in which the Poisson approximation method is a useful tool for a quick assessment of the magnitude of certain probabilities.

The exact solution to the birthday problem is easily derived, and the Poisson approximation is not necessarily required. This is different for the "almost" birthday problem: what is the probability that, within a randomly formed group of m people, two or more people will have birthdays within one day of each other? The derivation of an exact formula for this probability is far from simple, but a Poisson approximation is particularly simple to give. You must reconsider all the possible combinations of two people, that is, you must run $n = \binom{m}{2}$ trials. The probability of success in a given trial is now equal to $p = \frac{3}{365}$ (the probability that two given people will have birthdays within one day of each other). The number of successful trials is approximately Poisson distributed with an expected value of $\lambda = np$. In particular, the probability that two or more people will have birthdays within one day of each other is approximately equal to

$$1 - e^{-\frac{3}{2}m(m-1)/365}.$$

For $m = 14$, the approximate value is $1 - e^{-0.74795} = 0.5267$ (the exact value of the probability is 0.5375). The Poisson approximation method can be used to find solutions to many variants of the birthday problem.

A scratch-and-win lottery

A lottery organization distributes one million tickets every week. At one end of the ticket, there is a visible printed number consisting of six digits, say 070469. At the other end of the ticket, another six-digit number is printed, but this number is hidden by a layer of scratch-away silver paint. The ticket holder scratches the paint away to reveal the underlying number. If the number is the same as the number at the other end of the ticket, it is a winning ticket. The two six-digit numbers on each of the one million tickets printed each week are randomly generated in such a way that no two tickets are printed with the same visible numbers or the same hidden numbers. Assuming that all tickets are sold each week, the following questions are of interest to the lottery organizers. What is the probability distribution of the number of winners in any given week? In particular, what is the average number of winners per week?

The surprising answer is that the probability distribution of the number of winners in any given week is practically indistinguishable from a Poisson distribution with an expected value of 1. Even more astonishingly, the Poisson distribution with an expected value of 1 applies to any scratch lottery, regardless of whether the lottery issues one million six-digit tickets or 100 two-digit tickets. This is an astounding result that few will believe at first glance! However, the phenomenon can easily be explained by the Poisson-approximation approach. To do so, let's assume a scratch lottery that issues n different tickets with the printed numbers $1, \dots, n$ each week. Use the random variable X to denote the number of winners in any given week. The random variable X can be seen as the number of successes in n trials. In each trial the printed number and the hidden number on one of the tickets are compared. The success probability for each trial is $\frac{1}{n}$. If n is large enough, the dependence between the trials is weak enough to approximate the probability distribution of X by a Poisson distribution with an expected value of $\lambda = n \times \frac{1}{n} = 1$. In particular, the probability of no winner in any given week is approximately $\frac{1}{e} = 0.368$. It turns out that the Poisson distribution is indeed an excellent approximation to the exact probability distribution of X. The exact probability distribution will be given in Example 7.12 of Chapter 7. A numerical comparison of the exact distribution with the Poisson distribution reveals that $n = 10$ is sufficiently large in order for the Poisson probabilities to match the exact probabilities in at least eight decimals.

The scratch-and-win lottery problem is one of the many manifestations of the so-called *hat-check problem*. To explain this problem, imagine that, at a country wedding in France, all male guests throw their berets in a corner. After the reception, each guest takes a beret without bothering to check if it is his. The probability that at least one guest goes home with his own beret is approximately $1 - \frac{1}{e} = 0.632$. The origin of *matching problems* like the scratch-and-win lottery problem and the hat-check problem can be found in the book *Essay d'Analyse sur les Jeux de Hasard*, written in 1708 by Pierre Rémond de Montmort (1678–1719). In his book, Montmort solved a variant of the original card game *Jeu de Treize*, which is described in Problem 3.32. Montmort simplified this game by assuming that the deck of cards has only 13 cards of one suit. The dealer shuffles the cards and turns them up one at a time, calling out "Ace, two, three, . . . , king." A match occurs if the card that is turned over matches the rank called out by the dealer as he turns it over. The dealer wins if a match occurs. The probability of a match occurring is approximately $1 - \frac{1}{e} = 0.632$. A related problem was discussed in Marilyn vos Savant's column in *Parade Magazine* of August 21, 1994. An ordinary deck of 52 cards is thoroughly shuffled. The dealer turns over the cards one at a time, counting as he goes "ace, two, three, . . . , king, ace, two, . . . ," and so on, so that the dealer ends up calling out the 13 ranks four times each. A match occurs if the card that comes up matches the rank called out by the dealer as he turns it over. Using the Poisson-approximation method, it is easy to calculate an estimate of the probability of the occurrence of a match. There are $n = 52$ trials, and the success probability for each trial is $p = \frac{4}{52}$. The probability distribution of the number of matches is then approximated by a Poisson distribution with an expected value of $\lambda = 52 \times \frac{4}{52} = 4$. In particular, the probability of the dealer winning is approximated by $1 - e^{-4} = 0.9817$. This is again an excellent approximation. The exact value of the probability of the dealer winning is 0.9838, as can be calculated using the inclusion-exclusion rule in Chapter 7.

A lottery problem

What is the probability that, in 30 lottery drawings of six numbers from the numbers $1, \ldots, 45$, not each of these 45 numbers will be drawn at least once? This is the question that appears in Problem 4 of Chapter 1. To calculate this probability, a simple Poisson approximation can be given. The chance experiment of 30 lottery drawings of six different numbers from the numbers $1, \ldots, 45$ includes trials $1, \ldots, 45$. The ith trial determines whether the number i appears in any of the 30 drawings and is considered successful when the number i does *not* come up in any of the 30 drawings. For each trial the probability of the pertinent number not being drawn in any of the 30 drawings is equal to $p = \left(\frac{39}{45}\right)^{30} = 0.0136635$. This calculation uses the fact that the probability of a specific number i not coming up in a given drawing is equal to $\frac{44}{45} \times \frac{43}{44} \times \cdots \times \frac{39}{40} = \frac{39}{45}$. Although a slight dependence does exist between the trials, it seems reasonable to estimate the distribution of the amount of numbers that will not come up in 30 drawings by using a Poisson distribution with an expected value of $\lambda = 45 \times 0.0136635 = 0.61486$. This gives a surprising result: the probability that not each of the 45 numbers will come up in 30 drawings is approximately equal to $1 - e^{-0.61486} = 0.4593$. The exact value of the probability is 0.4722, as can be calculated using the inclusion-exclusion rule in Chapter 7. The methodology used for the lottery problem can also be applied to the coupon collector's problem set forth in Section 3.2.

The coupon collector's problem

How large should a group of randomly chosen persons be in order to have represented all of the 365 birthdays (excluding February 29) with a probability of at least 50%? This is in fact the coupon collector's problem from Chapter 3. To answer the question, take a group of size m with m fixed ($m > 365$). Imagine that you conduct a trial for each of the 365 days of the year. Trial i is said to be successful if day i is *not* among the birthdays of the m people in the group. Each trial has the same success probability of $p = \left(\frac{364}{365}\right)^m$. By the Poisson model for weakly dependent trials, the probability of no success among the 365 trials is approximately equal to $e^{-\lambda(m)}$ with $\lambda(m) = 365 \times p$. Thus, the probability of having represented all of the 365 birthdays in the group of m people is approximately equal to $e^{-\lambda(m)}$. The smallest value of m for which $e^{-\lambda(m)} \geq 0.5$ is $m = 2285$. Thus, the group size should be approximately equal to 2285 in order to have represented all of the 365 birthdays with a probability of at least 50%. The exact answer is 2287. You might wonder how the exact answer is calculated. The exact value of the probability that the group size should be more than m in order to have represented all of the 365 birthdays can be calculated from the inclusion-exclusion formula from Chapter 7. More advanced methods to solve the coupon collector's problem are the generating function approach from Chapter 14 and absorbing Markov chains from Chapter 15.

4.2.4 The Poisson process[†]

The Poisson process is inseparably linked to the Poisson distribution. This process is used to count events that occur randomly in time. Examples include: the emission of particles from a radioactive source, the arrival of claims at an insurance company, the occurrence of serious earthquakes, the occurrence of power outages, and the arrival of urgent calls to an emergency center. When does the process of counting events qualify as a Poisson process? To specify this, it is convenient to consider the Poisson process in terms of customers arriving at a facility. As such, it is necessary to begin with the assumption of a population *unlimited in size* of potential customers, in which the customers act independently of one another. The process of customer arrivals at a service facility is called a *Poisson process* if the process possesses the following properties:

A the customers arrive one at a time
B the numbers of arrivals during nonoverlapping time intervals are independent of one another
C the number of arrivals during any given time interval has a Poisson distribution of which the expected value is proportional to the duration of the interval.

Defining the *arrival intensity* of the Poisson process by

$$\alpha = \text{the expected value of the number of arrivals during a given time interval of unit length,}$$

[†] This section is earmarked for the more advanced student and may be set aside for subsequent readings by the novice.

then property **C** demands that, for each $t > 0$, it is true that

$$\begin{aligned}&P(k \text{ arrivals during a given time interval of duration } t)\\&= e^{-\alpha t}\frac{(\alpha t)^k}{k!} \qquad \text{for } k = 0, 1, \ldots .\end{aligned}$$

Also, by property **B**, the joint probability of j arrivals during a given time interval of length t and k arrivals during another given time interval of length u is equal to $e^{-\alpha t}\frac{(\alpha t)^j}{j!} \times e^{-\alpha u}\frac{(\alpha u)^k}{k!}$, provided that the two intervals are nonoverlapping.

The assumptions of the Poisson process are natural assumptions that hold in many practical situations.[†] The Poisson process is an example of a model that fulfills the dual objectives of realism and simplicity. The practical applicability of the Poisson process gets further support by the fact that condition **C** can be weakened to the requirement that the probability mass function of the number of arrivals in any time interval $(s, s + t)$ depends only on the length t of the interval and not on its position on the time axis. In conjunction with conditions **A** and **B**, this requirement suffices to prove that the number of arrivals in any time interval is Poisson distributed. Condition **B** expresses the absence of after-effects in the arrival process; that is, the number of arrivals in any time interval $(s, s + t)$ does not depend on the sequence of arrivals up to time s. The condition **B** is crucial for the Poisson process and cannot be satisfied unless the calling population of customers is very large. The absence of after-effects in the arrival process arises when the calling population is very large, customers act independently of each other, and any particular customer rarely causes an arrival. For example, this explains why a Poisson process can be used to describe the emission of particles by a radioactive source with its very many atoms, which act independently of one another and decay with a very small probability.

A construction of the Poisson process

A physical construction of the Poisson process is as follows. Split the time axis up into intervals of length Δt with Δt very small. Assume also that during a given interval of length Δt, the probability that precisely one customer will arrive is equal to $\alpha\Delta t$, and the probability that no customer will arrive is equal to $1 - \alpha\Delta t$, independently of what has happened before the interval in question. In this way, if a given interval of length t is split up into n smaller intervals each

[†] A nice illustration can be found in S. Chu, "Using soccer goals to motivate the Poisson process," *Informs Transactions on Education* **3** (2003): 62–68.

having length Δt, then the number of arrivals during the interval of length t has a binomial distribution with parameters $n = \frac{t}{\Delta t}$ and $p = \alpha \Delta t$. Now, let $\Delta t \to 0$, or equivalently $n \to \infty$. Because the Poisson distribution is a limiting case of the binomial distribution, it follows that the number of arrivals during an interval of length t has a Poisson distribution with an expected value of $np = \alpha t$. This construction of the Poisson process is especially useful and may be extended to include the situation in which customer arrival intensity is dependent on time.

The construction of a Poisson process on the line can be generalized to a Poisson process in the plane or other higher-dimensional spaces. The Poisson model defines a random way to distribute points in a higher-dimensional space. Examples are defects on a sheet of material and stars in the sky.

Relationship with the exponential distribution

In a Poisson arrival process the number of arrivals during a given time interval is a discrete random variable, but the time between two successive arrivals can take on any positive value and is thus a so-called continuous random variable. This can be seen in the following:

$$\begin{aligned}
&P(\text{time between two successive arrivals is greater than } y)\\
&\quad = P(\text{during an interval of duration } y \text{ there are no arrivals})\\
&\quad = e^{-\alpha y} \qquad \text{for each } y > 0.
\end{aligned}$$

Thus in a Poisson arrival process with an arrival intensity α, the time T between two successive arrivals has the probability distribution function

$$P(T \leq y) = 1 - e^{-\alpha y} \qquad \text{for } y \geq 0.$$

This continuous distribution is known as the *exponential distribution* (see also Chapter 10). The expected value of the interarrival time T is $\frac{1}{\alpha}$. Given property **B** of the Poisson process, it will not come as a surprise to anyone that the intervals between the arrivals of successive clients are independent from each other. A more surprising property of the Poisson process is as follows: for every fixed point in time, the waiting period from that point until the first arrival after that point has the *same* exponential distribution as the interarrival times, regardless of how long it has been since the last client arrived before that point in time. This extremely important *memoryless property* of the Poisson process can be shown with the help of property **B** of the Poisson process, which says that the number of arrivals in nonoverlapping intervals are independent from one another. The memoryless property is characteristic for the Poisson process.

Example 4.8 Out in front of Central Station, multiple-passenger taxicabs wait until they have either acquired four passengers or a period of ten minutes has passed since the first passenger stepped into the cab. Passengers arrive according to a Poisson process with an average of one passenger every three minutes.

(a) You are the first passenger to get into a cab. What is the probability that you will have to wait ten minutes before the cab gets underway?
(b) You were the first passenger to get into a cab and you have been waiting there for five minutes. In the meantime, two other passengers have entered the cab. What is the probability that you will have to wait another five minutes before the cab gets underway?

Solution. The answer to question (a) rests on the observation that you will only have to wait ten minutes if, during the next ten minutes, fewer than three other passengers arrive. This gives us:

$$\begin{aligned}&P(\text{you must wait ten minutes})\\&\quad= P(0,\ 1 \text{ or } 2 \text{ passengers arrive within the next ten minutes})\\&\quad= e^{-10/3} + e^{-10/3}\frac{(10/3)^1}{1!} + e^{-10/3}\frac{(10/3)^2}{2!} = 0.3528.\end{aligned}$$

Solving question (b) rests on the memoryless property of the Poisson process. The waiting period before the arrival of the next passenger is exponentially distributed with an expected value of three minutes, regardless of how long ago the last passenger arrived. You will have to wait another five minutes if this waiting period is longer than five minutes. Thus, the probability of having to wait another five minutes is then $e^{-5/3} = 0.1889$.

It is emphasized again that the Poisson process has both a discrete component (Poisson distribution for the number of arrivals) and a continuous component (exponential distribution for the interarrival times). Students mix up these two things sometimes. It may be helpful to think of the following situation. The emission of alpha-particles by a piece of radioactive material can be described by a Poisson process: the number of particles emitted in any fixed time interval is a discrete random variable with a Poisson distribution and the times between successive emissions are continuous random variables with an exponential distribution.

Clustering of arrival times

Customer arrival times reveal a tendency to cluster. This is clearly shown in Figure 4.1. This figure gives simulated arrival times in the time interval (0, 45)

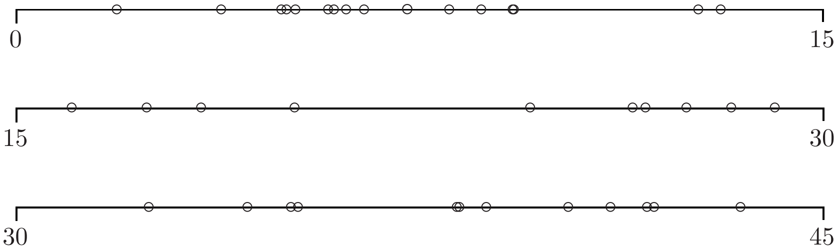


Fig. 4.1. Arrival times of a Poisson process.

for a Poisson process with arrival intensity $\alpha = 1$. A mathematical explanation of the clustering phenomenon can be given. As shown before, the interarrival time T has probability distribution function $P(T \leq y) = 1 - e^{-\alpha y}$ for $y \geq 0$. The derivative of the function $F(y) = 1 - e^{-\alpha y}$ is given by $f(y) = \alpha e^{-\alpha y}$. By definition, $f(y) = \lim_{\Delta y \to 0} [F(y + \Delta y) - F(y)] / \Delta y$. This implies that

$$P(y < T \leq y + \Delta y) \approx f(y)\Delta y \qquad \text{for } \Delta y \text{ small}$$

(see also Chapter 10). The function $f(y) = \alpha e^{-\alpha y}$ is largest at $y = 0$ and decreases from $y = 0$ onward. Hence the point y at which $P(y < T \leq y + \Delta y)$ is largest for fixed Δy is the point $y = 0$. Thus, short interarrival times occur relatively frequently and this suggests that a random series of arrivals will show a considerable tendency to cluster. The phenomenon of clustered arrival times casts an interesting light on a series of murders in Florida that caused a great deal of turmoil. In the period between October 1992 and October 1993, nine tourists of international origins were murdered in Florida. The murders were attributed to the fact that foreign tourists could easily be recognized as such because they drove rental cars. This could well have been one explanation for the explosion of murders, but it is also quite possible that one can only speak of a "normal" probability event when it is observable over a greater period of time. Assume that for each day there is a 1% probability of a foreign tourist being murdered somewhere in Florida. The random process showing the occurrence of foreign tourist murders in Florida over time can reasonably be modeled as a Poisson process with an intensity of 3.65 murders per year. Now, what is the probability that somewhere within a time frame of say, ten years, there will be one 12-month period containing nine or more foreign tourist murders? There is no easy formula for computing this probability, but a solution can easily be found by means of computer simulation (it will be explained later how to simulate arrival times in a Poisson process). The probability is approximately 36%. Over a period of 20 years, the probability of such a series of murders increases to approximately 60%. This contrasts with the probability of nine or more murders in a *given* 12-month period of 0.0127, a much smaller probability than the ones obtained for a moving time frame. In the situation of a moving time frame, the clustering phenomenon compounds the "law of coincidences": if you give something enough of a chance to happen, it eventually will. Also, the large number of shark attacks that took place in the summer of 2001 in Florida might be seen in a wider context through the clustering property of the Poisson process.

Example 4.9 In a given city, traffic accidents occur according to a Poisson process with an average of $\lambda = 10$ accidents per week. In a certain week, seven accidents have occurred. What is the probability that exactly one accident has occurred on each day of that week? Can you explain beforehand why this probability must be small?

Solution. Let the random variable $N(t)$ denote the number of accidents occurring in the time interval $(0, t)$, where a day is taken as time unit. Letting the epoch $t = u - 1$ correspond to the beginning of day u for $u = 1, 2, \ldots, 7$, the probability we are seeking is given by

$$P(N(u) - N(u-1) = 1 \text{ for } u = 1, \ldots, 7 \mid N(7) = 7)$$

with the convention $N(0) = 0$. By the properties **B** and **C** of the Poisson process, the random variables $N(1),\ N(2) - N(1), \ldots,\ N(7) - N(6)$ are independent and have a Poisson distribution with expected value $\lambda/7$. Also, by property **C**, the random variable $N(7)$ is Poisson distributed with expected value λ. Thus, the desired probability is equal to

$$\begin{aligned}&\frac{P(N(u) - N(u-1) = 1 \text{ for } u = 1, \ldots, 7)}{P(N(7) = 7)}\\ &\quad= \frac{P(N(1) = 1) \times P(N(2) - N(1) = 1) \times \cdots \times P(N(7) - N(6) = 1)}{P(N(7) = 7)}\\ &\quad= \frac{e^{-\lambda/7}(\lambda/7) \times e^{-\lambda/7}(\lambda/7) \times \cdots \times e^{-\lambda/7}(\lambda/7)}{e^{-\lambda}\lambda^7/7!} = \frac{7!}{7^7}.\end{aligned}$$

Hence, the desired probability is equal to 0.0162. Indeed, a small probability. The tendency of Poisson arrivals to cluster explains why the probability is so small. Incidentally, the probability $7!/7^7$ is the same as the probability of getting exactly one random number in each of the seven intervals $(0, \frac{1}{7}), (\frac{1}{7}, \frac{2}{7}), \ldots, (\frac{6}{7}, 1)$ when drawing seven independent random numbers from $(0, 1)$.

It can be proved that there is a close relationship between the Poisson arrival process and the uniform distribution: under the condition that exactly r arrivals have occurred in the fixed time interval $(0, t)$, then the r arrival epochs will be statistically indistinguishable from r random points that are independently chosen in the interval $(0, t)$. This result provides another explanation of the clustering phenomenon in the Poisson process: it is inherent to randomly chosen points in an interval that these points are not evenly distributed over the interval. The relation between the uniform distribution and the Poisson process on the line extends to the Poisson process in the plane or other higher-dimensional spaces: under the condition that exactly r entities (e.g., stars) are contained in a given bounded region, then the positions of the r entities will be distributed as r random points that are independently chosen in the region. This is a useful result for simulating a Poisson process in the plane or other higher-dimensional spaces. A simpler procedure to simulate a Poisson process on the line is as follows.

Simulating a Poisson process

There are several ways to simulate arrival times of a Poisson process. The easiest method is based on the result that the Poisson process with arrival intensity α can be equivalently defined by assuming single arrivals with interarrival times that are independent and have an exponential distribution with expected value $\frac{1}{\alpha}$. In Chapter 10 the reader will be asked to show that the random variable $X = -\frac{1}{\alpha}\ln(U)$ is exponentially distributed with expected value $\frac{1}{\alpha}$ if U is uniformly distributed on (0,1). This leads to the following algorithm for generating an interarrival time:

1. Generate a random number u between 0 and 1.
2. Take $x = -\frac{1}{\alpha}\ln(u)$ as the interarrival time.

This simple procedure gives the reader the power to verify the probabilities cited in the example of the Florida murders by means of a simulation study.

Merging and splitting Poisson processes

In applications of the Poisson process, it is frequently necessary to link two Poisson processes together, or to thin out one Poisson process. For example, consider a call center that functions as the telephone information facility for two completely different business organizations. Calls come in for the first company A according to a Poisson process with arrival intensity λ_A, and, independently of that, calls come in for the other company B according to a Poisson process with arrival intensity λ_B. The merging of these two arrival processes can be shown to give us a Poisson process with arrival intensity $\lambda_A + \lambda_B$. It can also be shown that any future telephone call will be for company A with a probability of $\frac{\lambda_A}{\lambda_A+\lambda_B}$ and will be for company B with a probability of $\frac{\lambda_B}{\lambda_A+\lambda_B}$.

In order to show how a Poisson process can be split up, we will refer to the example of a Poisson process with intensity λ that describes the occurrence of earthquakes in a certain region. Assume that the magnitudes of the earthquakes are independent from one another. Any earthquake is classified as being a high-magnitude earthquake with probability p and as being a low-magnitude earthquake with probability $1 - p$. Then, the process describing the occurrence of high-magnitude earthquakes is a Poisson process with intensity λp, and the occurrence of low-magnitude earthquakes is described by a Poisson process with intensity $\lambda(1 - p)$. It is surprising to find that these two Poisson processes are independent from one another!

Example 4.10 A piece of radioactive material emits alpha-particles according to a Poisson process with an intensity of 0.84 particle per second. A counter detects each emitted particle, independently, with probability 0.95. In a ten-second period the number of detected particles is 12. What is the probability that more than 15 particles were emitted in that period?

Solution. The process describing the emission of undetected particles is a Poisson process with an intensity of $0.84 \times 0.05 = 0.0402$ particle per second and the process is independent of the Poisson process describing the emission of detected particles. Thus, the number of emitted particles that were missed by the counter in the ten-second period has a Poisson distribution with expected value $10 \times 0.0402 = 0.402$, irrespective of how many particles were detected in that period. The desired probability is the probability of having more than three emissions of undetected particles in the ten-second period and is given by $1 - \sum_{j=0}^{3} e^{-0.402}(0.402)^j/j! = 0.00079$.

4.3 The hypergeometric distribution

The *urn* model is at the root of the hypergeometric distribution. In this model, you have an urn that is filled with R red balls and W white balls. You must randomly select n balls out of the urn without replacing any. What is the probability that, out of the n selected balls, r balls will be red? When the random variable X represents the number of red balls among the selected balls, this probability is given as follows

$$P(X = r) = \frac{\binom{R}{r}\binom{W}{n-r}}{\binom{R+W}{n}} \qquad \text{for } r = 0, 1, \ldots, n.$$

This is called the *hypergeometric distribution* with parameters R, W, and n. This comes with the understanding that $P(X = r) = 0$ for impossible combinations, or rather for values of r when $r > R$ or $n - r > W$. In skimming over the above formula, imagine that the R red balls are numbered $1, \ldots, R$ and the W white balls are numbered $R + 1, \ldots, R + W$. There are, in total, $\binom{R+W}{n}$ different ways to select n balls from the $R + W$ balls in the urn, and there are $\binom{R}{r}$ times $\binom{W}{n-r}$ different ways to select r balls from the R red balls and $n - r$ balls from the W white balls. Each of these outcomes is equally probable. When you divide the number of favorable outcomes by the total number of possible outcomes, you get the above formula.

The hypergeometric distribution has the expected value

$$E(X) = n\frac{R}{R + W}.$$

The proof is simple. Write $X = Y_1 + \cdots + Y_n$, where Y_i is equal to 1 if the ith drawn ball is red and 0 otherwise. For reasons of symmetry, each of the random variables Y_i has the same distribution as Y_1. Noting that $E(Y_1) = 1 \times \frac{R}{R+W}$ and using the fact that the expected value of a sum is the sum of the expected values, the desired result follows.

The hypergeometric distribution is often used when calculating the probability of winning prize money in a lottery.[†] The examples that follow show that when gambling with money, one is better off in a casino than taking part in a lottery. Lotteries often sell themselves by using psychological tricks to make one's chances of winning appear higher than they are in reality. Lottery organizers are perennial peddlers of hope! Providing hope to the masses is their ironclad sales objective. The purchasers of this laudable commodity, however, ordinarily see no more than 50% of their outlay return in the form of prize money.

Example 4.11 In the game "Lucky 10," 20 numbers are drawn from the numbers $1, \ldots, 80$. One plays this game by ticking one's choice of 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 number(s) on the game form. Table 4.3 indicates what the payoff rate is, depending on how many numbers are ticked and how many of those are correct. Also we give the chance of winning for each of the various combinations

[†] The modern lottery with prize money attached has its origins in the Netherlands: the oldest known lotteries of this kind have been traced as far back as 1444–1445, to the state lotteries of Brugge and Utrecht. Indeed, the local sovereign in Brugge, Philips de Goede (Philips the Good), moved quickly to require lottery organizers to obtain a license requiring them to hand one-third of the lottery profits over to him. Because such a high percentage of ticket sale monies went to the "house" (i.e., Philips de Goede), taking part in the lottery soon became tantamount to making a voluntary tax contribution.

Table 4.3. *Winning combinations in Lucky 10.*

Player's choice	Match	Payoff	Chance of winning
10 numbers	10/10	\$100,000	1.12×10^{-7}
	9/10	\$4,000	6.12×10^{-6}
	8/10	\$200	1.35×10^{-4}
	7/10	\$30	1.61×10^{-3}
	6/10	\$8	1.15×10^{-2}
	5/10	\$2	5.14×10^{-2}
	4/10	\$1	1.47×10^{-1}
	0/10	\$1	4.58×10^{-2}
9 numbers	9/9	\$25,000	7.24×10^{-7}
	8/9	\$2,000	3.26×10^{-5}
	7/9	\$200	5.92×10^{-4}
	6/9	\$15	5.72×10^{-3}
	5/9	\$3	3.26×10^{-2}
	4/9	\$1	1.14×10^{-1}
8 numbers	8/8	\$15,000	4.35×10^{-6}
	7/8	\$250	1.60×10^{-4}
	6/8	\$20	2.37×10^{-3}
	5/8	\$10	1.83×10^{-2}
	4/8	\$2	8.15×10^{-2}
7 numbers	7/7	\$2,000	2.44×10^{-5}
	6/7	\$80	7.32×10^{-4}
	5/7	\$12	8.64×10^{-3}
	4/7	\$2	5.22×10^{-2}
	3/7	\$1	1.75×10^{-1}
6 numbers	6/6	\$1,000	1.29×10^{-4}
	5/6	\$25	3.10×10^{-3}
	4/6	\$6	2.85×10^{-2}
	3/6	\$1	1.30×10^{-1}
5 numbers	5/5	\$200	6.45×10^{-4}
	4/5	\$8	1.21×10^{-2}
	3/5	\$3	8.39×10^{-2}
4 numbers	4/4	\$20	3.06×10^{-3}
	3/4	\$5	4.32×10^{-2}
	2/4	\$1	2.13×10^{-1}
3 numbers	3/3	\$16	1.39×10^{-2}
	2/3	\$2	1.39×10^{-1}
2 numbers	2/2	\$2	6.01×10^{-2}
	1/2	\$1	3.80×10^{-1}
1 number	1/1	\$2	2.50×10^{-1}

in Table 4.3. How are these chances of winning calculated and what are the expected payments per dollar staked when one ticks 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 number(s)?

Solution. In the case of ten numbers being ticked on the entry form, the following calculations apply (the same procedure is applicable in all of the other cases). Imagine that the 20 numbers drawn from the numbers $1, \ldots, 80$ are identified as $R = 20$ red balls in an urn and that the remaining 60, nonchosen numbers are identified as $W = 60$ white balls in the same urn. You have ticked ten numbers on your game form. The probability that you have chosen r numbers from the red group is simply the probability that r red balls will come up in the random drawing of $n = 10$ balls from the urn when no balls are replaced. This gives

$$P(r \text{ numbers correct out of 10 ticked numbers}) = \frac{\binom{20}{r}\binom{60}{10-r}}{\binom{80}{10}}.$$

Let us abbreviate this probability as $p_{r,10}$. With the data provided in Table 4.3, you will get an expected payoff of

$$\begin{aligned} &E(\text{payoff per dollar staked on ten ticked numbers}) \\ &\quad = 100{,}000 \times p_{10,10} + 4{,}000 \times p_{9,10} + 200 \times p_{8,10} + 30 \times p_{7,10} \\ &\qquad + 8 \times p_{6,10} + 2 \times p_{5,10} + 1 \times p_{4,10} + 1 \times p_{0,10}. \end{aligned}$$

When you enter the parameter values $R = 20$, $W = 60$, and $n = 10$ into a software module for the hypergeometric distribution, you get the numerical value of $p_{r,10}$ for each r. When these numerical values are filled in, you find

$$E(\text{payoff per dollar staked on ten ticked numbers}) = \$0.499.$$

In other words, the house percentage in the case of ten ticked numbers is 50.1%. The other house percentages in Table 4.4 are calculated in the same way.

In Table 4.4, the expected payoff per dollar staked on the total of ticked numbers is indicated. This is eye-opening information that you will not find on the Lucky 10 game form. The percentage of monies withheld on average by the lotto organizers says a lot. These house percentages linger in the neighborhood of 50% (and consider, on top of that, that many a lottery prize goes unclaimed!). That is quite a difference from the house percentage of 2.7% at a casino roulette wheel! But then, of course, when you play Lucky 10 you are not only lining the pockets of the lotto organizers, but you are also providing support for some worthy charities.

Table 4.4. *The average payoff on Lucky 10.*

Total numbers ticked	Average payoff per dollar staked	House percentage (%)
10	0.499	50.1
9	0.499	50.1
8	0.499	50.1
7	0.490	51.0
6	0.507	49.3
5	0.478	52.2
4	0.490	51.0
3	0.500	50.0
2	0.500	50.0
1	0.500	50.0

Example 4.12 The "New Amsterdam Lottery" offers the game "Take Five." In this game, players must tick five different numbers from the numbers 1, . . . , 39. The lottery draws five distinct numbers from the numbers 1, . . . , 39. For every one dollar staked, the payoff is \$100,000 for five correct numbers, \$500 for four correct numbers, and \$25 for three correct numbers. For two correct numbers, the player wins a free game. What is the house percentage for this lottery?

Solution. The hypergeometric model with $R = 5$, $W = 34$, and $n = 5$ is applicable in this case. This gives

$$P(\text{you have precisely } k \text{ numbers correct}) = \frac{\binom{5}{k}\binom{34}{5-k}}{\binom{39}{5}}.$$

The numerical value of the probability is 1.74×10^{-6}, 2.95×10^{-4}, 0.00974, and 0.10393 respectively for $k = 5, 4, 3$, and 2. The expected payoff per dollar staked is denoted with E. This results in

$$E = 1.74 \times 10^{-6} \times 100{,}000 + 2.95 \times 10^{-4} \times 500 + 0.00974 \times 25 \\ + 0.10393 \times E,$$

from which it follows that

$$E = \$0.456.$$

This means that the house percentage of the lottery is 54.4%. Nothing new under the sun: house percentages of lotteries the world over tend to be on the hefty side.

Table 4.5. *The winning combinations in the Powerball Lottery.*

You match	Payoff ($)	Chance of winning
5 white + Powerball	jackpot	8.30×10^{-9}
5 white	100,000	3.40×10^{-7}
4 white + Powerball	5,000	1.99×10^{-6}
4 white	100	0.0000816
3 white + Powerball	100	0.0000936
3 white	7	0.0038372
2 white + Powerball	7	0.0014350
1 white + Powerball	4	0.0080721
0 white + Powerball	3	0.0142068

Example 4.13[†] In the Powerball Lottery, five white balls are drawn from a drum containing 53 white balls numbered from $1, \ldots, 53$, and one red ball (Powerball) is drawn from 42 red balls numbered from $1, \ldots, 42$. This lottery is played in large sections of North America. On the game form, players must tick five "white" numbers from the numbers $1, \ldots, 53$ and one red number from the numbers $1, \ldots, 42$. The winning combinations with the corresponding payoffs and win probabilities are given in Table 4.5. The prizes are based on fixed monetary amounts except for the jackpot, which varies in its amounts and sometimes has to be divided among a number of winners. The amount of cash in the jackpot increases continuously until such time as it is won.

Solution. The calculation of the chances of winning rests on the hypergeometric distribution and the product formula for probabilities. The probability of choosing k white balls and the red Powerball correctly on one ticket is given as

$$\frac{\binom{5}{k}\binom{48}{5-k}}{\binom{53}{5}} \times \frac{1}{42},$$

while the probability of choosing just k white balls correctly is equal to

$$\frac{\binom{5}{k}\binom{48}{5-k}}{\binom{53}{5}} \times \frac{41}{42}.$$

The probability of winning the jackpot on one ticket is inconceivably small: 1 in 121 million. It is difficult to represent, in real terms, just how small this

[†] This example and the ensuing discussion are based on the teaching aid "Using lotteries teaching a chance course," available at *www.dartmouth.edu/~chance*.

probability is. It can best be attempted as follows: if you enter 12 Powerball tickets every week, then you will need approximately 134,000 years in order to have about a 50% chance of winning the jackpot at some time in your life (you can verify this for yourself by using the Poisson distribution!).

The Powerball game costs the player one dollar per ticket. The expected payoff for one ticket depends on the size of the jackpot and the total number of entries. The winning combinations, except the jackpot, make the following contribution to the expected value of the payoff for one ticket:

$$\begin{aligned}&100{,}000 \times 0.0000003402 + 5{,}000 \times 0.000001991 + 100 \times 0.00008164\\&\quad + 100 \times 0.00009359 + 7 \times 0.0038372 + 7 \times 0.001435\\&\quad + 4 \times 0.0080721 + 3 \times 0.0142068 = 0.1733 \text{ dollars.}\end{aligned}$$

In order to determine how much the jackpot contributes to the expected payoff, let's assume the following: there is a jackpot of 100 million dollars and 150 million tickets have been randomly filled out and entered. In calculating the jackpot's contribution to the expected payoff for any given ticket, you need the probability distribution of the number of winners of the jackpot among the remaining $n = 149{,}999{,}999$ tickets. The probability that any given ticket is a winning ticket is $p = 8.2969 \times 10^{-9}$. Thus, the probability distribution of the number of winning tickets is a Poisson distribution with an expected value of $\lambda = np = 1.2445$. This means that the contribution of the jackpot to the expected payoff of any given ticket is equal to

$$p \times \left(\sum_{k=0}^{\infty} \frac{1}{k+1} e^{-\lambda} \frac{\lambda^k}{k!}\right) \times 10^8 = 0.4746 \text{ dollars.}$$

The value of the expected payoff for any one dollar ticket, then, is equal to $0.1733 + 0.4746 = 0.6479$ dollars when the jackpot contains 100 million dollars and 150 million tickets are randomly filled out and entered.

Choosing lottery numbers

There is no reasonable way to improve your chances of winning at Lotto except to fill in more tickets. That said, it is to one's advantage not to choose "popular" numbers, i.e., numbers that a great many people might choose, when filling one's ticket in. If you are playing Lotto 6/45, for example, and you choose 1-2-3-4-5-6 or 7-14-21-28-35-42 as your six numbers, then you can be sure that you will have to share the jackpot with a huge number of others should it come up as the winning sequence. People use birth dates, lucky numbers,

Table 4.6. *Relative frequencies of numbers chosen.*

37	0.010	34	0.014	18	0.022	21	0.025	10	0.029
38	0.011	40	0.015	30	0.023	15	0.025	4	0.029
43	0.012	32	0.015	19	0.023	25	0.026	8	0.030
45	0.012	35	0.016	27	0.023	1	0.026	12	0.030
39	0.012	33	0.018	24	0.024	22	0.026	11	0.031
44	0.012	20	0.019	14	0.024	13	0.026	3	0.033
41	0.013	29	0.020	26	0.024	23	0.027	5	0.033
36	0.013	28	0.020	16	0.024	6	0.028	9	0.033
42	0.014	31	0.020	17	0.024	2	0.029	7	0.036

arithmetical sequences, etc., in order to choose lottery numbers. This is nicely illustrated in an empirical study done in 1996 for the Powerball Lottery. At that time, players of the Powerball Lottery chose six numbers from the numbers 1, . . . , 45 (before 1997 the Powerball Lottery consisted of the selection of five white balls out of a drum containing 45 white balls, and one red ball out of a drum containing 45 red balls). In total, a good 100,000 hand-written ticket numbers were analyzed. The relative frequencies of numbers chosen are given, in increasing order, in Table 4.6. No statistical tests are necessary in order to recognize that these people did not choose their numbers randomly. Table 4.6 indicates that people often use birth dates in choosing lottery numbers: the numbers 1 through 12, which may refer to both days of the month and months of the year, are frequently chosen. In lotteries where the majority of numbers in a series must be filled in by hand, it appears that the number of winners is largest when most of the six numbers drawn fall in the lower range.[†] When it comes to choosing nonpopular numbers in betting games, racetrack betting offers the reverse situation to the lottery: at the end of the day, when the last races are being run, one does well to bet on the favorites. The reason for this is that gamblers facing a loss for the day, and hoping to recover that loss before it is too late, will most often place bets on nonfavorites with a high payoff.

The fact that people do not choose their number sequences randomly decreases the probability that they will be the only winner, should they be lucky enough to win. On January 14, 1995, the UK National Lottery had a record number of 133 winners sharing a jackpot of £16,293,830. In this

[†] In the case of the majority of tickets being required to be filled in by hand, intelligent number choices can be found in N. Henze and H. Riedwyl's *How to Win More*, A. K. Peters, Massachusetts, 1998. These choices do not increase one's chances of winning, but they do increase the expected payoff for someone who is lucky enough to win.

lottery, six numbers must be ticked from the numbers $1, \ldots, 49$. Before the drawing in question, players had filled in 69.8 million tickets, the vast majority by hand. Had all of the tickets been filled in randomly, the probability of 133 or more winners would be somewhere on the order of 10^{-136} (verify this using the Poisson distribution with an expected value of $69{,}800{,}000/\binom{49}{6} = 4.99$). This inconceivably small probability indicates again that people do not choose their lottery numbers randomly. The winning sequence of the draw on Saturday January 14, 1995 was 7-17-23-32-38-42. The popularity of this sequence may be explained from the fact that the numbers 17, 32, and 42 were winning numbers in the draw two weeks before the draw on January 14, 1995.

Ticking the number sequences 1-2-3-4-5-6 or 7-14-21-28-35-42 is about the most foolish thing you can do in a lottery. In the improbable case that those six numbers actually come up winners, you can be quite sure that a massive number of players will be sharing the jackpot with you. This is what happened on June 18, 1983, in the Illinois Lotto Game, when 78 players won the jackpot with the number sequence 7-13-14-21-28-35. Today, most lotto games offer players “Quick Pick” or “Easy Pick” opportunities to choose their numbers randomly with the aid of a computer.[†] As the percentage of plays using random play grows, the number of winners becomes more predictable. The game then becomes less volatile and exorbitantly high jackpots are seen less frequently. As more hand-written tickets are entered into a lottery, the probability of a rollover of the jackpot will get larger. This is plausible if one considers the extreme case of all players choosing the same six numbers. In such an extreme case, the jackpot will never be won. Lottery officials are, to a certain point, not unhappy to see rollovers of the jackpot, as it will naturally be accompanied by increased ticket sales!

A coincidence problem

The coincidence problem presented in Question 7 of Chapter 1 can be solved according to the hypergeometric model. A bit of imagination will show that this problem reflects the hypergeometric model with a drum containing $R = 500$ red marbles and $W = 999{,}498$ white marbles, from which $n = 500$ marbles will be chosen. The probability we are looking for to solve our coincidence

[†] Approximately 70% of the tickets entered in the Powerball Lottery are nowadays Easy Picks. Players of the German Lotto, by contrast, chose no more than 4% of their number sequences by computer.

problem is equal to the probability of at least one red marble being drawn. This probability is equal to 0.2214. Stated in other terms, there is approximately a 22% probability that the two people in question have a common acquaintance. This answer, together with the answer to Question 1 of Chapter 1, reminds us that events are often less "coincidental" than we may tend to think.

4.4 Problems

4.1 During World War II, London was heavily bombed by V-2 guided ballistic rockets. These rockets, luckily, were not particularly accurate at hitting targets. The number of direct hits in the southern section of London has been analyzed by splitting the area up into 576 sectors measuring one quarter of a square kilometer each. The average number of direct hits per sector was 0.9323. The relative frequency of the number of sectors with k direct hits is determined for $k = 0, 1, \ldots$. In your opinion, which distribution is applicable in determining the frequency distribution of the number of direct hits? Is it a Poisson distribution or a geometric distribution?

4.2 What are the chances of getting at least one six in one throw of six dice, at least two sixes in one throw of 12 dice, and at least three sixes in one throw of 18 dice?[†] Do you think these chances are the same?

4.3 In an attempt to increase his market share, the maker of Aha Cola has formulated an advertising campaign to be released during the upcoming European soccer championship. The image of an orange ball has been imprinted on the underside of approximately one out of every one thousand cola can pop-tops. Anyone turning in such a pop-top on or before a certain date will receive a free ticket for the soccer tournament finale. Fifteen hundred cans of cola have been purchased for a school party, and all fifteen hundred cans will be consumed on the evening in question. What is the probability that the school party will deliver one or more free tickets?

4.4 A game of chance played historically by Canadian Indians involved throwing eight flat beans into the air and seeing how they fell. The beans were symmetrical and were painted white on one side and black on the other. The bean thrower would win one point if an odd number of beans came up white, two points if either zero or eight white beans came up, and would lose one point for any other configuration. Does the bean-thrower have the advantage in this game?

4.5 One hundred and twenty-five mutual funds have agreed to take part in an elimination competition being sponsored by Four Leaf Clover investment magazine. The competition will last for two years and will consist of seven rounds. At the beginning of each quarter, each fund remaining in the competition will put together a holdings portfolio. Funds will go through to the next round if, at the end of the quarter, they have performed above the market average. Funds finishing at or below market average will be eliminated from the competition. We can assume that the

[†] In a letter dated November 22, 1693, the gambler Samuel Pepys posed this question to Isaac Newton. It was not a trivial question for Newton.

funds' successive quarterly performances are independent from one another and that there is a probability of $\frac{1}{2}$ that a fund will perform above average during any given quarter. Calculate the probability that at least one fund will come through all seven rounds successfully. Calculate the probability that three or more funds will come through all seven rounds.

4.6 In 1989, American investment publication *Money Magazine* assessed the performance of 277 important mutual funds over the previous ten years. For each of those ten years they looked at which mutual funds performed better than the S&P index. Research showed that five of the 277 funds performed better than the S&P index for eight or more years. Verify that the expected value of the number of funds performing better than the S&P index for eight years or more is equal to 15.2 when the investment portfolios of each fund have been compiled by a blindfolded monkey throwing darts at the *Wall Street Journal*. Assume that each annual portfolio has a 50% probability of performing better than the S&P index.

4.7 The keeper of a certain king's treasure receives the task of filling each of 100 urns with 100 gold coins. While fulfilling this task, he substitutes one lead coin for one gold coin in each urn. The king suspects deceit on the part of the sentry and has two methods at his disposal of auditing the contents of the urns. The first method consists of randomly choosing one coin from each of the 100 urns. The second method consists of randomly choosing four coins from each one of 25 of the 100 urns. Which method provides the largest probability of uncovering the deceit?

4.8 A military early-warning installation is constructed in a desert. The installation consists of five main detectors and a number of reserve detectors. If fewer than five detectors are working, the installation ceases to function. Every two months an inspection of the installation is mounted and at that time all detectors are replaced by new ones. There is a probability of 0.05 that any given detector will cease to function during the period between inspections. The detectors function independently of one another. How many reserve detectors are needed to ensure a probability of less than 0.1% that the system will cease to function between inspections?

4.9 In a game called "26" a player chooses one number from the numbers $1, \ldots, 6$ as point number. After this, the player rolls a collection of ten dice 13 times in succession. If the player's point number comes up 26 times or more, the player receives five times the amount staked on the game. Is this game to the player's advantage?

4.10 Operating from within a tax-haven, some quick-witted businessmen have started an Internet Web site called Stockgamble. Through this Web site, interested parties can play the stock markets in a number of countries. Each of the participating stock markets lists 24 stocks available in their country. The game is played on a daily basis and for each market the six stocks that have performed the best are noted at the end of each day. Participants each choose a market and click on six of the 24 stocks available. The minimum stake is \$5 and the maximum stake is \$1,000. The payoff is 100 times the stake if all six of the top performing stocks have been clicked on. What would the expected pay-off be per dollar staked if this "game of skill" was purely a game of chance?

4.11 Decco is played with an ordinary deck of 52 playing cards. It costs \$1 to play this game. Having purchased a ticket on which the 52 playing cards of an ordinary deck are represented, each player ticks his choice of one card from each of the four suits (the ten of hearts, jack of clubs, two of spades and ace of diamonds, for example). On the corresponding television show, broadcast live, one card is chosen randomly from each of the four suits. If the four cards chosen by a player on his/her ticket are the same as the four chosen on the show, the player wins \$5,000. A player having three of the four cards correct wins \$50. Two correct cards result in a win of \$5. One correct card wins the player a free playing ticket for the next draw. What is the house percentage of this exciting game?

4.12 Consider an experiment with three possible outcomes 1, 2, and 3, which occur with probabilities of p_1, p_2, and $p_3 = 1 - p_1 - p_2$. For a given value of n, n independent trials of the experiment are performed. The random variable X_i gives the number of times that the outcome i occurs for $i = 1, 2, 3$. Verify that

$$P(X_1 = k_1, X_2 = k_2) = \binom{n}{k_1}\binom{n-k_1}{k_2} p_1^{k_1} p_2^{k_2} p_3^{n-k_1-k_2}$$

for all $k_1, k_2 \geq 0$ with $k_1 + k_2 \leq n$. This distribution is called the *multinomial distribution* (with $r = 3$ possible outcomes).

4.13 For the upcoming drawing of the Telenet Lottery, five extra prizes have been added to the pot. Each prize consists of an all-expenses paid vacation trip. The five winners of the extra prize may choose from among three possible destinations $i = 1, 2$, and 3. The winners choose independently of each other. The probability that a given winner chooses destination i is equal to p_i with $p_1 = 0.5$, $p_2 = 0.3$ and $p_3 = 0.2$. What is the probability that both destination 2 and destination 3 will be chosen? What is the probability that not all three destinations will be chosen?

4.14 A particular scratch-lottery ticket has 16 painted boxes on it, each box having one of the numbers 1, 2, 5, 10, 100 or 1,000 hidden under the paint. When the paint is scratched off and it appears that a same number shows up in seven or more boxes, the player wins an amount equal to that number multiplied by the purchasing price of the ticket. It is understood that, in cases where more than one number appears seven or more times, the higher number will serve as the winner. The preprinted number assortments are established randomly in accordance with the premise that on average 25% of the numbers will be a 1, 20% of the numbers will be a 2, another 20% will be a 5, 15% will be a 10, 10% will be a 100, and 10% will be a 1,000. Use the multinomial distribution to find the house percentage in this game.

4.15 A particular game is played with five poker dice. Each die displays an ace, king, queen, jack, ten and nine. Players may bet on two of the six images displayed. When the dice are thrown and the bet-on images turn up, the player receives three times the amount wagered. In all other cases, the amount of the wager is forfeited. Is this game advantageous for the player?

4.16 In the kingdom of Lightstone, the game of Lotto 6/42 is played. In Lotto 6/42 six numbers out of the numbers $1, \ldots, 42$ are drawn. At the time of an oil sheik's visit to Lightstone the jackpot for the next drawing is listed at 27.5 million dollars. The oil sheik decides to take a gamble and orders his retinue to fill in 15 million tickets in his name. These 15 million tickets do not have to be filled in by hand; rather a Lotto computer fills them in by randomly generating 15 million sequences of six distinct numbers (note that this manner of "random picks" allows for the possibility of the same sequence being generated more than once). Suppose that the local people have purchased ten million tickets for the same jackpot, and assume that the sequences for these tickets are also the result of "random picks." Each ticket costs \$1. What is the probability that the oil sheik will win the jackpot and what is the probability that the oil sheik will be the only winner? What is the probability that the oil sheik will win back his initial outlay?

4.17 The Brederode Finance Corporation has begun the following advertising campaign in Holland. Each new loan application submitted is accompanied by a chance to win a prize of \$25,000. Every month 100 zip codes will be drawn in a lottery. In Holland each house address has a unique zip code and there are about 2,500,000 zip codes. Each serious applicant whose zip code is drawn will receive a \$25,000 prize. Considering that Brederode Finance Corporation receives 200 serious loan applications each month, calculate the distribution of the monthly amount that they will have to give away.

4.18 You are at an assembly where 500 other persons are also present. The organizers of the assembly are raffling off a prize to be shared by all of those present whose birthday falls on that particular day. What is the probability that you will win the prize?

4.19 An organization running the Lotto 6/45 analyzes 100,000 tickets that were filled-in by hand. On each ticket of the Lotto 6/45 six different numbers from the numbers $1, \ldots, 45$ are filled in. A particular pick of six numbers occurred eight times in the 100,000 tickets. What is the probability of the same sequence of six numbers turning up eight or more times in the 100,000 tickets? Assuming that the tickets are randomly filled in, calculate a Poisson approximation for this variant of the birthday probability.

4.20 In the Massachusetts Numbers Game, one number is drawn each day from the 10,000 four-digit number sequence $0000, 0001, \ldots, 9999$. Calculate a Poisson approximation for the probability that the same number will be chosen two or more times in the upcoming 625 drawings. Before making the calculations in this variant of the birthday problem, can you say why this probability cannot be negligibly small?

4.21 What is a Poisson approximation for the probability that in a randomly selected group of 25 persons, three or more will have birthdays on a same day. What is a Poisson approximation for the probability that three or more persons from the group will have birthdays falling within one day of each other?

4.22 Ten married couples are invited to a bridge party. Bridge partners are chosen at random, without regard to gender. What is the probability of at least one couple being paired as bridge partners? Calculate a Poisson approximation for this probability.

4.23 A group of 25 students is going on a study trip of 14 days. Calculate a Poisson approximation for the probability that during this trip two or more students from the group will have birthdays on the same day.

4.24 Three people each write down the numbers $1, \ldots, 10$ in a random order. Calculate a Poisson approximation for the probability that the three people all have one number in the same position.

4.25 What is the probability of two consecutive numbers appearing in any given lotto drawing of six numbers from the numbers $1, \ldots, 45$? Calculate a Poisson approximation for this probability. Also, calculate a Poisson approximation for the probability of three consecutive numbers appearing in any given drawing of the Lotto 6/45.

4.26 Calculate a Poisson approximation for the probability that in a randomly selected group of 2,287 persons all of the 365 possible birthdays will be represented.

4.27 Sixteen teams remain in a soccer tournament. A drawing of lots will determine which eight matches will be played. Before the drawing takes place, it is possible to place bets with bookmakers over the outcome of the drawing. You are asked to predict all eight matches, paying no regard to the order of the two teams in each match. Calculate a Poisson distribution for the number of correctly predicted matches.

4.28 Calculate a Poisson approximation for the probability that in a thoroughly shuffled deck of 52 playing cards, it will occur at least one time that two cards of the same face value will succeed one another in the deck (two aces, for example). In addition, make the same calculation for the probability of three cards of the same face value succeeding one another in the deck.

4.29 A company has 75 employees in service. The administrator of the company notices, to his astonishment, that there are seven days on which two or more employees have birthdays. Verify, by using a Poisson approximation, whether this is so astonishing after all.

4.30 Argue that the following two problems are manifestations of the "hat-check" problem:

(a) In a particular branch of a company, the 15 employees have agreed that, for the upcoming Christmas party, each employee will bring one present without putting any name on it. The presents will be distributed blindly among them during the party. What is the probability of not one person ending up with his/her own present?

(b) A certain person is taking part in a blind taste test of ten different wines. The person has been made aware of the names of the ten wine producers beforehand, but does not know what order the wines will be served in. He may only name a wine producer one time. After the tasting session is over, it turns out that he has correctly identified five of the ten wines. Do you think he is a connoisseur?

4.31 A businessman parks his car illegally for one hour, twice a day, along the banks of an Amsterdam canal. During the course of an ordinary day, parking attendants monitor the streets according to a Poisson process with an average of α rounds per hour. What is the probability that the businessman will be ticketed and fined on any given day?

4.32 In the first five months of the year 2000, the tram hit and killed seven pedestrians in Amsterdam, each case caused by the pedestrian's own carelessness. In preceding years, such accidents occurred at a rate of 3.7 times per year. Simulate a Poisson process to estimate the probability that within a period of ten years, a block of five months will occur during which seven or more fatal tram accidents happen (you can simplify the problem by assuming that all months have the same number of days). Would you say that the disproportionately large number of fatal tram accidents in the year 2000 is the result of bad luck or would you categorize it in other terms?

4.33 Calls arrive at a computer-controlled exchange according to a Poisson process at a rate of two calls per second. Use computer simulation to find the probability that during the busy hour there will be some period of 30 seconds in which 90 or more calls arrive.

4.34 During the course of a summer day, tourist buses come and go in the picturesque town of Edam according to a Poisson process with an average arrival rate of five buses per hour. Each bus stays approximately two hours in Edam, which is famous for its cheese. What is the distribution of the number of buses to be found in Edam at 4 p.m.? What would the answer be if each bus stayed one hour with a probability of $\frac{1}{4}$ and two hours with a probability of $\frac{3}{4}$?

4.35 Paying customers (i.e., those who park legally) arrive at a large parking lot according to a Poisson process with an average of 45 cars per hour. Independently of this, nonpaying customers (i.e., those who park illegally) arrive at the parking lot according to a Poisson process with an average of five cars per hour. The length of parking time has the same distribution for legal as for illegal parking customers. At a given moment in time, there are 75 cars parked in the parking lot. What is the probability that 15 or more of these 75 cars are parked illegally?

4.36 Suppose that emergency response units are distributed throughout a large area according to a two-dimensional Poisson process. That is, the number of response units in any given bounded region has a Poisson distribution whose expected value is proportional to the area of the region, and the numbers of response units in disjoint regions are independent. An incident occurs at some arbitrary point. Argue that the probability of having at least one response unit within a distance r is $1 - e^{-\alpha\pi r^2}$ for some constant $\alpha > 0$ (this probability distribution is called the Rayleigh distribution).

4.37 Take another look at the lottery problem in Section 3.5. If we divide the lottery numbers $1, \ldots, 366$ into three equal groups, then we can see from Table 3.3 that 17 or more days in December fall into the first group of low numbers $1, \ldots, 122$. In a fair drawing, what would be the probability of 17 or more days in December falling into the first group?

4.38 You play Bingo together with 35 other people. Each player purchases one card with 24 different numbers that were selected at random out of the numbers $1, \ldots, 80$. The organizer of the game calls out random numbers between 1 and 80, one at a time. The first player to achieve a card with all of his/her 24 numbers shouts "Bingo" and collects the entire stake money. For $k = 65, 70$, and 75, calculate the probability that more than k numbers must be called out before one of the players shouts "Bingo." What is the probability that you will be the first one to achieve a full card?

4.39 In the German "Lotto am Samstag," six regular numbers and one reserve number are drawn from the numbers $1, \ldots, 49$. On the lottery ticket, players must tick six different numbers out of the numbers $1, \ldots, 49$. There is also an area of the lottery ticket reserved for something known as the Super Number. This is a number chosen from the sequence $0, 1, \ldots, 9$. So, in addition to drawing the regular and reserve numbers, a Super Number between $0, 1, \ldots, 9$ is also drawn in Lotto am Samstag. The Super Number only comes into play in combination with six correctly chosen regular numbers. The eight winning combinations are: six regular numbers correct

+ Super Number (jackpot), six regular numbers correct, five regular numbers correct and the reserve number, four regular numbers correct, three regular numbers correct + the reserve number, and three regular numbers correct.

(a) Calculate the probability of winning on one ticket for each of these combinations.

(b) You purchase 12 tickets every week for the German Lotto am Samstag. How many years will you need in order to have at least a 50% chance of ever winning the jackpot in your lifetime?

4.40 In Lottoland, there is a weekly lottery in which six (standard) numbers plus one bonus number are drawn from the numbers 1, . . . , 45. In addition to this, one color is randomly drawn out of six colors. On the lottery ticket, six numbers and one color must be chosen. The players use the computer for a random selection of the six numbers and the color. Each ticket costs $1.50. The number of tickets sold is about the same each week. The prizes are allotted as shown in the table. The jackpot begins with 4 million dollars; this is augmented each week by another half million dollars if the jackpot is not won. The lottery does not publish information regarding ticket sales and intake, but does publish a weekly listing in the newspaper of the number of winners for each of the six top prizes. The top six prizes from the table had 2, 10, 14, 64, 676, and 3,784 winners over the last 50 drawings.

6 + color	jackpot*
6	$1 million*
5 + bonus number + color	$250,000*
5 + bonus number	$150,000*
5 + color	$2,500
5	$1,000
4 + bonus number + color	$375
4 + bonus number	$250
4 + color	$37.50
4	$25
3 + bonus number + color	$15
3 + bonus number	$10
3 + color	$7.50
3	$5

* prize is divided by multiple winners

(a) Estimate the amount of the weekly intake.

(b) Estimate the average number of weeks between jackpots being won and estimate the average size of the jackpot when it is won.

(c) Estimate the percentage of the intake that gets paid out as prize money.

5
Probability and statistics

Chapter 2 was devoted to the law of large numbers. This law tells you that you may estimate the probability of a given event A in a chance experiment by simulating many independent repetitions of the experiment. Then the probability $P(A)$ is estimated by the proportion of trials in which the event A occurred. This estimate has an error. No matter how many repetitions of the experiment are performed, the law of large numbers will not tell you exactly how close the estimate is to the true value of the probability $P(A)$. How to quantify the error? For that purpose, you can use standard tools from statistics. Note that

simulation is analogous to a sampling experiment in statistics. An important concept in dealing with sample data is the central limit theorem. This theorem states that the histogram of the data will approach a bell-shaped curve when the number of observations becomes very large. The central limit theorem is the basis for constructing confidence intervals for simulation results. The confidence interval provides a probability statement about the magnitude of the error of the sample average. A confidence interval is useful not only in the context of simulation experiments, but in situations that also crop up in our daily lives. Consider the example of estimating the unknown percentage of the voting population that will vote for a particular political party. Such an estimate can be made by doing a random sampling of the voting population at large. Finding a confidence interval for the estimate is then essential: this is what allows you to judge how confident one might be about the prediction of the opinion poll.

The concepts of the normal curve and standard deviation are at the center of the central limit theorem. The normal curve is a bell-shaped curve that appears in numerous applications of probability theory. It is a sort of universal curve for displaying probability mass. The normal curve is symmetric around the expected value of the underlying probability distribution. The peakedness of the curve is measured in terms of the standard deviation of the probability distribution. The standard deviation is a measure for the spread of a random variable around its expected value. It says something about how likely certain deviations from the expected value are. When independent random variables each having the same distribution are averaged together, the standard deviation is reduced according to the square root law. This law is at the heart of the central limit theorem.

The concept of standard deviation is of great importance in itself. In finance, standard deviation is a key concept and is used to measure the volatility (risk) of investment returns and stock returns. It is common wisdom in finance that diversification of a portfolio of stocks generally reduces the total risk exposure of the investment. In the situation of similar but independent stocks the volatility of the portfolio is reduced according to the square root of the number of stocks in the portfolio. The square root law also provides useful insight in inventory control. Aggregation of independent demands at similar retail outlets by replacing the outlets with a single large outlet reduces the total required safety stock. The safety stock needed to protect against random fluctuations in the demand is then reduced according to the square root of the number of retail outlets aggregated.

In the upcoming sections, we take a look at the normal curve and standard deviation before the central limit theorem and its application to confidence intervals are discussed.

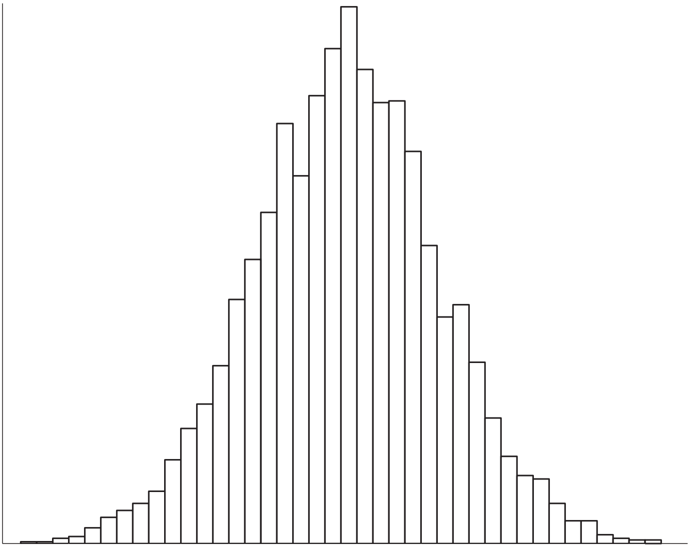


Fig. 5.1. Histogram of heights.

5.1 The normal curve

In many practical situations, histograms of measurements approximately follow a bell-shaped curve. A histogram is a bar chart that divides the range of values covered by the measurements into intervals of the same width, and shows the proportion of the measurements in each interval. For example, let's say you have the height measurements of a very large number of Dutch men between 20 and 30 years of age. To make a histogram, you break up the range of values covered by the measurements into a number of disjoint adjacent intervals each having the same width, say width Δ. The height of the bar on each interval $[j\Delta, (j+1)\Delta)$ is taken such that the area of the bar is equal to the proportion of the measurements falling in that interval (the proportion of measurements within the interval is divided by the width of the interval to obtain the height of the bar). The total area under the histogram in Figure 5.1 is thus standardized to one. Making the width Δ of the base intervals of the histogram smaller and smaller, the graph of the histogram will begin to look more and more like the bell-shaped curve shown in Figure 5.2.

The bell-shaped curve in Figure 5.2 can be described by a function $f(x)$ of the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}.$$

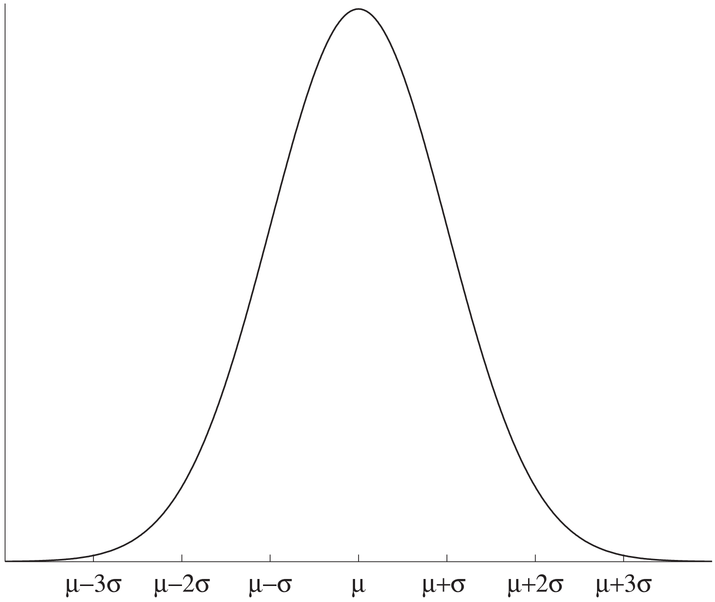


Fig. 5.2. The normal curve.

This function is defined on the real line and has two parameters μ and σ, where μ (the location parameter) is a real number and σ (the shape parameter) is a positive real number. The characteristic bell-shaped curve in Figure 5.2 is called the *normal curve*. It is also known as the Gaussian curve (of errors), after the famous mathematician/astronomer Carl Friedrich Gauss (1777–1855), who showed in a paper from 1809 that this bell curve is applicable with regard to the accidental errors that occur in the taking of astronomical measurements. It is usual to attribute the discovery of the normal curve to Gauss. However, the normal curve was discovered by the mathematician Abraham de Moivre (1667–1754) around 1730 when solving problems connected with games of chance. The pamphlet "Approximato ad Summani Terminorum Binomi $(a + b)^n$ in Seriem Expansis" containing this discovery was first made public in 1738 in the second edition of De Moivre's masterwork *Doctrine of Chance*. Also a publication of Pierre Simon Laplace (1749–1829) from 1778 contains the normal curve function and emphasizes its importance. De Moivre anticipated Laplace and the latter anticipated Gauss. One could say that the normal curve is a natural law of sorts, and it is worth noting that each of the three famous mathematical constants $\sqrt{2}$, $\pi = 3.141\ldots$ and $e = 2.718\ldots$ play roles in its makeup. Many natural phenomena, such as the height of men, harvest yields, errors in physical measurements, luminosity of stars, returns on stocks, can be described by a normal curve. The Belgian astronomer and statistician Adolphe Quetelet (1796–1894) was the first to recognize the universality of the normal curve and he fitted it to a large collection of data taken from all corners of science, including economics and the social sciences. Many in the eighteenth and nineteenth centuries considered the normal curve a God-given law. The universality of the bell-shaped Gaussian curve explains the popular use of the name normal curve for it. Later on in the text we shall present a mathematical explanation of the frequent occurrence of the normal curve with the help of the central limit theorem. But first we will give a few notable facts about the normal curve. It has a peak at the point $x = \mu$ and is symmetric around this point. Second, the total area under the curve is 1. Of the total area under the curve, approximately 68% is concentrated between points $\mu - \sigma$ and $\mu + \sigma$ and approximately 95% is concentrated between $\mu - 2\sigma$ and $\mu + 2\sigma$. Nearly the entire area is concentrated between points $\mu - 3\sigma$ and $\mu + 3\sigma$. For example, if the height of a certain person belonging to a particular group is normally distributed with parameters μ and σ, then it would be exceptional for another person from that same group to measure in at a height outside of the interval $(\mu - 3\sigma, \mu + 3\sigma)$.

5.1.1 Probability density function

Before giving further properties of the normal curve, it is helpful, informally, to discuss the concept of a probability density function. The function $f(x)$ describing the normal curve is an example of a probability density function. Any nonnegative function for which the total area under the graph of the function equals 1 is called a *probability density function*. Any probability density function underlies a so-called *continuous random variable*. Such a random variable can take on a continuum of values. The random variable describing the height of a randomly chosen person is an example of a continuous random variable if it is assumed that the height can be measured in infinite precision. Another example of a continuous random variable is the annual rainfall in a certain area or the time between serious earthquakes in a certain region. A probability density function can be seen as a "smoothed out" version of a probability histogram: if you take sufficiently many independent samples from a continuous random variable and the width of the base intervals of the histogram depicting the relative frequencies of the sampled values within each base interval is sufficiently narrow, then the histogram will resemble the probability density function of the continuous random variable. The probability histogram is made up of rectangles such that the area of each rectangle equals the proportion of the sampled values within

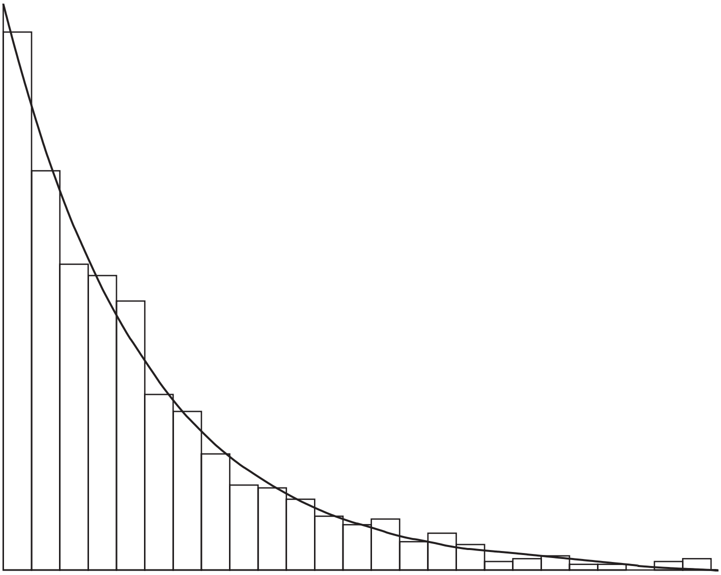


Fig. 5.3. Histogram of decay times.

the range of the base of the rectangle. For this normalization, the total area (or integral) under the histogram is equal to one. The area of any portion of the histogram is the proportion of the sampled values in the designated region. It is also the probability that a random observation from the continuous random variable will fall in the designated region. As an illustration, take the decay time of a radioactive particle. The decay time is a continuous random variable. Figure 5.3 displays the probability histogram of a large number of observations for the waiting times between counts from radioactive decay. Where the probability histogram in Figure 5.1 resembles a probability density function of the form $(\sigma\sqrt{2\pi})^{-1}e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$ for some values of the parameters μ and $\sigma > 0$, the probability histogram in Figure 5.3 resembles a probability density of the form $\lambda e^{-\lambda x}$ for some value of the parameter $\lambda > 0$. The area of the histogram between the base points t_1 and t_2 approximates the probability that the waiting time between counts will fall between t_1 and t_2 time units.

Taking the foregoing in mind, you may accept the fact that a continuous random variable X cannot be defined by assigning probabilities to individual values. For any number a, the probability that X takes on the value a is 0. Instead, a continuous random variable is described by assigning probabilities to intervals via a probability density function, where the probability assigned to an interval (a, b) is the same as the probability assigned to the interval $[a, b]$

that includes the point a. In Chapter 10, it will be proved that the probability $P(a \leq X \leq b)$, being the probability that the continuous random variable X takes on a value between a and b, satisfies

$$P(a \leq X \leq b) = \text{the area under the graph of the density function } f(x) \text{ between points } a \text{ and } b$$

for any real numbers a and b with $a < b$ when $f(x)$ is the probability density function of X. Readers who are familiar with integral calculus will recognize the area under the graph of $f(x)$ between a and b as the integral of $f(x)$ from a to b. Mathematically

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Any introductory course in integral calculus shows that the area under the graph of $f(x)$ between a and b can be approximated through the sum of the areas of small rectangles by dividing the interval $[a, b]$ into narrow subintervals of equal width. In particular, the area under the graph of $f(x)$ between the points $v - \frac{1}{2}\Delta$ and $v + \frac{1}{2}\Delta$ is approximately equal to $f(v)\Delta$ when Δ is small enough. In other words, $f(v)\Delta$ is approximately equal to the probability that the random variable X takes on a value in a small interval around v of width Δ. In view of this meaning, it is reasonable to define the *expected value* of a continuous random variable X by

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

This definition parallels the definition $E(X) = \sum_x x P(X = x)$ for a discrete random variable X.

5.1.2 Normal density function

A continuous random variable X is said to have a *normal distribution* with parameters μ and $\sigma > 0$ if

$$P(a \leq X \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}\,dx$$

for any real numbers a and b with $a \leq b$. The corresponding normal density function is given by

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

The notation "X is $N(\mu, \sigma^2)$" is often used as a shorthand for "X is normally distributed with parameters μ and σ." Theoretically, a normally distributed random variable has the whole real line as its range of possible values. However, a normal distribution can also be used for a nonnegative random variable provided that the normal distribution assigns a negligible probability to the negative axis. In Chapter 14, it will be shown for an $N(\mu, \sigma^2)$ random variable X that

$$E(X) = \mu \quad \text{and} \quad E[(X - \mu)^2] = \sigma^2.$$

Thus, the parameter μ gives the expected value of X and the parameter σ gives an indication of the spread of the random variable X around its expected value. The parameter σ is the so-called standard deviation of X. The concept of standard deviation will be discussed in more detail in Section 5.2.

An important result is:

if a random variable X is normally distributed with parameters μ and σ, then for each two constants $a \neq 0$ and b the random variable $U = aX + b$ is normally distributed with parameters $a\mu + b$ and $|a|\sigma$.

This result states that any linear combination of a normally distributed random variable X is again normally distributed. In particular, the random variable

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

is normally distributed with parameters 0 and 1. A normally distributed random variable Z with parameters 0 and 1 is said to have a *standard normal* distribution. The shorthand notation Z is $N(0, 1)$ is often used. The special notation

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}x^2}\, dx$$

is used for the cumulative probability distribution function $P(Z \leq z)$. The derivative of $\Phi(z)$ is the standard normal density function and is given by

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

The quantity $\Phi(z)$ gives the area under the graph of the standard normal density function left from the point $x = z$. No closed form of the cumulative distribution function $\Phi(z)$ exists. In terms of calculations, the integral for $\Phi(z)$ looks terrifying, but mathematicians have shown that the integral can be approximated with extreme precision by the quotient of two suitably chosen polynomials. This means that in practice the calculation of $\Phi(x)$ for a given value of z presents no difficulties at all and can be accomplished very quickly.

All calculations for an $N(\mu, \sigma^2)$-distributed random variable X can be reduced to calculations for the $N(0, 1)$-distributed random variable Z by using the linear transformation $Z = (X - \mu)/\sigma$. Writing $P(X \leq a) = P((X - \mu)/\sigma \leq (a - \mu)/\sigma)$ and noting that $\Phi(z) = P(Z \leq z)$, it follows that

$$P(X \leq a) = \Phi\left(\frac{a - \mu}{\sigma}\right).$$

An extremely useful result is the following:

> **the probability that a normally distributed random variable will take on a value that lies z or more standard deviations above the expected value is equal to $1 - \Phi(z)$ for $z > 0$, as is the probability of a value that lies z or more standard deviations below the expected value.**

This important result is the basis for a rule of thumb that is much used in statistics when testing hypotheses (see Section 5.6). The proof of the result is easy. Letting Z denote the standard normal random variable, it holds that

$$\begin{aligned} P\left(X \geq \mu + z\sigma\right) &= P\left(\frac{X - \mu}{\sigma} \geq z\right) = P(Z \geq z) = 1 - P(Z < z) \\ &= 1 - \Phi(z). \end{aligned}$$

The reader should note that $P(Z < z) = P(Z \leq z)$, because Z is a continuous random variable and so $P(Z = z) = 0$ for any value of z. Since the graph of the normal density function of X is symmetric around $x = \mu$, the area under this graph left from the point $\mu - z\sigma$ is equal to the area under the graph right from the point $\mu + z\sigma$. In other words, $P\left(X \leq \mu - z\sigma\right) = P\left(X \geq \mu + z\sigma\right)$. This completes the proof.

5.1.3 Percentiles

In applications of the normal distribution, percentiles are often used. For a fixed number p with $0 < p < 1$, the $100p\%$ *percentile* of a normally distributed random variable X is defined as the number x_p for which

$$P(X \leq x_p) = p.$$

In other words, the area under the graph of the normal density function of X left from the percentile point x_p is equal to p. The percentiles of the $N(\mu, \sigma^2)$ distribution can be expressed in terms of the percentiles of the $N(0, 1)$ distribution. The $100p\%$ percentile of the standard normal distribution is denoted as z_p and is thus the solution of the equation

$$\Phi(z_p) = p.$$

It is enough to tabulate the percentiles of the standard normal distribution. If the random variable X has an $N(\mu, \sigma^2)$ distribution, then it follows from

$$P(X \leq x_p) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x_p-\mu}{\sigma}\right) = \Phi\left(\frac{x_p-\mu}{\sigma}\right)$$

that its $100p\%$ percentile x_p satisfies $(x_p - \mu)/\sigma = z_p$. Hence

$$x_p = \mu + \sigma z_p.$$

A much used percentile of the standard normal distribution is the 95% percentile

$$z_{0.95} = 1.6449.$$

Let's illustrate the use of percentiles by means of the following example: of the people calling in for travel information, how long do 95% of them spend on the line with an agent when the length of a telephone call is normally distributed with an expected value of four minutes and a standard deviation of half a minute? The 95% percentile of the call-conclusion time is $4 + 0.5 \times z_{0.95} = 4.82$ minutes. In other words, on average 95% of the calls are concluded within 4.82 minutes.

In inventory control, the normal distribution is often used to model the demand distribution. Occasionally, one finds oneself asking experts in the field for educated guesses with regard to the expected value and standard deviation of the normal demand distribution. But even such experts often have difficulty with the concept of standard deviation. They can, however, provide an estimate (educated guess) for the average demand, and they can usually even estimate the threshold level of demand that will only be exceeded with a 5% chance, say. Let's say you receive an estimated value of 75 for this threshold, against an estimated value of 50 for the average level of demand. From this, you can immediately derive what the expected value μ and the standard deviation σ of the normal demand distribution are. Obviously, the expected value μ is 50. The standard deviation σ follows from the relationship $x_p = \mu + \sigma z_p$ with $x_p = 75$ and $z_p = 1.6449$. This gives $\sigma = 15.2$. The same idea of estimating μ and σ through an indirect approach may be useful in financial analysis. Let the random variable X represent the price of a stock next year. Suppose that an investor expresses his/her belief in the future stock price by assessing that there is a 25% probability of a stock price being below \$80 and a 25% probability of a stock price being above \$120. Estimates for the expected value μ and standard deviation σ of the stock price next year are then obtained from the equations $(80 - \mu)/\sigma = z_{0.25}$ and $(120 - \mu)/\sigma = z_{0.75}$, where $z_{0.25} = -0.67449$ and $z_{0.75} = 0.67449$. This leads to $\mu = 100$ and $\sigma = 5.45$.

5.2 The concept of standard deviation

The expected value of a random variable X is an important feature of this variable. Say, for instance, that the random variable X represents the winnings in a certain game. The law of large numbers teaches us that the average win per game will be equal to $E(X)$ when a very large number of independent repetitions are completed. However, the expected value reveals little about the value of X in any one particular game. To illustrate, say that the random variable X takes on the two values 0 and 5,000 with corresponding probabilities 0.9 and 0.1. The expected value of the random variable is then 500, but this value tells us nothing about the value of X in any one game. The following example shows the danger of relying merely on average values in situations involving uncertainty.

Pitfalls for averages[†]

A retired gentleman would like to place \$100,000 in an investment fund in order to ensure funding for a variety of purposes over the coming 20 years. How much will he be able to draw out of the account at the end of each year such that the initial investment capital, which must remain in the fund for 20 years, will not be disturbed? In order to research this issue, our man contacts Legio Risk, a well-known investment fund corporation. The advisor with whom he speaks tells him that the average rate of return has been 14% for the past 20 years (the one-year *rate of return* on a risky asset is defined as the beginning price of the asset minus the end price divided by the beginning price). The advisor shows him that with a *fixed* yearly return of 14%, he could withdraw \$15,098 at the end of each of the coming 20 years given an initial investment sum of \$100,000 (one can arrive at this sum by solving x from the equation $(1+r)^{20}A - \sum_{k=0}^{19}(1+r)^k x = 0$ yielding $x = [r(1+r)^{20}A]/[(1+r)^{20}-1]$ with $A = \$100{,}000$ and $r = 0.14$). This is music to the ears of our retired friend, and he decides to invest \$100,000 in the fund. His wife does not share his enthusiasm for the project, and cites Roman philosopher and statesman Pliny the Elder to support her case: the only certainty is that nothing is certain. Her husband ignores her concerns and says that there will be nothing to worry about so long as the average value of the rate of return remains at 14%. Can our retiree count on a yearly payoff of \$15,098 at the end of each of the coming 20 years if the rate of return fluctuates, from year to year, around 14% such

[†] This example is borrowed from Sam Savage in his article, "The flaw of averages," October 8, 2000 in the *San Jose Mercury News*.

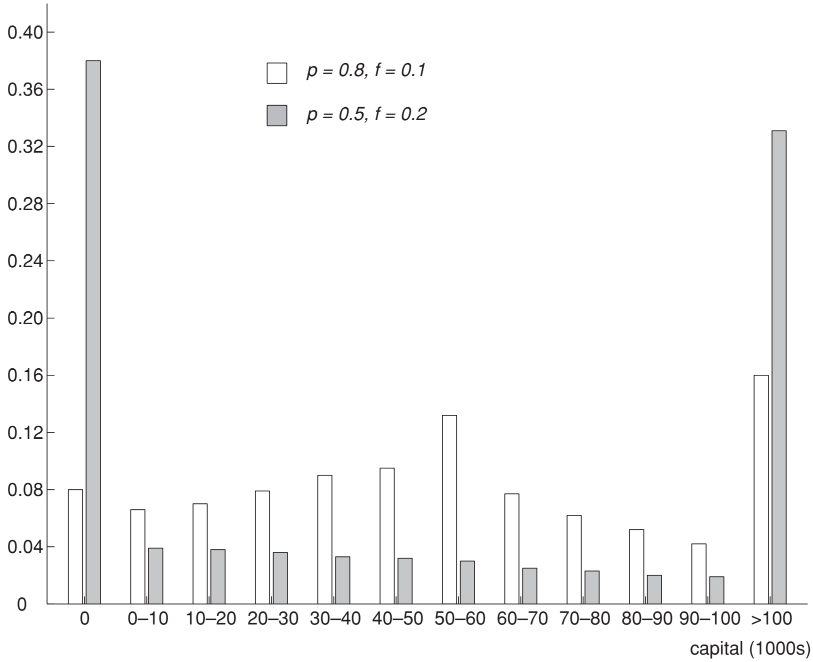


Fig. 5.4. Distribution of invested capital after 15 years.

that the average rate of return really is 14%? The answer is a resounding no! In this case, there is a relatively high chance of the capital being used up before the 20-year term is over (on the other hand, there is also a chance that after 20 years a hefty portion of the initial investment will still be left). In situations of uncertainty you cannot depend on average values. Statisticians like to tell the story of the man who begins walking across a particular lake, having ascertained beforehand that it is, on average, 30 centimeters deep. Suddenly, he encounters an area where the lake is approximately 3 meters deep, and, a nonswimmer, he falls in and drowns. In Figure 5.4, we illustrate the consequences of uncertainty, using a simple probability model for the development of the rate of return. If last year the rate of return was $r\%$, then over the course of the coming year the rate of return will be $r\%$, $(1 + f)r\%$, or $(1 - f)r\%$ with respective probabilities p, $\frac{1}{2}(1 - p)$, and $\frac{1}{2}(1 - p)$. In this model the expected value of the rate of return is the same for every year and is equal to the initial value of the rate of return. When we assume an initial investment capital of \$100,000 and the knowledge of a 14% rate of return for the year preceding the initial investment, we arrive at the data presented in Figure 5.4. This figure displays the distribution of the invested capital after 15 years, when at the end of each of those

15 years a total of \$15,098 is withdrawn from the fund (when there is less than \$15,098 in the fund, the entire amount remaining is withdrawn). The nonshaded distribution corresponds to $p = 0.8$ and $f = 0.1$, and the shaded distribution corresponds to $p = 0.5$ and $f = 0.2$. These distributions are calculated by simulation (4 million runs). The nonshaded distribution is less spread out than the shaded distribution. Can you explain this? In addition, the simulation reveals surprising pattern similarities between the random walk describing the course of the invested capital at the end of each year and the random walk describing the difference between the number of times heads comes up and the number of times tails comes up in the experiment of the recurring coin toss (see Problem 5.11 and the arc-sine law in Section 2.1). The fair coin is a familiar figure in the world of finance!

5.2.1 Variance and standard deviation

Let X be any random variable with expected value

$$\mu = E(X).$$

A measure of the spread of the random variable X around the expected value μ is the variance. The variance of X is defined as the expected value of the random variable $(X - \mu)^2$ and is denoted by $\sigma^2(X)$. That is

$$\sigma^2(X) = E[(X - \mu)^2].$$

Another common notation for the variance of X is var(X).[†] Why not use $E(|X - \mu|)$ as the measuring gauge for the spread? The answer is simply that it is much easier to work with $E[(X - \mu)^2]$ than with $E(|X - \mu|)$. The variance $\sigma^2(X) = E[(X - \mu)^2]$ can also be seen in the famed Chebyshev's inequality

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2(X)}{a^2}$$

for every constant $a > 0$. This inequality is generally applicable regardless of what form the distribution of X takes. It can even be sharpened to

$$P(X > \mu + a) \leq \frac{\sigma^2(X)}{\sigma^2(X) + a^2} \quad \text{and} \quad P(X < \mu - a) \leq \frac{\sigma^2(X)}{\sigma^2(X) + a^2}$$

[†] How do you compute $E[(X - \mu)^2]$? Let's assume for simplicity that X is a discrete random variable with I as its set of possible values. Then, you can use the generally valid formula $E[(X - \mu)^2] = \sum_{x \in I} (x - \mu)^2 P(X = x)$. This formula is a special case of the substitution rule that will be discussed in Chapter 9. For example, let X be the score of a single roll of one die. Then, $P(X = i) = \frac{1}{6}$ for $i = 1, \ldots, 6$ and so $\mu = E(X) = \sum_{i=1}^{6} i P(X = i) = 3.5$ and $E[(X - \mu)^2] = \sum_{i=1}^{6} (i - 3.5)^2 P(X = i) = 2.917$.

for every constant $a > 0$. This one-sided version of Chebyshev's inequality is a practical and useful result. In practical situations, it commonly occurs that only the expected value $E(X)$ and the variance $\sigma^2(X)$ of the distribution of X are known. In such situations, you can still establish an upper limit for a probability of the form $P(X > \mu + a)$ or $P(X < \mu - a)$. For example, imagine that X is the return on a certain investment and that you only know that the return has an expected value of 100 and a variance of 150. In this case, the probability of the return X taking on a value less than 80 will always be capped off at $150/(150 + 20^2) = 0.273$, regardless of what the distribution of X is.

The variance $\sigma^2(X)$ does not have the same dimension as the values of the random variable X. For example, if the values of X are expressed in dollars, then the dimension of $\sigma^2(X)$ will be equal to $(\text{dollars})^2$. A measure for the spread that has the same dimension as the random variable X is the *standard deviation.* It is defined as

$$\sigma(X) = \sqrt{E[(X - \mu)^2]}.$$

Referring back to the distribution in Figure 5.4, the nonshaded distribution corresponding to the case of $p = 0.8$ and $f = 0.1$ has an expected value of approximately \$58,000 and a standard deviation of approximately \$47,000, whereas the shaded distribution corresponding to the case of $p = 0.5$ and $f = 0.2$ has an expected value of approximately \$142,000 and a standard deviation of approximately \$366,000. The results for the case of $(p = 0.5,\ f = 0.2)$ are quite surprising and go against intuitive thinking! The explanation lies in the sharp movement of the yearly rate of return. This comes out in a standard deviation of the capital after 15 years that is relatively large with regard to the expected value (so we get, for example, a strong 6% probability of an invested capital of more than \$500,000 after 15 years lining up right next to a probability of 38% that the capital will be depleted after 15 years).

In the field of investment, smaller standard deviations are considered to be highly preferable when the expected value remains stable. Nevertheless, it is not always wise to base decisions on expected value and standard deviation alone. Distributions having the same expected value and the same standard deviation may display strong differences in the tails of the distributions. We illustrate this in the following example: investment A has a 0.8 probability of a \$2,000 profit and a 0.2 probability of a \$3,000 loss. Investment B has a 0.2 probability of a \$5,000 profit and a 0.8 probability of a zero profit. The net profit is denoted by the random variable X for investment A and by the random variable Y for investment B. Then

$$E(X) = 2{,}000 \times 0.8 - 3{,}000 \times 0.2 = \$1{,}000$$

and

$$\sigma(X) = \sqrt{(2{,}000 - 1{,}000)^2 \times 0.8 + (-3{,}000 - 1{,}000)^2 \times 0.2} = \$2{,}000.$$

Similarly, $E(Y) = \$1{,}000$ and $\sigma(Y) = \$2{,}000$ (verify!). Hence both investments have the same expected value and the same standard deviation for the net profit. In this situation, it is important to know the entire distribution in order to choose wisely between the two investments.

We will now present a number of properties of the standard deviation.

Property 1. *For every two constants a and b*

$$\sigma^2\,(aX + b) = a^2\sigma^2(X).$$

This property comes as the result of applying the definition of variance and using the fact that the expected value of a sum is the sum of the expected values, we leave its derivation to the reader. To illustrate Property 1, let's say that an investor has a portfolio half of which is made up of liquidities and half of equities. The liquidities show a fixed 4% return. The equities show an uncertain return with an expected value of 10% and a standard deviation of 25%. The return on the portfolio, then, has an expected value of $\frac{1}{2} \times 4\% + \frac{1}{2} \times 10\% = 7\%$ and a standard deviation of $\sqrt{\frac{1}{4} \times 625\%} = 12.5\%$.

In contrast to expected value, it is not always the case with variance that the variance of the sum of two random variables is equal to the sum of the variances of the two individual random variables. In order to give a formula for the variance of the sum of two random variables, we need the concept of covariance. The *covariance* of two random variables X and Y is denoted and defined by

$$\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))].$$

The value of $\text{cov}(X, Y)$ gives an indication of how closely connected the random variables X and Y are. If random variable Y tends to take on values smaller (larger) than $E(Y)$ whenever X takes on values larger (smaller) than $E(X)$, then $\text{cov}(X, Y)$ will usually be negative. Conversely, if the random variables X and Y tend to take on values on the same side of $E(X)$ and $E(Y)$, then $\text{cov}(X, Y)$ will usually be positive. Two random variables X and Y are said to be *positively (negatively) correlated* if the covariance has a positive (negative) value.[†] We can now state Property 2 whose proof can be found in Chapter 11.

[†] The correlation coefficient of two random variables X and Y is defined by $\rho(X, Y) = \frac{\text{cov}(X,Y)}{\sigma(X)\sigma(Y)}$. This is a dimensionless quantity with $-1 \leq \rho(X, Y) \leq 1$ (see Chapter 11). The correlation coefficient measures how strongly X and Y are correlated. The farther $\rho(X, Y)$ is from 0, the stronger the correlation between X and Y.

Property 2. *For every two random variables X and Y,*

$$\sigma^2(X+Y)=\sigma^2(X)+\sigma^2(Y)+2\,\mathrm{cov}(X,Y).$$

5.2.2 Independent random variables

Random variables X and Y are said to be *uncorrelated* if $\mathrm{cov}(X,Y)=0$. In Chapter 11, it will be shown that a sufficient (but not necessary) condition for uncorrelatedness of two random variables X and Y is that X and Y are independent random variables. The concept of independent random variables is very important. Intuitively, two random variables X and Y are independent if learning that Y has taken on the value y gives no additional information about the value that X will take on and, conversely, learning that X has taken on the value x gives no additional information about the value that Y will take on. In the experiment of throwing two dice, the two random variables giving the number of points shown by the first die and the second die are independent, but the two random variables giving the largest and the smallest number shown are dependent. Formally, independence of random variables is defined in terms of independence of events. Two random variables X and Y are said to be *independent* if the event of X taking on a value less than or equal to a and the event of Y taking on a value less than or equal to b are independent for all possible values of a and b. The independence of two events A and B is defined by $P(AB)=P(A)P(B)$. It can be shown that $\mathrm{cov}(X,Y)=0$ if X and Y are independent (see Chapter 11). Thus, Property 2 implies that

$$\sigma^2(X+Y)=\sigma^2(X)+\sigma^2(Y)\qquad\text{for independent } X \text{ and } Y.$$

5.2.3 Illustration: investment risks

Property 2 quantifies an important fact that investment experience supports: spreading investments over a variety of funds (diversification) diminishes risk. To illustrate, imagine that the random variable X is the return on every invested dollar in a local fund, and random variable Y is the return on every invested dollar in a foreign fund. Assume that random variables X and Y are independent and both have a normal distribution with expected value 0.15 and standard deviation 0.12. If you invest all of your money in either the local or the foreign fund, the probability of a negative return on your investment is equal to the probability that a normally distributed random variable takes on a value that is $\frac{0.15}{0.12}=1.25$ standard deviations below the expected value. This probability is equal to 0.106. Now imagine that your money is equally distributed over the two funds. Then, the expected return remains at 15%, but the probability of a negative return falls from 10.6% to 3.9%. To explain this, we need the fact that the sum of two independent *normally* distributed random variables is *normally* distributed (see Chapter 14). This means that $\frac{1}{2}(X + Y)$ is normally distributed with expected value 0.15 and standard deviation

$$\sqrt{\frac{1}{4}(0.12)^2 + \frac{1}{4}(0.12)^2} = \frac{0.12}{\sqrt{2}} = 0.0849.$$

The probability that a normally distributed random variable takes on a value that is $\frac{0.15}{0.0849} = 1.768$ standard deviations below the expected value is equal to 0.039. By distributing your money equally over the two funds, you reduce your downward risk, but you also reduce the probability of doubling your expected return (this probability also falls from 10.6% to 3.9%). In comparison with the distributions of X and Y, the probability mass of $\frac{1}{2}(X + Y)$ is concentrated more around the expected value and less at the far ends of the distribution. The centralization of the distribution as random variables are averaged together is a manifestation of the central limit theorem.

The example is based on the assumption that returns X and Y are independent from each other. In the world of investment, however, risks are more commonly reduced by combining *negatively correlated* funds (two funds are negatively correlated when one tends to go up as the other falls). This becomes clear when one considers the following hypothetical situation. Suppose that two stock market outcomes ω_1 and ω_2 are possible, and that each outcome will occur with a probability of $\frac{1}{2}$. Assume that domestic and foreign fund returns X and Y are determined by $X(\omega_1) = Y(\omega_2) = 0.25$ and $X(\omega_2) = Y(\omega_1) = -0.10$. Each of the two funds then has an expected return of 7.5%, with equal probability for actual returns of 25% and −10%. The random variable $Z = \frac{1}{2}(X + Y)$ satisfies $Z(\omega_1) = Z(\omega_2) = 0.075$. In other words, Z is equal to 0.075 with certainty. This means that an investment that is equally divided between the domestic and foreign funds has a guaranteed return of 7.5%.

We conclude this section with another example showing that you cannot always rely on averages only.

5.2.4 Waiting-time paradox[†]

You are in Manhattan for the first time. Having no prior knowledge of the bus schedules, you happen upon a bus stop located on Fifth Avenue. According to the timetable posted, buses are scheduled to run at ten-minute intervals. So, having reckoned on a waiting period of five minutes, you are dismayed to find

[†] This section is highly specialized and may be skipped over.

that after waiting for more than 20, there is still no bus in sight. The following day you encounter a similar problem at another busy spot in the city. How is this possible? Is it just bad luck? No, you have merely encountered the bus waiting paradox: when arrival/departure times at the various stops cannot be strictly governed (due to traffic problems, for example), then a person arriving randomly at a bus stop may wind up waiting longer than the average time scheduled between the arrival of two consecutive buses! It is only when buses run *precisely* at ten-minute intervals that the average wait will equal the expected five-minute period. We can elucidate the waiting-time paradox further by looking at a purely fictional example. Suppose that buses run at 30-minute intervals with a probability of 20%, and at one-second intervals with a probability of 80%. The average running time, then, should be six minutes, but the average waiting period for the person arriving randomly at the bus stop is approximately 15 minutes! The paradox can be explained by the fact that one has a higher probability of arriving at the bus stop during a long waiting interval than during a short one. A simple mathematical formula handsomely shows the effect of variability in running times on the average wait for a person turning up randomly at a bus stop. This formula involves the concept of coefficient of variation of a random variable. The coefficient of variation is the ratio of the standard deviation and the expected value. If the random variable T represents the amount of time elapsing between two consecutive buses, then the *coefficient of variation* of T is denoted and defined by

$$c_T = \frac{\sigma(T)}{E(T)}.$$

The coefficient of variation is dimensionless and is often a better measure for variability than the standard deviation (a large value of the standard deviation does not necessarily imply much variability when the expected value is large as well). Supposing that buses run at independent intervals that are distributed as the random variable T, it can be proved that

$$\frac{1}{2}\left(1 + c_T^2\right)E(T)$$

gives the average time a person must wait for a bus if the person arrives at the bus stop at a random point in time. If the buses run *precisely* on schedule ($c_T = 0$), then the average wait period is equal to $\frac{1}{2}E(T)$ as may be expected. Otherwise, the average wait period is always larger than $\frac{1}{2}E(T)$. The average wait period is even larger than $E(T)$ if the interstop running time T has $c_T > 1$!

Variability is also the reason why a small increase in demand for an already busy cash register at the supermarket leads to a disproportionately large increase in the queue for that cashier. This is nicely explained by the

Pollaczek-Khintchine formula from queueing theory

$$L_q = \frac{1}{2}\left(1 + c_S^2\right)\frac{[\lambda E(S)]^2}{1 - \lambda E(S)}.$$

This formula refers to the situation in which customers arrive at a service facility according to a Poisson process with intensity λ (the Poisson process was discussed in Section 4.2.4). The service times of the customers are independent of each other and have an expected value of $E(S)$ and a coefficient of variation of c_S. There is a single server who can handle only one customer at a time. Assuming that the average number of arrivals during a service time is less than 1, it can be shown that the long-run average number of customers waiting in the queue is given by the Pollaczek-Khintchine formula for L_q. This formula clearly shows the danger of increasing the load on a highly loaded system. Normalizing the average service time as $E(S) = 1$ and assuming a highly loaded system with $\lambda = 0.9$, then a 5% increase in the arrival rate λ leads to a 100.5% increase in the average queue size. In stochastic service systems, one should never try to balance the input with the service capacity of the system! This is an important lesson from the Pollaczek-Khintchine formula.

5.3 The square-root law

This section deals with a sequence $X_1, X_2, \ldots, X_n$ of independent random variables each having the same probability distribution with standard deviation σ. Letting X be a random variable defined on the sample space of a chance experiment, it is helpful to think of $X_1, X_2, \ldots, X_n$ as the representatives of X in n independent repetitions of the experiment. A repeated application of the formula $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$ for two independent random variables X and Y gives

$$\sigma^2(X_1 + \cdots + X_n) = \sigma^2(X_1) + \cdots + \sigma^2(X_n) = n\sigma^2.$$

Consequently,

Property 3. *For each $n \geq 1$*

$$\sigma\left(X_1 + X_2 + \cdots + X_n\right) = \sigma\sqrt{n}.$$

Properties 1 and 3 make it immediately apparent that

Property 4. *For every $n \geq 1$*

$$\sigma\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{\sigma}{\sqrt{n}}.$$

This property is called *the square-root law*. In finance, diversification of a portfolio of stocks generally achieves a reduction in the overall risk exposure with no reduction in expected return. Suppose that you split an investment budget equally between n similar but independent funds instead of concentrating it all in only one. Then, Property 4 states that the standard deviation of the rate of return falls by a factor $1/\sqrt{n}$ in comparison with the situation that the full budget is invested in a single fund. Insurance works according to the same mechanism.

The sample mean of $X_1, X_2, \ldots, X_n$ is denoted and defined by

$$\overline{X}(n) = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

We know already, based on the law of large numbers, that the sample mean becomes more and more concentrated around the expected value $\mu = E(X)$ as n increases. The square-root law specifies further that

> **the standard deviation of the sample mean $\overline{X}(n)$ is proportional to $\frac{1}{\sqrt{n}}$ when n is the sample size.**

In other words, in order to reduce the standard deviation of the sample mean *by half*, a sample size *four times as large* is required. The central limit theorem to be discussed in the next section specifies precisely how the probability mass of the sample mean $\overline{X}(n)$ is distributed around the expected value $\mu = E(X)$ when the sample size n is large.

In Figure 5.5, we give an experimental demonstration of the square-root law. A standard normal distribution is taken for the underlying random variable X. For each of the respective sample sizes $n = 1, 4, 16$, and 64, there are 100 outcomes of the sample average $\overline{X}(n)$ simulated. Figure 5.5 shows that the bandwidths within which the simulated outcomes lie are indeed reduced by an approximate factor of 2 when the sample sizes are increased by a factor of 4.

5.4 The central limit theorem

The central limit theorem is without a doubt the most important finding in the fields of probability theory and statistics. This theorem postulates that the sum (or average) of a sufficiently large number of independent random variables approximately follows a normal distribution. Suppose that $X_1, X_2, \ldots, X_n$ represents a sequence of independent random variables, each having the same distribution as the random variable X. Think of X as a random variable defined on the sample space of a chance experiment and think of $X_1, X_2, \ldots, X_n$ as

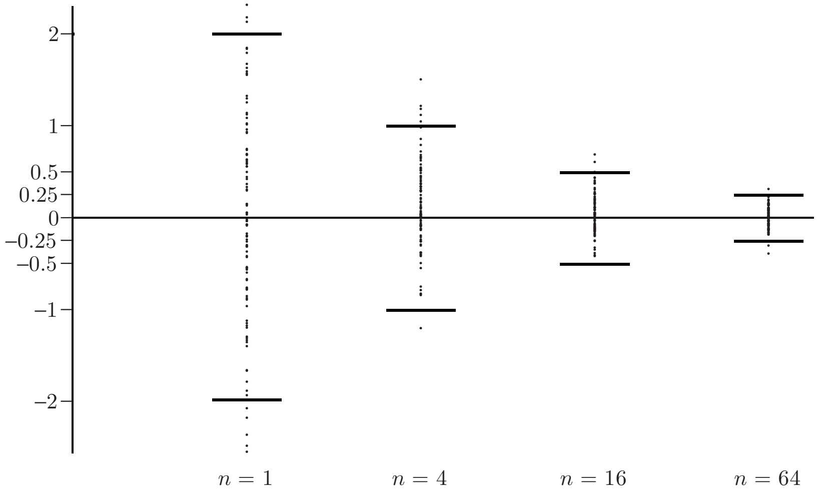


Fig. 5.5. Simulation outcomes for the sample mean.

the representatives of X in n independent repetitions of the experiment. The notation

$$\mu = E(X) \quad \text{and} \quad \sigma = \sigma(X)$$

is used for the expected value and the standard deviation of the random variable X. Mathematically, the central limit theorem states:

Central Limit Theorem. *For any real numbers a and b with $a < b$,*

$$\lim_{n\to\infty} P\left(a \leq \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \Phi(b) - \Phi(a),$$

where the standard normal distribution function $\Phi(x)$ is given by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2}\, dy.$$

Thus, the standardized variable $(X_1 + X_2 + \cdots + X_n - n\mu)/(\sigma\sqrt{n})$ has an approximately standard normal distribution. A mathematical proof of the central limit theorem will be outlined in Chapter 14. In Section 5.1, it was pointed out that $V = \alpha Z + \beta$ has a normal distribution with expected value β and standard deviation α when Z is $N(0, 1)$ distributed and α, β are constants with $\alpha > 0$. In ordinary words, we would be able to re-formulate the central limit theorem as follows:

> **if $X_1, X_2, \ldots, X_n$ are independent random variables each having the same distribution with expected value μ and standard deviation σ, then the sum $X_1 + X_2 + \cdots + X_n$ has an approximately normal distribution with expected value $n\mu$ and standard deviation $\sigma\sqrt{n}$ when n is sufficiently large.**

In terms of averaging random variables together, the central limit theorem tells us that:

> **if $X_1, X_2, \ldots, X_n$ are independent random variables each having the same distribution with expected value μ and standard deviation σ, then the sample mean $\overline{X}(n) = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ approximately has a normal distribution with expected value μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ when n is sufficiently large.**

This remarkable finding holds true no matter what form the distribution of the random variables X_k takes. How large n must be before the normal approximation is applicable depends, however, on the form of the underlying distribution of X_k. We return to this point in Section 5.5, where we show that it makes a big difference whether or not the probability mass of the underlying distribution is symmetrically accrued around the expected value.

In the central limit theorem it is essential that the random variables X_k are independent, but it is not necessary for them to have the same distribution. When the random variables X_k exhibit different distributions, the central limit theorem still holds true in general terms when we replace $n\mu$ and $\sigma\sqrt{n}$ with $\sum_{k=1}^{n} \mu_k$ and $(\sum_{k=1}^{n} \sigma_k^2)^{\frac{1}{2}}$, where $\mu_k = E(X_k)$ and $\sigma_k = \sigma(X_k)$. This generalized version of the central limit theorem elucidates the reason why, in practice, so many random phenomena, such as the rate of return on a stock, the cholesterol level of an adult male, the duration of a pregnancy, are approximately normally distributed. Each of these random quantities can be seen as the result of a large number of small independent random effects that add together.

The central limit theorem has an interesting history. The first version of this theorem was postulated by the French-born English mathematician Abraham de Moivre, who, in a remarkable article published in 1733, used the normal distribution to approximate the distribution of the number of heads resulting from many tosses of a fair coin. This finding was far ahead of its time, and was nearly forgotten until the famous French mathematician Pierre-Simon Laplace rescued it from obscurity in his monumental work *Théorie Analytique des Probabilités*, which was published in 1812. Laplace expanded De Moivre's finding by approximating the binomial distribution with the normal distribution. But as with De Moivre, Laplace's finding received little attention in his own time. It was not until the nineteenth century was at an end that the importance of the central limit theorem was discerned, when, in 1901, Russian mathematician Aleksandr Lyapunov defined it in general terms and proved precisely how it worked mathematically. Nowadays, the central limit theorem is considered to be the unofficial sovereign of probability theory.

5.4.1 Deviations

Do you believe a friend who claims to have tossed heads 5,250 times in 10,000 tosses of a fair coin? The central limit theorem provides an answer to this question.[†] For independent random variables $X_1, \ldots, X_n$, the central limit theorem points out how probable deviations of the sum $X_1 + X_2 + \cdots + X_n$ are from its expected value. The random variable $X_1 + \cdots + X_n$ is approximately normally distributed with expected value $n\mu$ and standard deviation $\sigma\sqrt{n}$. Also, as pointed out in Section 5.1, the probability of a normally distributed random variable taking on a value that is more than c standard deviations above or below the expected value is equal to $1 - \Phi(c) + 1 - \Phi(c) = 2\{1 - \Phi(c)\}$. Thus, for any constant $c > 0$

$$P(|(X_1 + X_2 + \cdots + X_n) - n\mu| > c\sigma\sqrt{n}) \approx 2\{1 - \Phi(c)\}$$

when n is sufficiently large. In particular

$$P(|(X_1 + \cdots + X_n) - n\mu| > c\sigma\sqrt{n}) \approx \begin{cases} 0.317 & \text{for } c = 1 \\ 0.046 & \text{for } c = 2 \\ 2.7 \times 10^{-3} & \text{for } c = 3 \\ 6.3 \times 10^{-5} & \text{for } c = 4 \\ 5.7 \times 10^{-7} & \text{for } c = 5. \end{cases}$$

Thus, the outcome of the sum $X_1 + \cdots + X_n$ will seldom be three or more standard deviations removed from the expected value $n\mu$. Coming back to the issue of whether or not the claim of having tossed 5,250 heads in 10,000 fair coin tosses is plausible, the answer is no. This cannot be explained as a chance variation. In order to justify this assertion, one should calculate the probability of 5,250 or more heads appearing in 10,000 tosses of a fair coin (and not the probability of exactly 5,250 heads). The number of times the coin lands on heads can be written as the sum $X_1 + X_2 + \cdots + X_{10{,}000}$, where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th toss turns heads} \\ 0 & \text{otherwise.} \end{cases}$$

[†] It appears that, for many students, there is a world of difference between a technical understanding of the central limit theorem and the ability to use it in solving a problem at hand.

Using the fact that $E(X_i) = 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{1}{2}$, the standard deviation of the Bernoulli variable X_i is

$$\sigma = \sqrt{\left(0 - \frac{1}{2}\right)^2 \times \frac{1}{2} + \left(1 - \frac{1}{2}\right)^2 \times \frac{1}{2}} = \frac{1}{2}.$$

Tossing 5,250 or more heads, then, lies

$$\frac{(5{,}250 - 5{,}000)}{\frac{1}{2}\sqrt{10{,}000}} = 5$$

or more standard deviations above the expected value of 5,000 heads. The chance of this happening is approximately 1 in 3.5 million. The claim of your friend is fakery! In the situation considered, you base your judgment on the probability of getting 5,250 or more heads in 10,000 tosses and not on the probability of getting exactly 5,250 heads. However, in the situation of a suspicious claim of 5,001 heads in 10,000 tosses of a fair coin, you calculate the probability that the absolute difference between the actual number of heads in 10,000 tosses and the expected number will be at most 1 (see also Problem 5.20 that deals with one of the most remarkable fakeries in the history of statistics).

5.5 Graphical illustration of the central limit theorem

The mathematical proof of the central limit theorem is far from simple and is also quite technical. Moreover, the proof gives no insight into the issue of how large n must actually be in order to get an approximate normal distribution for the sum $X_1 + X_2 + \cdots + X_n$. Insight into the working of the central limit theorem can best be acquired through empirical means. Simulation can be used to visualize the effect of adding random variables. For any fixed value of n, one runs many simulation trials for the sum $X_1 + X_2 + \cdots + X_n$ and creates a histogram by plotting the outcomes of the simulation runs. Then, for increasing values of n, it will be seen that the histogram approaches the famous bell-shaped curve. The disadvantage of this empirical approach is that the law of large numbers interferes with the central limit theorem. For a fixed value of n, one needs many simulation trials before the simulated distribution of $X_1 + X_2 + \cdots + X_n$ is sufficiently close to its actual distribution. This complication can be avoided by taking a different approach. In the case where the random variables $X_1, X_2, \ldots$ have a discrete distribution, it is fairly simple to calculate the probability mass function of the sum $X_1 + X_2 + \cdots + X_n$ exactly for any value of n. This can be done by using the convolution formula for the sum of discrete random variables. The convolution formula will be discussed in

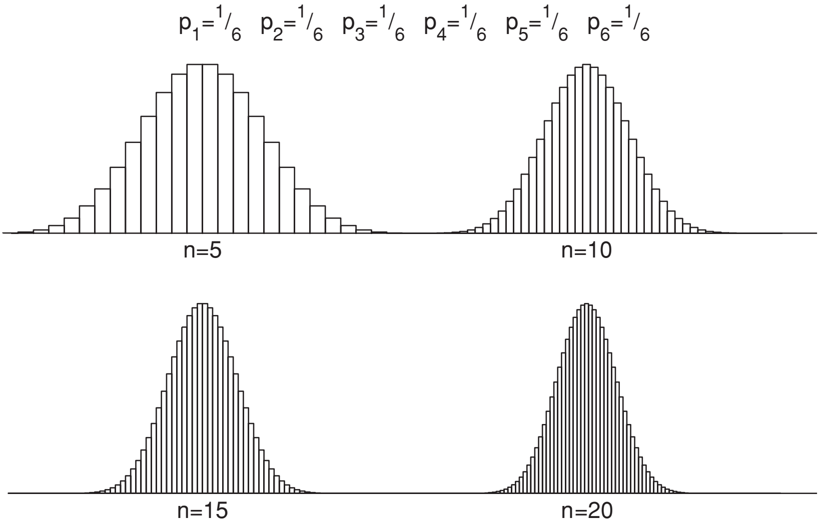


Fig. 5.6. Probability histogram for the unbiased die.

Chapter 9. In this way, you can determine empirically how large n must be in order to ensure that the probability histogram of the sum $X_1 + \cdots + X_n$ will take on the bell shape of the normal curve. You will see that the answer to the question of how large n must be strongly hinges on how "symmetrical" the probability mass of the random variable X_i is distributed around its expected value. The more *skewed* the probability mass is, the *larger n* must be in order for the sum of $X_1 + X_2 + \cdots + X_n$ to be approximately normally distributed. This can be nicely illustrated by using the chance experiment of rolling a (biased) die. Let's assume that one roll of the die turns up j points with a given probability p_j for $j = 1, \ldots, 6$. Playing with the probabilities p_j, one can construct both a symmetrical die and an asymmetrical die. Define the random variable X_k by

$$X_k = \text{number of points obtained by the } k\text{th roll of the die.}$$

The random variables $X_1, X_2, \ldots$ are independent and are distributed according to the probability mass function $(p_1, \ldots, p_6)$. The sum $X_1 + \cdots + X_n$ gives the total number of points that have been obtained in n rolls of the die. Figures 5.6 and 5.7 show the probability histogram of the sum $X_1 + \cdots + X_n$ for $n =$ 5, 10, 15, and 20 rolls of the die. This is shown for both the unbiased die with the symmetrical distribution $p_1 = \cdots = p_6 = \frac{1}{6}$ and a biased die with the asymmetrical distribution $p_1 = 0.2$, $p_2 = 0.1$, $p_3 = p_4 = 0$, $p_5 = 0.3$ and $p_6 = 0.4$. The two figures speak for themselves. It is quite apparent that for

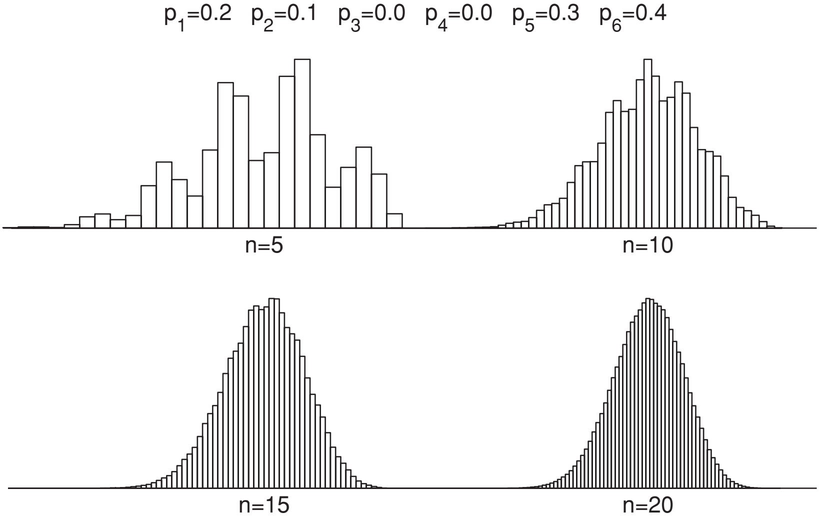


Fig. 5.7. Probability histogram for a biased die.

both distributions the diagram of $X_1 + \cdots + X_n$ ultimately takes on a normal bell-shaped curve, but that it occurs much earlier in the case of a symmetrical distribution than it does in the case of an asymmetrical distribution.

5.6 Statistical applications

The central limit theorem has numerous applications in probability theory and statistics. In this section, we discuss a few illustrative applications.

5.6.1 Normal approximation of the binomial distribution

Suppose that X is a binomially distributed random variable with parameters n and p. The random variable X can be interpreted as the total number of successes in n independent repetitions of a Bernoulli experiment with a success probability of p (see Section 4.1). Consequently, the random variable X can be represented as

$$X = I_1 + I_2 + \cdots + I_n,$$

where the indicator variable I_k is defined by

$$I_k = \begin{cases} 1 & \text{if the } k\text{th trial leads to a success} \\ 0 & \text{otherwise.} \end{cases}$$

It is left to the reader to verify that the Bernoulli variable I_k has an expected value of $\mu = p$ and a standard deviation of $\sigma = \sqrt{p(1-p)}$. The random variables $I_1, \ldots, I_n$ are independent. Using Property 2 from Section 5.2, it now follows that the expected value and the standard deviation of the binomially distributed random variable X are given by

$$E(X) = np \quad \text{and } \sigma(X) = [p(1-p)]^{1/2}\sqrt{n}.$$

The central limit theorem now tells us that the (discrete) binomial distribution can be approximated by the (continuous) normal distribution when n is large.[†] A guideline is to use the normal approximation when $np > 5$ and $n(1-p) > 5$, but of course this would depend on the accuracy required. The normal probability density function then has sufficient space to "unfurl" around the expected value np without too much of the probability mass falling into the negative axis.

Example 5.1 A student has passed a final exam by supplying correct answers for 26 out of 50 multiple-choice questions. For each question, there was a choice of three possible answers, of which only one was correct. The student claims not to have learned anything in the course and not to have studied for the exam, and says that his correct answers are the product of guesswork. Do you believe him?

Solution. This problem can be approached as follows: take as hypothesis that the student did guess at all the answers and calculate the probability of identifying 26 or more correct answers through guesswork. If this probability is below a threshold value you have chosen in advance, you judge that the student is bluffing. If all the answers are guessed at, then the number of correct answers can be seen as the number of successes in $n = 50$ independent trials of a Bernoulli experiment having a success probability of $p = \frac{1}{3}$. The binomial probability model is thus applicable. A generally useful method of determining whether 26 correct answers is exceptional is based on finding out how many standard deviations lie between the observed number of correct answers achieved and the expected number. To do so, a quick approach is to approximate the binomial distribution with parameters $n = 50$ and $p = \frac{1}{3}$ by a normal distribution with expected value $np = 16\frac{2}{3}$ and standard deviation $\sqrt{np(1-p)} = 3\frac{1}{3}$. Next, use the rule of thumb stating that the probability of a normally distributed random variable taking on a value lying three or more standard deviations above the expected value is very small (the probability is 0.0013). The observed value

[†] This approximation can be improved by using the so-called *continuity correction*: approximate $P(X > k)$ by $P(U \geq k + \frac{1}{2})$ and $P(X < k)$ by $P(U \leq k - \frac{1}{2})$, where U is a normal random variable with expected value np and standard deviation $[p(1-p)]^{1/2}\sqrt{n}$.

of 26 correct answers lies $(26 - 16\frac{2}{3})/3\frac{1}{3} = 2.8$ standard deviations above the expected value. The probability of such a deviation occurring is quite small. There is very good reason, therefore, to suppose that the student is bluffing, and that he in fact did prepare for the exam.

5.6.2 The z-value

The key to finding the solution to the problem in Example 5.1 is to measure the number of standard deviations separating the observed value from the expected value. The normal distribution will allow you to establish whether the difference between the observed value and the expected value can be explained as a chance variation or not. The z-value is defined as

$$z = \frac{\text{observed value} - \text{expected value}}{\text{standard deviation}}.$$

It is often used in the testing of hypotheses. This is illustrated with the famous example of the Salk vaccine.

Example 5.2 The Salk vaccine against polio was tested in 1954 in a carefully designed field experiment. Approximately 400,000 children took part in this experiment. Using a randomization procedure, the children were randomly divided into two groups of equal size, a treatment group and a control group. The vaccine was given only to the children in the treatment group; the control group children received placebo injections. The children did not know which of the two groups they had been placed into. The diagnosticians also lacked this information (double-blind experiment). Fifty-seven children in the treatment group went on to contract polio, while 142 children in the control group contracted the illness. Based on these results, how reliable is the claim that the vaccine worked?

Solution. This famous experiment is commonly misperceived. It is often claimed that such an experiment including the participation of 400,000 children cannot deliver reliable conclusions when those conclusions are based on the fact that two relatively small groups of 142 and 57 children contracted polio. People subscribing to this train of thought are misled by the magnitude of the group at large; what they should really be focusing on is the difference between the number of polio occurrences in each of the two groups as compared to the total number of polio occurrences. The test group must be large because statistically founded conclusions can only be drawn when a sufficiently large number of cases have been observed. Incidentally, since the probability of contracting polio is very small and the group sizes are very large, the realized number of polio occurrences in each group can be seen as an outcome of a Poisson distribution.

In order to find out whether the difference in outcomes between the two groups is a significant difference and not merely the result of a chance fluctuation, the following reasoning is used. Suppose that assignment to treatment or control had absolutely no effect on the outcome. Under this hypothesis, each of the 199 children was doomed to contract polio regardless of which group he/she was in. Now we have to ask ourselves this question: what is the probability that of the 199 affected children only 57 or less will belong to the treatment group? This problem strongly resembles the problem of determining the probability of not more than 57 heads turning up in 199 tosses of a fair coin. This problem can be solved with the binomial model with parameters $n = 199$ and $p = \frac{1}{2}$. This binomial model can be approximated by the normal model. For the z-value we find

$$z = \frac{57 - 199 \times 0.5}{\sqrt{199 \times 0.5 \times 0.5}} = -6.03.$$

Thus, the observed number of polio cases in the treatment group registers at more than six standard deviations below the expected number. The probability of this occurring in a normal distribution is on the order of 10^{-9}. It is therefore extremely unlikely that the difference in outcomes between the two groups can be explained as a chance variation. This in turn makes clear that the hypothesis is incorrect and that the vaccine does, in fact, work.

5.6.3 The z-value and the Poisson distribution

For many of the everyday situations of a statistical nature that occur, we only have averages available from which to draw conclusions. For example, records are kept of the average number of traffic accidents per year, the average number of bank robberies per year, etc. When working with this kind of information, the Poisson model is often suitable. The Poisson distribution is completely determined by its expected value. In Chapter 9, it will be shown that a Poisson-distributed random variable with expected value λ has a standard deviation of $\sqrt{\lambda}$. Also, for λ sufficiently large (say $\lambda \geq 25$), the Poisson distribution with expected value λ can be approximated by a normal distribution with expected value λ and standard deviation $\sqrt{\lambda}$. The explanation is that the Poisson distribution is a limiting case of the binomial distribution (see Chapter 4). In the beginning of this section, we saw that the binomial distribution can be approximated by the normal distribution. The normal approximation to the Poisson distribution and the concept of the z-value allow one to make statistical claims in situations such as those mentioned above. Suppose, for example, you read in the paper that, based on an average of 1,000 traffic deaths per year in previous years, the number of traffic deaths for last year rose 12%. How can you evaluate this? The number of traffic deaths over a period of one year can be modeled as a Poisson-distributed random variable with expected value 1,000 (why is this model reasonable?). An increase of 12% on an average of 1,000 is an increase of 120, or rather an increase of $120/\sqrt{1{,}000} = 3.8$ standard deviations above the expected value 1,000. The probability that a normally distributed random variable will take on a value of more than three standard deviations above the expected value is quite small. In this way, we find justification for the conclusion that the increase in the number of traffic deaths is not coincidental, but that something for which concrete explanations can be found has occurred. What would your conclusions have been if, based on an average of 100 traffic deaths per year, a total of 112 traffic deaths occurred in the past year?

5.7 Confidence intervals for simulations

In the preceding chapters we encountered several examples of simulation studies for stochastic systems. In these studies we obtained point estimates for unknown probabilities or unknown expected values. It will be seen in this more technical section that the central limit theorem enables us to give a probabilistic judgment about the accuracy of the point estimate. Simulation of a stochastic system is in fact a statistical experiment in which one or more unknown parameters of the system are estimated from a sequence of observations that are obtained from independent simulation runs of the system. Let's first consider the situation in which we wish to estimate the unknown expected value $\mu = E(X)$ of a random variable X defined for a given stochastic system (e.g., the expected value of the random time until a complex electronic system fails for the first time). Later on, when we encounter the estimating of probabilities, we will see that this turns out to be none other than a special case of estimating an expected value.

Let X be a random variable defined on the sample space of a chance experiment. The goal is to estimate the unknown expected value $\mu = E(X)$. Suppose that n independent repetitions of a chance experiment are performed. The kth performance of the experiment yields the representative X_k of the random variable X. An estimator for the unknown expected value $\mu = E(X)$ is given by the *sample mean*

$$\overline{X}(n) = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

It should be noted that this statistic, being the arithmetic mean of the random sample $X_1, \ldots, X_n$, is a random variable. The central limit theorem tells us that for n large

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

has an approximately standard normal distribution, where $\sigma = \sigma(X)$ is the standard deviation of the random variable X. Dividing the numerator and the denominator of the above expression by n, we find that

$$\frac{\overline{X}(n) - \mu}{\sigma/\sqrt{n}}$$

has an approximately standard normal distribution. For any number α with $0 < \alpha < 1$ the percentile $z_{1-\frac{1}{2}\alpha}$ is defined as the unique number for which the area under the standard normal curve between the points $-z_{1-\frac{1}{2}\alpha}$ and $z_{1-\frac{1}{2}\alpha}$ equals $100(1-\alpha)\%$. The percentile $z_{1-\frac{1}{2}\alpha}$ has the values 1.960 and 2.324 for the often-used values 0.05 and 0.01 for α. Since $[\overline{X}(n) - \mu]/(\sigma/\sqrt{n})$ is an approximately standard normal random variable, it follows that

$$P\left(-z_{1-\frac{1}{2}\alpha} \le \frac{\overline{X}(n) - \mu}{\sigma/\sqrt{n}} \le z_{1-\frac{1}{2}\alpha}\right) \approx 1 - \alpha$$

or, stated differently,

$$P\left(\overline{X}(n) - z_{1-\frac{1}{2}\alpha}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X}(n) + z_{1-\frac{1}{2}\alpha}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$

Voila! You have now delimited the unknown expected value μ on two ends. Both endpoints involve the standard deviation σ of the random variable X. In most situations σ will be unknown when the expected value μ is unknown, but fortunately, this problem is easily circumvented by replacing σ by an estimator based on the sample values $X_1, \ldots, X_n$. Just as the unknown expected value $\mu = E(X)$ is estimated by the sample mean $\overline{X}(n) = (1/n)\sum_{k=1}^{n} X_k$, the standard deviation σ is estimated by the square root of the *sample variance*. This statistic is denoted and defined by

$$S^2(n) = \frac{1}{n}\sum_{k=1}^{n}[X_k - \overline{X}(n)]^2.$$

The definition of the statistic $S^2(n)$ resembles the definition of the variance $\sigma^2(X) = E[(X-\mu)^2]$ (usually one defines $S^2(n)$ by dividing through $n-1$ rather than n, but for large n the two definitions of $S^2(n)$ boil down to the same thing). Using the law of large numbers, it can be shown that the statistic $S^2(n)$ converges to σ^2 as n tends to infinity.

The sample variance enables us to give a probability judgment about the quality or accuracy of the estimate $\overline{X}(n)$ for the unknown expected value $\mu = E(X)$. It can be proved that the central limit theorem remains valid when σ is replaced by its estimator $S(n)$. That is, for n large

$$P\left(-z_{1-\frac{1}{2}\alpha} \leq \frac{\overline{X}(n) - \mu}{S(n)/\sqrt{n}} \leq z_{1-\frac{1}{2}\alpha}\right) \approx 1 - \alpha$$

or, stated differently,

$$P\left(\overline{X}(n) - z_{1-\frac{1}{2}\alpha}\frac{S(n)}{\sqrt{n}} \leq \mu \leq \overline{X}(n) + z_{1-\frac{1}{2}\alpha}\frac{S(n)}{\sqrt{n}}\right) \approx 1 - \alpha.$$

This result is the basis for an interval estimate of the unknown parameter μ rather than a point estimate. Such an interval estimate is called a *confidence interval*. The following important result holds

> **for n large, an approximate $100(1-\alpha)\%$ confidence interval for the unknown expected value $\mu = E(X)$ is given by**
>
> $$\overline{X}(n) \pm z_{1-\frac{1}{2}\alpha}\frac{S(n)}{\sqrt{n}}.$$

When speaking about large n, it is better to think in terms of values of n on the order of tens of thousands than on the order of hundreds.[†] In practice, one often chooses $\alpha = 0.05$ and thus constructs a 95% confidence interval. The percentile $z_{1-\frac{1}{2}\alpha}$ is 1.960 for $\alpha = 0.05$.

When n independent simulation runs are performed to estimate the unknown expected value μ of the random variable X, then

$$\begin{aligned}&\text{the width of the approximate } 100(1-\alpha)\% \text{ confidence interval}\\ &= 2z_{1-\alpha/2}\frac{S(n)}{\sqrt{n}}\\ &= \frac{2z_{1-\alpha/2}}{\sqrt{n}} \times (\text{estimate for the unknown standard deviation of } X).\end{aligned}$$

The estimator $S(n)$ of the unknown standard deviation σ of X will not change much after some initial period of the simulation. This means that the width of the confidence interval is nearly proportional to $1/\sqrt{n}$ for n sufficiently large.

[†] In the special case of the random variables X_i themselves being normally distributed, it is possible to give a confidence interval that is not only exact but also applies to small values of n. This exact confidence interval is based on the so-called Student t-distribution instead of the standard normal distribution (see Chapter 10).

This conclusion leads to a practically important rule of thumb

> **to reduce the width of a confidence interval by a factor of two, about four times as many observations are needed.**

5.7.1 Interpretation of the confidence interval

Let's say we have determined by simulation a 95% confidence interval (25.5, 27.8) for an unknown expected value μ. In this case, we cannot actually say that there is a 95% probability of μ falling within the interval (25.5, 27.8). Why not? The reason is simply that the unknown μ is a constant and not a random variable. Thus, either the constant μ falls within the interval (25.5, 27.8) or it does not. In other words, the probability of μ falling within the interval (25.5, 27.8) is 1 or 0. If the values of the simulation data $X_1, \ldots, X_n$ had been different, the confidence interval would also have been different. Some simulation studies will produce confidence intervals that cover the true value of μ and others will not. Before the simulation runs are done, it can be said that the 95% confidence interval that will result will cover the true value of μ with a probability of 95%. After the data are obtained, it can only be said that "we are 95% confident that the resultant interval covers the true value of μ." A more concrete interpretation of the $100(1-\alpha)\%$ confidence interval is provided by the frequentist approach. If you construct a large number of $100(1-\alpha)\%$ confidence intervals, each based on the same number of simulation runs, then the proportion of intervals covering the unknown value of μ is approximately equal to $1-\alpha$. To illustrate this, consider Figure 5.8. This figure relates to a problem known as the newsboy problem and displays 100 95% confidence intervals for the expected value of the daily net profit of the newsboy. In this well-known inventory problem, a newsboy decides at the beginning of each day how many newspapers he will purchase for resale. Let's assume that the daily demand for newspapers is uniformly distributed between 150 and 250 papers. Demand on any given day is independent of demand on any other day. The purchase price per newspaper is one dollar. The resale price per newspaper is two dollars; the agency will buy back any unsold newspapers for fifty cents apiece. The performance measure we are interested in is the expected value μ of the daily net profit when at the beginning of each day the newsboy purchases 217 newspapers. We constructed 100 approximate 95% confidence intervals for μ by simulating the sales 100 times over $n = 2{,}000$ days. The resulting 95% confidence intervals for the expected value μ are given in Figure 5.8. It is instructive to have a look at the figure. Indeed, in approximately 95 of the 100 cases, the true value of μ is contained within the confidence interval (the true value of μ can analytically be shown to be equal to \$183.17).

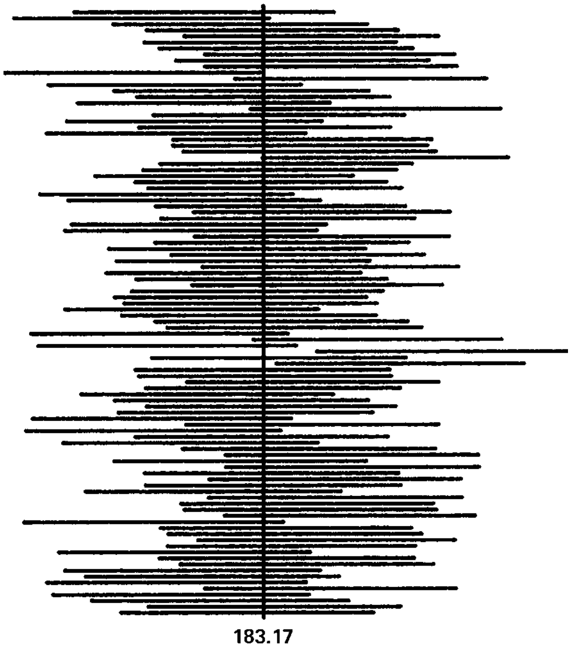


Fig. 5.8. One hundred 95% confidence intervals.

5.7.2 Confidence interval for a probability

The goal of many simulation studies is to estimate an unknown probability $P(E)$ for a given event E in a chance experiment. We shall demonstrate that the probability $P(E)$ can be seen as an expected value of an indicator variable. This implies that a confidence interval for a probability is a special case of a confidence interval for an expected value. Suppose that n independent repetitions of the experiment are simulated. Define the random variable X_i as

$$X_i = \begin{cases} 1 & \text{if event } E \text{ occurs in the } i\text{th trial} \\ 0 & \text{otherwise.} \end{cases}$$

The indicator variables $X_1, \ldots, X_n$ are independent Bernoulli variables each having the same distribution. Note that

$$E(X_i) = 0 \times P(X_i = 0) + 1 \times P(X_i = 1) = P(X_i = 1).$$

Since $P\,(X_i = 1) = P(E)$, it follows that the probability $P(E)$ is equal to the expected value of the indicator variables X_i. Thus, the sample mean

$$\overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

provides a point estimate for the unknown probability $P(E)$. The corresponding $100(1-\alpha)\%$ confidence interval $\overline{X}(n) \pm z_{1-\frac{1}{2}\alpha}\, S(n)/\sqrt{n}$ takes the insightful and simple form of

$$\overline{X}(n) \pm z_{1-\frac{1}{2}\alpha} \frac{\sqrt{\overline{X}(n)[1-\overline{X}(n)]}}{\sqrt{n}}.$$

The explanation is that $S^2(n) = \overline{X}(n)[1-\overline{X}(n)]$ for variables X_i that take on only the values 0 and 1. It is a matter of simple algebra to verify this fact. The simplified expression for $S^2(n)$ is in agreement with the fact that a Bernoulli variable X has variance $E(X)[1-E(X)]$. It follows from the structure of the confidence interval for an unknown probability that it suffices to know the sample mean in order to construct the confidence interval.

As an illustration, suppose that someone tells you that he/she simulated the so-called game of ace-jack-two 2,500 times and found the point estimate of 0.8092 for the probability of the player winning. Then you know enough to conclude that the half width of a 95% confidence interval for the probability of winning equals $1.96\sqrt{0.8092(1-0.8092)}/\sqrt{2{,}500} = 0.015$. The game of ace-jack-two is played this way: 17 times in a row, a player chooses three cards from a deck of 52 thoroughly shuffled cards. Every time the group of three cards contains an ace, jack, or two, the player accrues one point; otherwise the bank wins a point. An analytical calculation of the player's probability of accruing the most points and winning the game is far from easily achieved. That's why simulation has been used for this problem.

The confidence interval for an unknown probability also gives insight into the necessary simulation efforts for *extremely small* probabilities. Let's say that the unknown probability $p = P(E)$ is on the order of 10^{-6}. How large must the number of simulation runs be before the half width of the 95% confidence interval is smaller than $f \times 10^{-6}$ for a given value of f between 0 and 1? To answer this question, note that, for n large,

$$\sqrt{\overline{X}(n)[1-\overline{X}(n)]} \approx \sqrt{p(1-p)} \approx \sqrt{p}$$

because $1-p \approx 1$. The formula for the confidence interval for p now gives that the required number n of runs must satisfy $\frac{1.96\sqrt{p}}{\sqrt{n}} \approx f \times p$, or

$$n \approx \left(\frac{1.96}{f}\right)^2 \times \frac{1}{p}.$$

For $p = 10^{-6}$ and $f = 0.1$, this means approximately 400 million simulation runs. This shows how careful one must be when estimating extremely small probabilities with precision.

Example 5.3 A random sample of 1,000 voters is garnered from a population of 250,000 voters inhabiting a particular area. Interviews of the 1,000 voters in the sample group were conducted after which it was apparent that 520 of them voted Democrat in the last election. What is the 95% confidence interval for the fraction of Democrats among the voters in that area?

Solution. We can conceive of the process of interviewing 1,000 randomly chosen individuals as a simulation study with $n = 1{,}000$ independent trials of a Bernoulli experiment. Define the random variable X_i as

$$X_i = \begin{cases} 1 & \text{if the } i\text{th interviewee is Democrat} \\ 0 & \text{otherwise.} \end{cases}$$

The observed value of the sample mean $\overline{X}(n) = (1/n)\sum_{k=1}^{n} X_k$ is 520/1,000. Letting p represent the unknown fraction of Democrats among the voters, this fraction is estimated by the value

$$\overline{X}(n) = 0.52$$

with corresponding 95% confidence interval

$$\overline{X}(n) \pm 1.96\frac{\sqrt{\overline{X}(n)[1 - \overline{X}(n)]}}{\sqrt{n}} = 0.52 \pm 0.03.$$

Imagine that you are asked how large the sample size must be in order to get a 95% confidence interval with a margin of ± 0.01. The answer to this question tells us that you would then need a sample size of about 9,000 people. Increasing the sample size by a factor of 9 reduces the margin of the confidence interval with a factor of about 3.

Example 5.4 It is commonly presumed that an unborn child has a 50% probability of being female. But is this really the case? Let's take a look at birth statistics for the Netherlands for the years 1989, 1990, and 1991. According to the Central Bureau of Statistics, there were, in total, 585,609 children born during the span of those years, of which 286,114 were girls. What is the estimate for the probability that a newborn child will be a girl and what is the corresponding 95% confidence interval?

Solution. We can model this problem as $n = 585{,}609$ independent trials of a Bernoulli experiment with an unknown success probability of p, where success is defined as the birth of a girl. Let the random variable X_i be equal to 1 if the ith trial of the experiment delivers a success and let X_i be otherwise equal to 0. Then the unknown probability p is estimated by the value

$$\overline{X}(n) = \frac{286{,}114}{585{,}609} = 0.4886.$$

The corresponding 95% confidence interval is

$$\overline{X}(n) \pm 1.96 \frac{\sqrt{\overline{X}(n)[1-\overline{X}(n)]}}{\sqrt{n}} = 0.4886 \pm 0.0013.$$

In reality, then, the probability of a child being born female is slightly under 50% (the value 0.5 is also well outside the 99.99% confidence interval 0.4886 ± 0.0025). This probability appears to alter very little over time and applies to other countries as well. The celebrated French probability theorist Laplace, who also did much empirical research, investigated births over a long period in the eighteenth century and found that the probability of a newborn child being a girl had the value $\frac{21}{43} = 0.4884$ in each of the cities of Paris, London, Naples, and St. Petersburg. Interestingly, Laplace initially found a slightly deviating value for Paris, but in the end, after adjusting for the relatively large number of provincial girls placed in Parisian foundling homes, that probability was also reckoned at approximately $\frac{21}{43}$.

5.8 The central limit theorem and random walks

Random walks are among the most useful models in probability theory. They find applications in all parts of science. In Chapter 2, we introduced the random walk model based on the simple coin-tossing experiment and the random walk model of a gambler's fortune under the Kelly betting system. These elementary models uncover interesting, and occasionally profound, insights into the study of more complicated models. In this section, the central limit theorem will be used to reveal further properties of random walk models and to establish a link between random walks and the Brownian motion process. The Brownian motion process is widely applied to the modeling of financial markets. In particular, the famous Black-Scholes formula for the pricing of options will be discussed.

5.8.1 Fluctuations in a random walk

The central limit theorem allows us to make a mathematical statement about the behavior of the random walk in Figure 2.1. This random walk describes the evolution of the actual number of heads tossed minus the expected number when a fair coin is repeatedly tossed. The actual number of heads minus the expected number can be represented as $Z_n = X_1 + X_2 + \cdots + X_n - \frac{1}{2}n$, where the random variable X_i is equal to 1 if the ith toss of the coin shows heads and X_i is otherwise equal to 0. The random variable Z_n has approximately a normal distribution with expected value 0 and standard deviation $\sigma\sqrt{n}$ when

n is large, where $\sigma = \frac{1}{2}$ is the standard deviation of the X_i. This fact explains the phenomenon that the range of the difference between the number of heads and the number of tails tossed in n fair coin tosses shows a tendency to grow proportionally with $\sqrt{n}$ as n increases (this difference is given by $X_1 + \cdots + X_n - (1 - X_1 + \cdots + 1 - X_n) = 2(X_1 + \cdots + X_n - \frac{1}{2}n))$. The proportion of heads in n coin tosses is $(X_1 + \cdots + X_n)/n$. The probability distribution of $(X_1 + \cdots + X_n)/n$ becomes more and more concentrated around the value 0.5 as n increases, where the deviations from the expected value of 0.5 are on the order of $1/\sqrt{n}$. A similar phenomenon appears in lotto drawings. The difference between the number of times the most frequently drawn number comes up and the number of times the least frequently drawn number comes up in n drawings shows a tendency to increase proportionally with $\sqrt{n}$ as n increases. This phenomenon can be explained using the multivariate central limit theorem (see Chapter 12).

5.8.2 Casino profits

The square-root law and the central limit theorem give further mathematical support to an earlier claim that operating a casino is in fact a risk-free undertaking. However small the house advantage may be, it is fairly well assured of large and stable profits if it spreads its risk over a very large number of gamblers. To illustrate this, let's consider the casino game of red-and-black. The plays are independent and at each play the gambler wins with probability p and loses with probability $1 - p$, where the win probability p satisfies $p < \frac{1}{2}$. The payoff odds are 1 to 1. That is, in case of a win the player gets paid out twice the stake; otherwise, the player loses the stake. Suppose that the player places n bets and stakes the same amount of cash on each bet. The total number of bets won by the player can be represented as the sum $X_1 + \cdots + X_n$, where the random variable X_i is equal to 1 if the player wins the ith bet and X_i is otherwise equal to 0. The expected value and the standard deviation of the Bernoulli variable X_i are given by

$$E(X_i) = 0 \times (1 - p) + 1 \times p = p$$

and

$$\sigma(X_i) = \sqrt{(0 - p)^2 \times (1 - p) + (1 - p)^2 \times p} = \sqrt{p(1 - p)}.$$

The central limit theorem tells us that the random variable $X_1 + \cdots + X_n$ has approximately a normal distribution with expected value np and standard deviation $[p(1 - p)]^{\frac{1}{2}}\sqrt{n}$ if n is sufficiently large. The casino loses money to the player only if the player wins $\frac{1}{2}n + 1$ or more bets (assume that n is even). In other words, the casino only loses money to the player if the number of bets the player wins exceeds the expected value np by

$$\beta_n = \frac{\frac{1}{2}n + 1 - np}{[p\,(1-p)]^{1/2}\sqrt{n}}$$

or more standard deviations. The probability of this is approximately equal to $1 - \Phi\,(\beta_n)$. Because β_n increases proportionally to $\sqrt{n}$ as n increases, the probability $1 - \Phi\,(\beta_n)$ tends very rapidly to zero as n gets larger. In other words, it is practically impossible for the casino to lose money to the gambler when the gambler continues to play. The persistent gambler will always lose in the long run. The gambler's chances are the same as those of a lamb in the slaughterhouse. Assuming that the player stakes one dollar on each bet, then for n plays the profit of the casino over the gambler equals

$$W_n = n - 2(X_1 + \cdots + X_n).$$

Using Property 1 from Section 5.2 and the fact that $X_1 + \cdots + X_n$ has expected value np and standard deviation $[p(1-p)]^{\frac{1}{2}}\sqrt{n}$, it follows that

$$E(W_n) = n\,(1-2p) \quad \text{and} \quad \sigma(W_n) = 2\,[p\,(1-p)]^{\frac{1}{2}}\,\sqrt{n}.$$

The random variable W_n is approximately normally distributed for large n, because $X_1 + \cdots + X_n$ is approximately normally distributed and a linear transformation of a normally distributed random variable is again normally distributed. The fact that W_n is normally distributed allows us to give an insightful formula for the profit that the casino will grab with, say, 99% certainty. The standard normal density has 99% of its probability mass to the right of point -2.326. This means that, with a probability of approximately 99%, the profit of the casino over the player is greater than

$$n\,(1-2p) - 2.326 \times 2\,[p\,(1-p)]^{\frac{1}{2}}\,\sqrt{n}$$

dollars if the player places n bets of one dollar a piece.

In European roulette, the player has a win probability of $p = \frac{18}{37}$ when betting on red. It is interesting to see how quickly the probability of the casino losing money to the player tends to zero as the number (n) of bets placed by the player increases. For European roulette, the casino's loss probability has the values:

loss probability = 0.3553	if n = 100
loss probability = 0.1876	if n = 1,000
loss probability = 0.0033	if n = 10,000
loss probability = 6.1×10^{-18}	if n = 100,000
loss probability = 3.0×10^{-161}	if n = 1,000,000.

This clearly illustrates that the casino will not lose over the long run, notwithstanding the fact that, in the short run, an individual player has a reasonable chance of leaving the casino with a profit. Casinos are naturally more interested in long-run findings because over the long run a great many players will be encountered. The above calculations show that, in European roulette, the casino has a 99% probability of winning an amount of more than $0.02703n - 2.325\sqrt{n}$ dollars from a player when that player bets on n spins of the wheel and stakes one dollar on red each time. This is a steadily growing riskless profit!

All-or-nothing play at the casino

In the short run, an individual player has a good chance of leaving the casino with a profit, but in the long run no player can beat the house percentage of 2.7% for European roulette. A nice illustration of this fact is provided by the all-or-nothing game. Suppose that a player enters the casino with \$80 and has the goal of reaching \$100. The player bets at the roulette table until he has either reached his goal of \$100 or lost everything. As pointed out in Section 2.6, bets may be placed on either single numbers or combinations of numbers, where the combinations of numbers involve 2, 3, 4, 6, 12, or 18 numbers. The payoff odds of a roulette bet with k numbers are $(36/k) - 1$ to 1 and the probability of winning the bet is $k/37$. Using the so-called method of dynamic programming from mathematical optimization theory, an optimal strategy for the all-or-nothing game can be calculated. For the case of an initial capital of \$80 and the goal of reaching \$100, the probability of reaching the goal is 0.78996 when using an optimal betting strategy (e.g., stake \$4 on a six-numbers bet if your current bankroll is \$80 and stake \$1 on red if your current bankroll is \$76). This winning probability is considerably larger than 50%. But wait before you rush to the casino with the idea of making a fortune. If you play the all-or-nothing game *repeatedly* by using the optimal strategy for a single performance of the game, you will lose in the long run 2.7 dollar cents for every dollar staked. This inevitable fact can be mathematically argued as follows. Despite the high success probability of 0.78996 for each game, the expected value of your gain is negative and equals $\$20 \times 0.78996 - \$80 \times 0.21004 = -\$1.004$. This negative expected value of –\$1.004 can be translated into the house percentage of 2.7%. The expected value of the total number of dollars you stake each game can be calculated to be equal to \$37.142. Yes, \$1.004 divided by \$37.142 results in the inevitable house percentage of 2.7%! Over the long term you cannot beat the house percentage of roulette.

5.8.3 Drunkard's walk

In "walking the line," a drunkard repeatedly takes a step to the right with a probability of $\frac{1}{2}$ or a step to the left with a probability of $\frac{1}{2}$. Each step the drunkard takes is of unit length. The consecutive steps are made independently from one another. Let random variable D_n represent the distance of the drunkard from the starting point after n steps. In Section 2.4, it was claimed that

$$E(D_n) \approx \sqrt{\frac{2}{\pi} n}$$

for n large. This claim can easily be proven correct with the help of the central limit theorem. Toward that end, D_n is represented as

$$D_n = |X_1 + \cdots + X_n|,$$

where the random variable X_i is equal to 1 if the ith step of the drunkard goes to the right and is otherwise equal to -1. The random variables $X_1, \ldots, X_n$ are independent and have the same distribution with expected value $\mu = 0$ and standard deviation $\sigma = 1$ (verify!). The central limit theorem now tells us that $X_1 + \cdots + X_n$ is approximately normally distributed with expected value 0 and standard deviation $\sqrt{n}$ for n large. In Example 10.7 in Section 10.3, the probability distribution of $V = |X|$ will be derived for a normally distributed random variable X with expected value zero. Using the results in Example 10.7, we have that $E(D_n) \approx \sqrt{2n/\pi}$ for n large and

$$P(D_n \leq x) \approx \Phi\left(\frac{x}{\sqrt{n}}\right) - \Phi\left(\frac{-x}{\sqrt{n}}\right) \qquad \text{for } x > 0.$$

5.8.4 Kelly betting

In Chapter 2, we saw that the Kelly betting system is an attractive system in a repeated sequence of favorable games. This system prescribes betting the same fixed fraction of your current bankroll each time. It maximizes the long-run rate of growth of your bankroll, and it has the property of minimizing the expected time needed to reach a specified but large size of your bankroll. In Section 2.7, the long-run rate of growth was found with the help of the law of large numbers. The central limit theorem enables you to make statements about the number of bets needed to increase your bankroll with a specified factor. Let's recapitulate the Kelly model. You face a sequence of favorable betting opportunities. Each time you can bet any amount up to your current bankroll. The payoff odds are $f - 1$ to 1. That is, in case of a win, the player gets paid out f times the amount staked; otherwise, the player loses the amount staked. The win probability p of the player is typically less than $\frac{1}{2}$, but it is assumed that the product pf is larger than 1 (a favorable bet). Under the Kelly system you bet the same fixed fraction α of your current bankroll each time. Assuming an initial capital of V_0, define the random variable V_n as

$$V_n = \text{ the size of your bankroll after } n \text{ bets.}$$

We ask ourselves the following two questions:

(a) What is the smallest value of n such that

$$E(V_n) \geq aV_0$$

for a given value of $a > 1$?

(b) What is the smallest value of n such that

$$P(V_n \geq aV_0) \geq 0.95$$

for a given value of $a > 1$?

The key to the answers to these questions is the relation

$$V_n = (1 - \alpha + \alpha R_1) \times \cdots \times (1 - \alpha + \alpha R_n)V_0,$$

where $R_1, \ldots, R_n$ are independent random variables with

$$P(R_i = f) = p \quad \text{and} \quad P(R_i = 0) = 1 - p.$$

This relation was obtained in Section 2.7. Next note that

$$\ln(V_n) = \ln(1 - \alpha + \alpha R_1) + \cdots + \ln(1 - \alpha + \alpha R_n) + \ln(V_0).$$

Hence, except for the term $\ln(V_0)$, the random variable $\ln(V_n)$ is the sum of n independent random variables each having the same distribution. Denoting by μ and σ^2 the expected value and the variance of the random variables $\ln(1 - \alpha + \alpha R_i)$, then

$$\mu = p \ln(1 - \alpha + \alpha f) + (1 - p) \ln(1 - \alpha)$$

and

$$\sigma^2 = p[\ln(1 - \alpha + \alpha f) - \mu]^2 + (1 - p)[\ln(1 - \alpha) - \mu]^2.$$

The central limit theorem tells us that $\ln(V_n)$ is approximately $N(n\mu + \ln(V_0), n\sigma^2)$ distributed for n large. Next, we invoke a basic result that will be proved in Chapter 10. If the random variable U has a $N(\nu, \tau^2)$ distribution, then the random variable e^U has a so-called lognormal distribution with expected value $e^{\nu+\tau^2/2}$. This means that the random variable V_n is approximately lognormally distributed with expected value $e^{n\mu+\ln(V_0)+n\sigma^2/2}$ for n large. Thus, Question (a) reduces to finding the value of n for which

$$e^{n\mu+\ln(V_0)+n\sigma^2/2} \approx aV_0,$$

or, $n\mu + n\sigma^2/2 \approx \ln(a)$. For the data $V_0 = 1, a = 2, f = 3$ and $p = 0.4$, the optimal Kelly fraction is $\alpha = 0.1$. After some calculations, we find the answer $n = 36$ bets for Question (a). In order to answer Question (b), note that

$$\begin{aligned} P(V_n \geq aV_0) &= P\big(\ln(V_n) \geq \ln(a) + \ln(V_0)\big) \\ &= P\left(\frac{\ln(V_n) - n\mu - \ln(V_0)}{\sigma\sqrt{n}} \geq \frac{\ln(a) - n\mu}{\sigma\sqrt{n}}\right). \end{aligned}$$

The standardized variable $[\ln(V_n) - n\mu - \ln(V_0)]/[\sigma\sqrt{n}]$ has approximately a standard normal distribution for n large. Thus, the answer to Question (b) reduces to find the value of n for which

$$1 - \Phi\left(\frac{\ln(a) - n\mu}{\sigma\sqrt{n}}\right) \approx 0.95.$$

For the data $V_0 = 1, a = 2, f = 3$ and $p = 0.4$, the optimal Kelly fraction is $\alpha = 0.1$. After some calculations, we find the answer $n = 708$ bets for Question (b).

In his book *A Mathematician Plays the Stock Market,* Basic Books, 2003, John Allen Paulos discusses the following scenario. Hundreds of new dotcom companies are brought to the stock market each year. It is impossible to predict in which direction the stock prices will move, but for half of the companies the stock price will rise 80% during the first week after the stock is introduced and for half of the other companies the price will fall 60% during this period. You have an initial bankroll of \$10,000 for investments. Your investing scheme is to invest your current bankroll in a new dotcom company each Monday morning and sell the stock the following Friday afternoon. Paulos argues that your \$10,000 would likely be worth all of \$1.95 after 52 weeks, despite the fact your expected gain in any week is positive and equals 10% of your investment. His argument is that the most likely paths are the paths in which the stock price rises during half of the time and falls the other half of the time. In such a scenario your bankroll realizes the value of $1.8^{26} \times 0.4^{26} \times \$10{,}000 = \$1.95$ after 52 weeks. How to calculate the probability of ending up with a bankroll more than \$1.95? How can you do better than investing all of your money each week? The answer to the second question is that you better invest a fixed fraction of your bankroll each week rather than your whole bankroll. Using results from Problem 2.10, the optimal Kelly betting fraction is

$$\alpha^* = \frac{0.5 \times 1.8 + 0.5 \times 0.4 - 1}{(1.8 - 1)(1 - 0.4)} = \frac{5}{24}.$$

The above discussion about the Kelly system tells you how to calculate the probability distribution of your bankroll after 52 weeks if you invest the same fraction α of your bankroll each week. Denoting by V_n the size of your bankroll after n weeks, a minor modification of the above analysis shows that $\ln(V_n/V_0)$ is approximately normally distributed with expected value $n\mu_1$ and standard deviation $\sigma_1\sqrt{n}$ for n large enough, where

$$\mu_1 = 0.5\ln(1+0.8\alpha) + 0.5\ln(1-0.6\alpha)$$

and

$$\sigma_1 = \sqrt{0.5[\ln(1+0.8\alpha)-\mu_1]^2 + 0.5[\ln(1-0.6\alpha)-\mu_1]^2}.$$

This gives with $V_0 = 10{,}000$ that

$$\begin{aligned} P(V_n > x) &= P\left(\ln(V_n/V_0) > \ln(x/V_0)\right) \\ &= P\left(\frac{\ln(V_n/V_0) - n\mu_1}{\sigma_1\sqrt{n}} > \frac{\ln(x/10^4) - n\mu_1}{\sigma_1\sqrt{n}}\right) \\ &\approx 1 - \Phi\left(\frac{\ln(x/10^4) - n\mu_1}{\sigma_1\sqrt{n}}\right). \end{aligned}$$

Using this result, we can now answer the first question. If you invest your whole bankroll each week ($\alpha = 1$), then the probability of having a bankroll of more than \$1.95 after 52 weeks is approximately equal to 0.500. This probability is practically equal to 1 if you use the Kelly strategy with $\alpha = \frac{5}{24}$. Denoting by $P(x)$ the probability of having a bankroll larger than x dollars after 52 weeks, it is interesting to compare the values of $P(x)$ for the two strategies with $\alpha = 1$ and $\alpha = \frac{5}{24}$. For $x = 10{,}000$, 20,000, and 50,000, the probability $P(x)$ has the approximate values 0.058, 0.044, and 0.031 when you invest your whole bankroll each week and the approximate values 0.697, 0.440, and 0.150 when you invest a fraction $\frac{5}{24}$ of your bankroll each week. The probabilities obtained from the normal approximation are very accurate, as has been verified by simulation.

5.8.5 Brownian motion[†]

Random movements are abundant in nature: butterfly movement, smoke particles in the air or pollen particles in a water droplet. In 1828, the British botanist Robert Brown noticed that while studying tiny particles of plant pollen in water

[†] This section contains advanced material.

under a microscope, these pieces of pollen traveled about randomly. This apparently obscure phenomenon played a key role in the revolution that occurred in the field of physics in the first decade of the twentieth century. In a landmark 1905 paper, Einstein explained the motion of a tiny particle of pollen was the result of its collisions with water molecules. The rules describing this random motion are pretty similar to the rules describing the random walk of a drunkard. The random walk model and the Brownian motion model are among the most useful probability models in science. Brownian motion appears in an extraordinary number of places. It plays not only a crucial role in physics, but it is also widely applied to the modeling of financial markets. Think of a stock price as a small particle which is "hit" by buyers and sellers. The first mathematical description of stock prices utilizing Brownian motion was given in 1900 by the French mathematician Louis Bachelier (1870–1946), who can be considered as the founding father of modern option pricing theory. His innovativeness, however, was not fully appreciated by his contemporaries, and his work was largely ignored until the 1950s.

This section is aimed at giving readers a better perception of Brownian motion. We present an intuitive approach showing how Brownian motion can be seen as a limiting process of random walks. The central limit theorem is the link between the random walk model and the Brownian motion model. Let's assume a particle that makes every Δ time units either an upward jump or a downward jump of size δ with probabilities p and $1 - p$, where δ and p depend on Δ. The idea is to choose smaller and smaller step sizes for the time and to make the displacements of the random walk smaller as well. As the time-step size Δ gets closer and closer to zero and the displacements decrease proportionally to $\sqrt{\Delta}$, the discrete-time random walk looks more and more like a continuous-time process, called the Brownian motion. To make this more precise, fix for the moment the time-step size Δ. For any $t > 0$, let

$$X^{\Delta}(t) = \text{the position of the particle at time } t.$$

It is assumed that the initial position of the particle is at the origin. The random variable $X^{\Delta}(t)$ can be represented as the sum of independent random variables X_i with

$$X_i = \begin{cases} \delta & \text{with probability } p \\ -\delta & \text{with probability } 1 - p. \end{cases}$$

Letting $\lfloor u \rfloor$ denote the integer that results by rounding down the number u, it holds for any $t > 0$ that

$$X^{\Delta}(t) = X_1 + \cdots + X_{\lfloor t/\Delta \rfloor}.$$

Invoking the central limit theorem, it follows that the random variable $X^{\Delta}(t)$ is approximately normally distributed for t large. Using the fact that

$$E(X_i) = (2p-1)\delta \quad \text{and} \quad \mathrm{Var}(X_i) = 4p(1-p)\delta^2$$

(verify!), the expected value and the variance of $X^{\Delta}(t)$ are given by

$$E[X^{\Delta}(t)] = \lfloor t/\Delta \rfloor (2p-1)\delta \quad \text{and} \quad \mathrm{Var}[X^{\Delta}(t)] = \lfloor t/\Delta \rfloor 4p(1-p)\delta^2.$$

By choosing the displacement size δ and the displacement probability p in a proper way as function of the time-step size Δ and letting Δ tend to zero, we can achieve for any $t > 0$ that

$$\lim_{\Delta\to 0} E[X^{\Delta}(t)] = \mu t \quad \text{and} \quad \lim_{\Delta\to 0} \mathrm{Var}[X^{\Delta}(t)] = \sigma^2 t$$

for given numbers μ and σ with $\sigma > 0$. These limiting relations are obtained by taking

$$\delta = \sigma\sqrt{\Delta} \quad \text{and} \quad p = \frac{1}{2}\left\{1 + \frac{\mu}{\sigma}\sqrt{\Delta}\right\}.$$

It is a matter of simple algebra to verify this result. The details are left to the reader.

We now have made plausible that the random variable $X^{\Delta}(t)$ converges in a probabilistic sense to an $N(\mu t, \sigma^2 t)$-distributed random variable $X(t)$ when the time-step size Δ tends to zero and δ and p are chosen according to $\delta = \sigma\sqrt{\Delta}$ and $p = \frac{1}{2}\{1 + \frac{\mu}{\sigma}\sqrt{\Delta}\}$. The random variable $X(t)$ describes the position of the particle in a continuous-time process at time t. The process $\{X(t)\}$ was constructed as a limiting process by rescaling a discrete random walk in such a way that the time between transitions shrinks to zero and simultaneously the size of the jumps contracts appropriately to zero. Using deep mathematics, it can be shown that the random process $\{X(t)\}$ has the following properties:

(a) the sample paths of the process are continuous functions of t

(b) the increments $X(t_1) - X(t_0),\ X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1})$ are independent for all $0 \leq t_0 < t_1 < \cdots < t_{n-1} < t_n$ and $n > 1$

(c) $X(s+t) - X(s)$ is $N(\mu t, \sigma^2 t)$ distributed for all $s \geq 0$ and $t > 0$.

A random process $\{X(t)\}$ having these properties is called a *Brownian motion* with drift parameter μ and variance parameter σ^2. The parameter μ reflects the expected change in the process per unit of time and is therefore called the *drift* of the process. The parameter σ is a measure for the standard deviation of the change per unit time and is often called the *volatility* of the process. The Brownian motion process is often referred to as the Wiener process after the

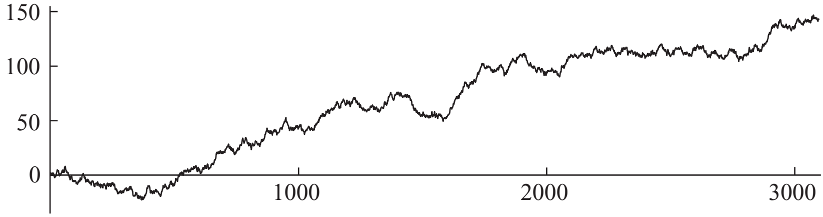


Fig. 5.9. A realization of Brownian motion.

American mathematician Norman Wiener who laid the mathematical foundation of Brownian motion and showed the existence of a random process $X(t)$ satisfying properties (a)–(c). The random variable $X(t_i) - X(t_{i-1})$ is called the increment in the process $\{X(t)\}$ between the times t_{i-1} and t_i. Since the distribution of the increment $X(t_i) - X(t_{i-1})$ depends only on the length $t_i - t_{i-1}$ of the interval $[t_{i-1}, t_i)$ and not on the times t_{i-1} and t_i, the process $\{X(t)\}$ is said to have stationary increments. Also, by property (b), the increments are independent. The Poisson process from Chapter 4 is another example of a random process with stationary and independent increments. In the Poisson process the increments have a Poisson distribution, whereas in the Brownian motion process the increments are normally distributed.

A peculiar feature of Brownian motion is that the probability of occurrence of a sample path being either decreasing or increasing on any finite time interval is zero, no matter how short the interval is. In other words, the sample paths are very kinky and nowhere differentiable, although they are continuous functions of the time t (see Figure 5.9). An intuitive explanation of this remarkable property is as follows. Divide any given small time interval of length L in many smaller disjoint subintervals of length Δ and note that the increments of the Brownian motion in the disjoint subintervals are independent. Each increment is normally distributed with mean $\mu\Delta$ and thus takes on a positive or a negative value each with an approximate probability of 0.5 as Δ tends to zero. The probability of having increments of the same sign in all of the L/Δ subintervals is thus of the order $0.5^{L/\Delta}$ and tends to zero as Δ approaches zero. This explains why a typical Brownian path is nowhere differentiable, in agreement with the phenomenon that a Brownian particle jiggles about randomly. The irregular behavior of Brownian motion can also be explained from the fact that the standard deviation of the change of the process over a small time interval of length Δ is significantly larger than the expected value of the change. The standard deviation is proportional to $\sqrt{\Delta}$ and the expected value is proportional to Δ. For Δ small, $\sqrt{\Delta}$ is significantly larger than Δ.

It is instructive to simulate Brownian motion on the computer. In Monte Carlo simulation, the position of the particle is numerically advanced with the update equation $X(t+\Delta) = X(t) + I(\Delta)$ for a small time-step Δ, where $I(\Delta)$ is $N(\mu\Delta, \sigma^2\Delta)$ distributed. An effective method to simulate random observations from a normal distribution is given in Section 11.3.1. Figure 5.9 displays a simulated realization of Brownian motion.

As pointed out before, Brownian motion has applications in a wide variety of fields. In particular, the application of Brownian motion to the field of finance received a great deal of attention. It was found that the logarithms of common-stock prices can often be very well modeled as Brownian motions. In agreement with this important finding is the result we found in a previous paragraph for the wealth process in the situation of Kelly betting. Other important applications of Brownian motion arise by combining the theory of fractals and Brownian motion. Fractals refer to images in the real world that tend to consist of many complex patterns that recur at various sizes. The fractional Brownian motion model regards naturally occurring rough surfaces such as mountains and clouds as the end result of random walks.

5.8.6 Stock prices and Brownian motion

Let's assume a stock whose price changes every Δ time units, where Δ denotes a small increment of time. Each time the price of the stock goes up by the factor δ with probability p or goes down by the same factor δ with probability $1-p$, where δ and p are given by

$$\delta = \sigma\sqrt{\Delta} \quad \text{and} \quad p = \frac{1}{2}\left\{1 + \frac{\mu}{\sigma}\sqrt{\Delta}\right\}$$

for given values of μ and σ with $\sigma > 0$. It is assumed that Δ is small enough such that $0 < \delta < 1$ and $0 < p < 1$. The initial price of the stock is S_0. If the time-step Δ tends to zero, what happens to the random process describing the stock price? Letting S_t denote the stock price at time t in the limiting process, the answer is that the random process describing $\ln(S_t/S_0)$ is a Brownian motion with drift parameter $\mu - \frac{1}{2}\sigma^2$ and variance parameter σ^2.

An intuitive explanation is as follows. Denote by S_t^Δ the stock price at time t when the stock price changes every Δ time units. Then, for any $t > 0$

$$S_t^\Delta = (1 + X_1) \times \cdots \times \left(1 + X_{\lfloor t/\Delta \rfloor}\right) S_0.$$

Here $X_1, X_2, \ldots$ are independent random variables with

$$P(X_i = \sigma\sqrt{\Delta}) = p \quad \text{and} \quad P(X_i = -\sigma\sqrt{\Delta}) = 1 - p,$$

where the displacement probability $p = \frac{1}{2}(1 + \frac{\mu}{\sigma}\sqrt{\Delta})$. Hence

$$\ln\left(\frac{S_t^\Delta}{S_0}\right) = \sum_{i=1}^{\lfloor t/\Delta \rfloor} \ln(1 + X_i).$$

The next step is to use a basic result from calculus

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots \qquad \text{for } |x| < 1.$$

For fixed t, we may assume that t/Δ is an integer if we let Δ tend to zero in an appropriate way. Since $|X_i| = \sigma\sqrt{\Delta}$, we have $|X_i| < 1$ for Δ small and

$$\sum_{i=1}^{t/\Delta} \frac{X_i^2}{2} = \frac{t}{\Delta}\frac{\sigma^2\Delta}{2} = \frac{1}{2}\sigma^2 t.$$

Also, for Δ small, $\sum_{i=1}^{t/\Delta} \frac{1}{3}X_i^3$ is on the order of $\sqrt{\Delta}$, $\sum_{i=1}^{t/\Delta} \frac{1}{4}X_i^4$ is on the order of Δ, and so on. Hence, the contribution of these terms becomes negligible as $\Delta \to 0$. Thus, using the expansion of $\ln(1 + X_i)$, we find that $\ln(S_t^\Delta/S_0)$ is approximately distributed as

$$\sum_{i=1}^{t/\Delta} X_i - \frac{1}{2}\sigma^2 t$$

for Δ small. It was argued earlier that the random walk process $\sum_{i=1}^{t/\Delta} X_i$ becomes a Brownian motion with drift parameter μ and variance parameter σ^2 when the time-step Δ tends to zero. The sum of an $N(\nu, \tau^2)$ random variable and a constant a has an $N(a + \nu, \tau^2)$ distribution. This is the last step in the intuitive explanation that the process describing $\ln(S_t/S_0)$ is a Brownian motion with drift parameter $\mu - \frac{1}{2}\sigma^2$ and variance parameter σ^2.†

5.8.7 Black-Scholes formula

The Black-Scholes formula is the most often-used formula with probabilities in finance. It shows how to determine the value of an option. An option is a financial product written on another financial product. The latter is typically

† The process $\{S_t\}$ is a so-called geometric Brownian motion: a random process $\{Y(t)\}$ is said to be a geometric Brownian motion with parameters α and σ^2 if $Y(t) = y_0 e^{X(t)}$ with $\{X(t)\}$ is a Brownian motion with drift parameter $\alpha - \frac{1}{2}\sigma^2$ and variance parameter σ^2 (it can be shown that $E[Y(t)] = y_0 e^{\alpha t}$ and so α is the growth rate of the $\{Y(t)\}$ process). In geometric Brownian motion the relative changes $Y(t_1)/Y(t_0), \ldots, Y(t_n)/Y(t_{n-1})$ over nonoverlapping time intervals are independent and have identical distributions over time intervals of the same length. This is a reasonable description for behavior of stock prices.

referred to as the "underlying." A call option gives the holder the right, but not the obligation, to buy some underlying stock at a given price, called the exercise price, on a given date. The buyer pays the seller a premium for this right. The premium is the value of the option. Taking $t = 0$ as the current date, let

$$\begin{aligned} T &= \text{time to maturity of the option (in years)} \\ S_0 &= \text{current stock price (in dollars)} \\ K &= \text{exercise price of the option (in dollars)} \\ r &= \text{risk-free interest rate (annualized)} \\ \sigma &= \text{underlying stock volatility (annualized).} \end{aligned}$$

The risk-free interest rate is assumed to be continuously compounded. Thus, if $r = 0.07$, this means that in one year \$1 will grow to $e^{0.07}$ dollars. The volatility parameter σ is nothing else than the standard deviation parameter of the Brownian motion process that is supposed to describe the process $\ln(S_t/S_0)$ with S_t denoting the price of the underlying stock at time t. On the basis of economical considerations, the drift parameter η of this Brownian motion process is chosen as

$$\eta = r - \frac{1}{2}\sigma^2.$$

An intuitive explanation for this choice is as follows. In an efficient market, it is reasonable to assume that betting on the price change of the stock in a short time interval is a fair bet. That is, the condition $E(S_\Delta) - e^{r\Delta}S_0 = 0$ is imposed for Δ small. Letting $W = \ln(S_\Delta/S_0)$ and using the fact that $e^{\ln(a)} = a$, we have $E(S_\Delta/S_0) = E(e^W)$. Since W is $N(\eta\Delta, \sigma^2\Delta)$ distributed, a basic result for the normal distribution tells us that $E(e^W) = e^{\eta\Delta + \frac{1}{2}\sigma^2\Delta}$ (see Example 14.4 in Chapter 14). Thus, the condition $E(S_\Delta) - e^{r\Delta}S_0 = 0$ is equivalent to $e^{\eta\Delta+\frac{1}{2}\sigma^2\Delta} = e^{r\Delta}$, yielding $\eta = r - \frac{1}{2}\sigma^2$. In the real world, the volatility parameter σ is estimated from the sample variance of the observations $\ln(S_{i\Delta}/S_{(i-1)\Delta})$ for $i = 1, 2, \ldots, h$ (say, $h = 250$) over the last h trading days of the stock.

We now turn to the determination of the price of the option. To do so, it is assumed that the option can only be exercised at the maturity date T. Furthermore, it is assumed that the stock will pay no dividend before the maturity date. The option will be exercised at the maturity date T only if $S_T > K$. Hence, at maturity, the option is worth

$$C_T = \max(0, S_T - K).$$

The net present value of the worth of the option is $e^{-rT}C_T$. Using the fact that $\ln(S_T/S_0)$ has an $N((r - \frac{1}{2}\sigma^2)T, \sigma^2T)$ distribution, it is matter of integral calculus and standard formulas to evaluate the expression for the option price $e^{-rT}E(C_T)$. This gives the *Black-Scholes formula*

$$e^{-rT}E(C_T) = \Phi(d_1)S_0 - \Phi(d_2)Ke^{-rT}$$

with

$$d_1 = \frac{\ln(S_0/K) + \big(r + \frac{1}{2}\sigma^2\big)T}{\sigma\sqrt{T}} \quad \text{and} \quad d_2 = d_1 - \sigma\sqrt{T},$$

where $\Phi(x)$ is the standard normal distribution function. This beautiful mathematical formula was developed by Fisher Black, Robert Merton, and Myron Scholes. Its publication in 1973 removed the guesswork and reliance on individual brokerage firms from options pricing and brought it under a theoretical framework that is applicable to other derivative products as well. The Black-Scholes formula changed the world, financial markets, and indeed capitalism as well. It helped give rise to a standardized options industry dealing in the hundreds of billions of dollars.

As a numerical illustration, consider a European call option on 100 shares of a nondividend-paying stock ABC. The option is struck at \$50 and expires in 0.3 years. ABC is trading at \$51.25 and has 30% implied volatility. The risk-free interest is 7%. What is the value of the option? Applying the Black-Scholes formula with $S_0 = 51.25$, $K = 50$, $T = 0.3$, $r = 0.07$, and $\sigma = 0.3$, the value of the option per share of ABC is \$4.5511. The call option is for 100 shares and so it is worth \$455.11. In doing the calculations, the values of the standard normal distribution function $\Phi(x)$ were calculated from the approximation

$$\Phi(x) \approx 1 - \frac{1}{2}\left(1 + d_1x + d_2x^2 + d_3x^3 + d_4x^4 + d_5x^5 + d_6x^6\right)^{-16}, \quad x \geq 0,$$

where the constants $d_1, \ldots, d_6$ are given by

$$\begin{aligned} d_1 &= 0.0498673470, \; d_2 = 0.0211410061, \; d_3 = 0.0032776263, \\ d_4 &= 0.0000380036, \; d_5 = 0.0000488906, \; d_6 = 0.0000053830. \end{aligned}$$

The absolute error of this approximation is less than 1.5×10^{-7} for all $x \geq 0$. For $x < 0$, $\Phi(x)$ can be calculated from $\Phi(x) = 1 - \Phi(-x)$.

5.9 Falsified data and Benford's law

Most people have preconceived notions of randomness that often differ substantially from true randomness. Truly random datasets often have unexpected properties that go against intuitive thinking. These properties can be used to test whether datasets have been tampered with when suspicion arises. To illustrate, suppose that two people are separately asked to toss a fair coin 120 times and take note of the results. Heads is noted as a “one” and tails as a “zero.” The following two lists of compiled zeros and ones result

1	1	0	0	1	0	0	1	0	1	1	0	0	1	0	0	0	1	1	0
1	0	1	0	0	1	1	0	1	0	0	1	0	1	0	1	1	0	1	1
0	0	1	1	0	1	1	1	0	1	0	0	1	0	0	1	1	0	1	0
0	1	1	0	1	0	0	1	1	0	1	0	1	1	0	0	1	1	1	0
0	1	0	1	0	1	0	0	0	1	0	1	0	1	0	1	0	1	0	1
1	0	0	1	0	0	1	0	1	1	0	0	1	0	0	1	1	0	1	1

and

1	1	1	0	0	0	1	1	1	0	1	0	1	1	1	1	1	1	0	1
0	0	0	1	1	0	0	1	1	0	1	0	1	0	0	0	1	1	0	1
0	0	1	1	1	0	1	0	0	0	0	1	0	1	1	1	0	1	1	0
0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	0	1	1	0	1
0	1	1	1	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	1
1	1	1	0	1	1	1	1	0	1	0	1	1	0	1	1	0	1	0	1

One of the two individuals has cheated and has fabricated a list of numbers without having tossed the coin. Which is the fabricated list? The key to solving this dilemma lays in the fact that in 120 tosses of a fair coin, there is a very large probability that *at some point* during the tossing process, a sequence of five or more heads or five or more tails will naturally occur. The probability of this is 0.9865. In contrast to the second list, the first list shows no such sequence of five heads in a row or five tails in a row. In the first list, the longest sequence of either heads or tails consists of three in a row. In 120 tosses of a fair coin, the probability of the longest sequence consisting of three or less in a row is equal to 0.000053, which is extremely small indeed. Thus, the first list is almost certainly a fake. Most people tend to avoid noting long sequences of consecutive heads or tails. Truly random sequences do not share this human tendency!

5.9.1 Success runs[†]

How can we calculate the probability of the occurrence of a success run of a certain length in a given number of coin tosses? Among other things, this probability comes in handy when tackling questions such as the one posed in Chapter 1: what is the probability of a basketball player with a 50% success rate shooting five or more baskets in a row in 20 attempts? We learned in

[†] This section contains advanced material.

Section 2.1.3, with the help of computer simulation, that the player has approximately a 25% probability of achieving such a lengthy success run. However, this probability can also be exactly calculated. In this paragraph, we give an exact method to use in answering the following question: what is the probability of getting a run of r heads in n fair coin tosses? To answer this question, let's say that the tossing process is in state (i, k) when there are still k tosses to go and heads came up in the last i tosses but so far a run of r heads has not occurred. Define

$$\begin{aligned} u_k(i) = \; & \text{the probability of getting a run of } r \text{ heads during } n \text{ tosses} \\ & \text{when the current state of the tossing process is } (i, k). \end{aligned}$$

The index k runs through $0, 1, \ldots, n$ and the index i through $0, 1, \ldots, r$. The probability $u_n(0)$ is being sought. To set up a recursion equation for the probability $u_k(i)$, we condition on the outcome of the next toss after state (i, k). Heads comes up in the next toss with probability $\frac{1}{2}$. If this happens, the next state of the tossing process is $(i + 1, k - 1)$; otherwise, the next state is $(0, k - 1)$. Thus, by the law of conditional probabilities, we find the following recursion for $k = 1, 2, \ldots, n$

$$u_k(i) = \frac{1}{2}u_{k-1}(i + 1) + \frac{1}{2}u_{k-1}(0) \qquad \text{for } i = 0, 1, \ldots, r - 1.$$

This recursion equation has the boundary conditions

$$u_0(i) = 0 \qquad \text{for } 0 \le i \le r - 1 \quad \text{and} \quad u_k(r) = 1 \qquad \text{for } 0 \le k \le n - 1.$$

The recursion equation leads to a simple method in order to calculate the probability $u_n(0)$ exactly. Beginning with $u_0(i) = 0$ for $0 \le i \le r - 1$ and $u_0(i) = 1$ for $i = r$, we first calculate $u_1(i)$ for $0 \le i \le r$, then $u_2(i)$ for $0 \le i \le r$ and going on recursively, we eventually arrive at the desired probability $u_n(0)$.

Applying the recursion with $n = 20$ and $r = 5$ leads to the value 0.2499 for the probability that in 20 shots a basketball player with a successful shot rate of 50% will shoot five or more baskets in a row (Question 2 from Chapter 1). This is the same value as was found earlier with computer simulation in Chapter 2. Isn't it fascinating to see how two fundamentally different approaches lead to the same answer? Yet another approach for success runs will be discussed in Section 15.3.

A similar recursion can be given to calculate the probability that in n fair coin tosses a run of r heads or r tails occurs. In this case, we say that the tossing process is in state (i, k) when there are k tosses still to go and the last i tosses all showed the same outcome but so far no run of r heads or r tails has occurred. The probability $v_k(i)$ is defined as

$v_k(i) =$ the probability of getting a run of r heads or r tails during n tosses when the current state of the tossing process is (i, k).

The probability $v_{n-1}(1)$ is being sought (why?). Verify for yourself that the following recursion applies for $k = 1, 2, \ldots, n$

$$v_k(i) = \frac{1}{2} v_{k-1}(i+1) + \frac{1}{2} v_{k-1}(1) \qquad \text{for } i = 1, \ldots, r-1.$$

The boundary conditions are $v_0(i) = 0$ for $1 \leq i \leq r-1$ and $v_j(r) = 1$ for $0 \leq j \leq n-1$. If you apply the recursion with $n = 120$ and $r = 5$, then you arrive at the earlier found value $v_{n-1}(1) = 0.9865$ for the probability of tossing five heads or five tails in a row in 120 fair coin tosses. The recursion with $n = 120$ and $r = 4$ gives the value $1 - v_{n-1}(1) = 0.000053$ for the probability that in 120 fair coin tosses the longest run of either heads or tails has a length of no more than three. More about success runs in Section 15.3.

5.9.2 Benford's law

In 1881, the astronomer/mathematician Simon Newcomb published a short article in which he noticed that the pages of logarithm tables with small initial digits were dirtier than those with larger initial digits. Apparently, numbers beginning with 1 were more often looked up than numbers beginning with 2, and numbers beginning with 2 more often than numbers beginning with 3, etc. Newcomb quantified this surprising observation in a logarithmic law giving the frequencies of occurrence of numbers with given initial digits. This law became well known for the first time, many years later, as Benford's law. In 1938, physicist Frank Benford rediscovered the "law of anomalous numbers," and published an impressive collection of empirical evidence supporting it. Benford's law says that in many naturally occurring sets of numerical data, the first significant (nonzero) digit of an arbitrarily chosen number is not equally likely to be any one of the digits $1, \ldots, 9$, as one might expect, but instead is closely approximated by the logarithmic law

$$P(\text{first significant digit} = d) = \log_{10}\left(1 + \frac{1}{d}\right) \qquad \text{for } d = 1, 2, \ldots, 9.$$

Figure 5.10 shows the values of these probabilities. Benford's empirical evidence showed that this logarithmic law was fairly accurate for the numbers on the front pages of newspapers, the lengths of rivers, stock prices, universal constants in physics and chemistry, numbers of inhabitants of large cities, and many other tables of numerical data. It appeared that the logarithmic law was

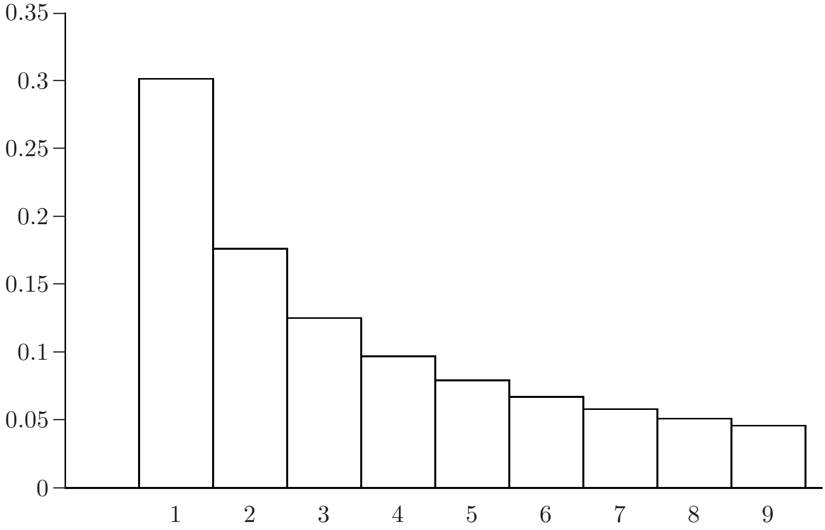


Fig. 5.10. Probability distribution of the first significant digit.

a nearly perfect approximation if all these different datasets were combined. Of course, not every dataset follows Benford's law. For example, consider the times for the Olympic 400-meter race. Very few of those times will begin with a 1! The same is true for the telephone numbers in New York City. Surprisingly enough, Benford's law applies to the Fibonacci numbers.

This led to the question of what properties "natural" datasets must satisfy in order to follow Benford's law. It can be proven mathematically that, if a collection of numbers satisfies Benford's law, then the same collection still satisfies this law if every number in the collection is multiplied by the same positive constant. This shows, for example, that, for Benford's law, it does not matter whether the lengths of rivers are expressed in miles or in kilometers. Moreover, the logarithmic distribution is the only distribution that is scale invariant. This still does not explain why so many "natural" datasets satisfy Benford's law. An explanation for this phenomenon was recently given by the American mathematician Ted Hill. Roughly speaking, Hill showed the following: if numbers are selected at *random* from *different* arbitrarily chosen collections of data, then the numbers in the *combined* sample will tend to follow Benford's law. The larger and more varied the sample from the different datasets, the more likely it is that the relative frequency of the first significant digits will tend to obey Benford's law. This result offers a plausible theoretical explanation, for example, of the fact that the numbers from the front pages of newspapers are a very good fit to Benford's law. Those numbers typically arise from many sources, and are influenced by many factors.

Benford's law, which at first glance appears bizarre, does have practical applications. The article by Ted Hill, "The difficulty of faking Data," *Chance Magazine* **12** (1999): 27–31, discusses an interesting application of Benford's law to help detect possible fraud in tax returns. Empirical research in the United States has shown, for example, that in actual tax returns that correctly reported income, the entries for interest paid and interest received are a very good fit to Benford's law. Companies' returns that deviate from this law, over the course of many years, appear to be fraudulent in many cases.

5.10 The normal distribution strikes again

How to pick a winning lottery number is the subject of many a book about playing the lottery. The advice extended in these entertaining books is usually based on the so-called secret of balanced numbers. Let's take Lotto $6/49$ as an example. In the Lotto $6/49$, the player must choose six different numbers from the numbers $1, \ldots, 49$. For this lottery, players are advised to choose six numbers whose sum add up to a number between 117 and 183. The basic idea here is that the sum of six randomly picked numbers in the lottery is approximately normally distributed. Indeed, this is the case. In Lotto $6/49$, the sum of the six winning numbers is approximately normally distributed with expected value 150 and a spread of 32.8. It is known that a sample from the normal distribution with expected value μ and spread σ will be situated between $\mu - \sigma$ and $\mu + \sigma$, with a probability of approximately 68%. This is the basis for advising players to choose six numbers that add up to a number between 117 and 183. The reasoning is that playing such a number combination raises the probability of winning a $6/49$ Lottery prize. This is nothing but poppycock. It is true enough that we can predict which stretch the sum of the six winning numbers will fall into, but the combination of six numbers that adds up to a given sum can in no way be predicted. In Lotto $6/49$, there are in total 165,772 combinations of six numbers that add up to the sum of 150. The probability of the winning numbers adding up to a sum of 150 is equal to 0.0118546. If you divide this probability by 165,772, then you get the exact value of $1/\binom{49}{6}$ for the probability that a *given* combination of six numbers will be drawn!

We speak of the Lotto r/s when r different numbers are randomly drawn from the numbers $1, \ldots, s$. Let the random variable X_i represent the ith number drawn. The random variables $X_1, \ldots, X_r$ are dependent, but for reasons of symmetry, each of these random variables has the same distribution. From

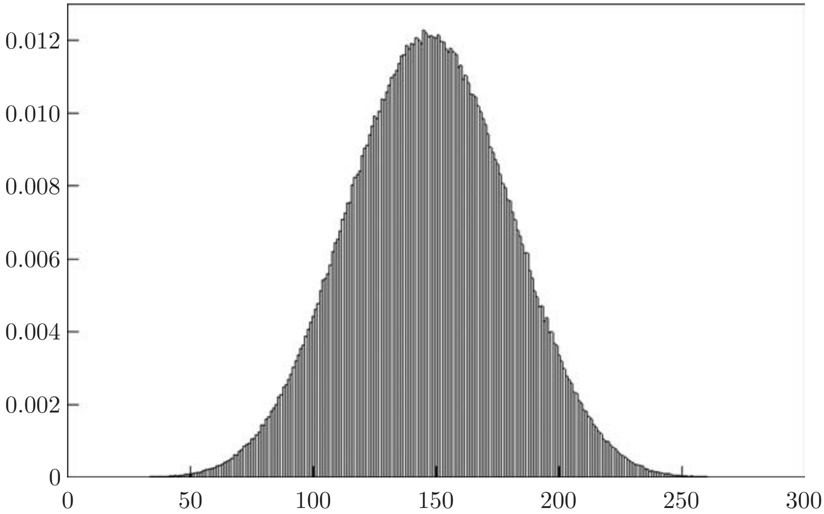


Fig. 5.11. Probability histogram for $r = 6$ and $s = 49$.

$P(X_1 = k) = 1/s$ for $k = 1, \ldots, s$, it follows that $E(X_1) = \frac{1}{2}(s+1)$. This leads to

$$E(X_1 + \cdots + X_r) = \frac{1}{2}r(s+1).$$

It is stated without proof that

$$\sigma^2(X_1 + \cdots + X_r) = \frac{1}{12}r(s-r)(s+1).$$

Also, for r and $s - r$ both sufficiently large, it can proved that the sum $X_1 + \cdots + X_r$ is approximately normally distributed with expected value $\frac{1}{2}r(s+1)$ and variance $\frac{1}{12}r(s-r)(s+1)$. Figure 5.11 displays the simulated frequency diagram of $X_1 + \cdots + X_r$ for $r = 6$ and $s = 49$. The simulation consisted of one million runs. A glance at Figure 5.11 confirms that the probability histogram of $X_1 + \cdots + X_r$ can indeed be approximated by the normal density function.

5.11 Statistics and probability theory

In this chapter, we have already introduced several statistical problems. Statistics and probability theory are distinct disciplines. Probability theory is a branch of mathematics. In mathematics, we reason from the general to the specific. Given a number of axioms, we can derive general propositions that we can then apply to specific situations. This is called *deductive* reasoning. The deductive nature of probability theory is clearly demonstrated in Chapter 7. Statistics, on the other hand, works the other way around by reasoning from the specific to the general. Statistics is therefore a science based on *inductive* reasoning. In statistics we attempt to draw more generally valid conclusions based on data obtained from a specific situation. For example, statisticians attempt to discern the general effectiveness of new medicines based on their effectiveness in treating limited groups of test patients. To do so, statisticians must select a method based on one of two schools of thought. Most statisticians base their methods on the *classical* approach, whereas others base their methods on the *Bayesian* approach. In the classical approach, the test of the null hypothesis is based on the idea that any observed deviation from what the null hypothesis predicts is solely the product of chance. If something that is unusual under the null hypothesis happens, then the null hypothesis is rejected. It is common to use a significance level of 5% or 1% as a benchmark for the probability to be judged. Note that in the classical approach, the probability of rejecting the null hypothesis is not the same as the probability that the null hypothesis is false.

The Bayesian approach assumes an *a priori* probability distribution (called the *prior*) as to whether or not the null hypothesis is true. The prior distribution is then updated in the light of the new observations. Simply put, the classical approach is based on $P(\text{data}|H_0)$, whereas the Bayesian approach makes use of $P(H_0|\text{data})$. The need to specify a prior distribution before analyzing the data introduces a subjective element into the analysis and this is often regarded as a weakness of the Bayesian approach. It is, however, important to keep in mind that the classical approach is not entirely objective either. The choice whether to reject the null hypothesis at the 5% significance level instead of at, for example, the 0.1% level is also subjective. The fundamental difference between the classical and Bayesian approach can be best illustrated via Example 5.1. This example deals with a multiple-choice exam consisting of 50 questions, each of which has three possible answers. A student receives a passing grade if he/she correctly answers more than half of the questions. Take the case of a student who manages to answer 26 of the 50 questions correctly and claims not to have studied, but rather to have obtained 26 correct answers merely by guessing. In Section 5.6, we see that the classical approach is based on the probability that 26 or more of the 50 questions could be answered by luck alone. The Bayesian approach is based on a different probability: the probability that the student could have guessed each answer given that he/she answered 26 of the 50 questions correctly. The Bayesian approach to determining this probability requires that we first specify a prior distribution for the various ways the student may have prepared for the exam. This distribution concerns the situation *before* the exam and can be a purely subjective assessment (although it can also be based on information of the student's earlier academic performance on homework or previous exams). Let us assume for simplicity's sake that there are only two possibilities: either the student was totally unprepared (hypothesis H) or that the student was well prepared (the complementary hypothesis $\overline{H}$). We furthermore assume that the assessment before the exam was that with a probability of 50% the student was well prepared. In other words, $P(H) = P(\overline{H}) = \frac{1}{2}$. Using the binomial distribution with $n = 50$ and $p = \frac{1}{3}$, it then follows that with a probability of 0.00492 the student could pass the exam if the student did not study (and therefore guessed the answer to all of the questions). Let us now make the additional assumption that based on experience it is known that a well-prepared student passes an exam 70% of the time. In the Bayesian approach (see Chapter 8), we then conclude that with a probability of 0.7% the student did not study and could only complete the exam by guessing, given the fact that he/she passed the exam. Although this represents at least partially a subjective estimate, it is in any case based on a "reasonable" choice for the prior probabilities. If we had instead assumed the prior probabilities $P(H) = 0.8$ and $P(\overline{H}) = 0.2$, then our estimated Bayesian probability would have been 2.7%. Generally speaking we come to the same conclusion we found using the classical approach: it is very likely that the student is bluffing if he/she claims to have passed the exam without studying.

The following example also clearly demonstrates the differences between classical and Bayesian statistics.[†] Imagine that you participate in a game requiring you to guess the number of heads resulting from 50 coin tosses. We would expect approximately 25 heads, but imagine that the actual result is 18 heads. Is this result the product of chance, or is the coin not a fair one? The Bayesian approach makes it possible to estimate the probability that heads results less than 50% of the time given the observation of 18 heads. The approach requires the specification of the prior probability of obtaining heads based only on information available before the game begins. In the classical approach, on the other hand, we determine the probability of 18 or fewer heads given the hypothesis that the coin is fair. This, however, does not result in a statement about the probability of the coin being fair. For such a statement, we need Bayesian anal-

[†] An especially readable book on Bayesian statistics with emphasis on medical applications is D.A. Berry's *Statistics: A Bayesian Perspective*, Duxbury Press, 1996. Of particular relevance in the case of medical applications, and contrary to the classical approach, the Bayesian approach permits us to draw intermediate conclusions based on partial results from an ongoing experiment and, as a result, to modify the future course of the experiment in light of these conclusions.

ysis. Bayesian statistics is discussed in more detail in Chapter 8. The spirit of Reverend Bayes (1702–1761) is still very much alive!

5.12 Problems

5.1 You draw 12 random numbers from (0,1) and average these 12 random numbers. Which of the following statements is then correct?

(a) the average has the same uniform distribution as each of the random numbers

(b) the distribution of the average becomes more concentrated in the middle and less at the ends.

5.2 Someone has written a simulation program in an attempt to estimate a particular probability. Five hundred simulation runs result in an estimate of 0.451 for the unknown probability with 0.451±0.021 as the corresponding 95% confidence interval. One thousand simulation runs give an estimate of 0.453 with a corresponding 95% confidence interval of 0.453±0.010. Give your opinion:

(a) there is no reason to question the programming

(b) there is an error in the simulation program.

5.3 The annual rainfall in Amsterdam is normally distributed with an expected value of 799.5 mm and a standard deviation of 121.4 mm. Over many years, what is the proportion of years that the annual rainfall in Amsterdam is below 550 mm?

5.4 The cholesterol level for an adult male of a specific racial group is normally distributed with an expected value of 5.2 mmol/l and a standard deviation of 0.65 mmol/l. Which cholesterol level is exceeded by 5% of the population?

5.5 Gestation periods of humans are normally distributed with an expected value of 266 days and a standard deviation of 16 days. What is the percentage of births that are more than 20 days overdue?

5.6 In a single-product inventory system a replenishment order will be placed as soon as the inventory on hand drops to the level s. You want to choose the reorder point s such that the probability of a stockout during the replenishment lead time is no more than 5%. Verify that s should be taken equal to $\mu + 1.645\sigma$ when the total demand during the replenishment lead time is $N(\mu, \sigma^2)$ distributed.

5.7 Suppose that the rate of return on stock A takes on the values 30%, 10%, and -10% with respective probabilities 0.25, 0.50, and 0.25 and on stock B the values 50%, 10%, and -30% with the same probabilities 0.25, 0.50, and 0.25. Each stock, then, has an expected rate of return of 10%. Without calculating the actual values of the standard deviation, can you argue why the standard deviation of the rate of return on stock B is twice as large as that on stock A?

5.8 You wish to invest in two funds, A and B, both having the same expected return. The returns of the funds are negatively correlated with correlation coefficient ρ_{AB}. The standard deviations of the returns on funds A and B are given by σ_A and σ_B. Demonstrate that you can achieve a portfolio with the lowest standard deviation by investing a fraction f of your money in fund A and a fraction $1 - f$ in fund B, where the optimal fraction f is given by $(\sigma_B^2 - \sigma_A\sigma_B\rho_{AB})/(\sigma_A^2 + \sigma_B^2 - 2\sigma_A\sigma_B\rho_{AB})$. *Remark*: use the fact that $\text{cov}(aX, bY) = ab\text{cov}(X, Y)$.

5.9 You want to invest in two stocks A and B. The rates of return on these stocks in the coming year depend on the development of the economy. The economic prospects for the coming year consist of three equally likely case scenarios: a strong economy, a normal economy, and a weak economy. If the economy is strong, the rate of return on stock A will be equal to 34% and the rate of return on stock B will be equal to -20%. If the economy is normal, the rate of return on stock A will be equal to 9.5% and the rate of return on stock B will be equal to 4.5%. If the economy is weak, stocks A and B will have rates of return of -15% and 29%, respectively.

(a) Is the correlation coefficient of the rates of return on the stocks A and B positive or negative? Calculate this correlation coefficient.

(b) How can you divide the investment amount between two stocks if you desire a portfolio with a minimum variance? What are the expected value and standard deviation of the rate of return on this portfolio?

5.10 Suppose that the random variables $X_1, X_2, \ldots, X_n$ are defined on a same probability space. In Chapter 11, it will be seen that

$$\sigma^2\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma^2(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \operatorname{cov}(X_i, X_j).$$

For the case that $X_1, \ldots, X_n$ all have the same variance σ^2 and $\operatorname{cov}(X_i, X_j)$ is equal to a constant $c \neq 0$ for all i, j with $i \neq j$, verify that the variance of $\overline{X}(n) = (1/n)\sum_{k=1}^{n} X_k$ is given by

$$\sigma^2\left(\overline{X}(n)\right) = \frac{\sigma^2}{n} + \left(1 - \frac{1}{n}\right)c.$$

In investment theory, the first term σ^2/n is referred to as the nonsystematic risk and the second term $(1 - 1/n)c$ is referred to as the systematic risk. The nonsystematic risk can be significantly reduced by diversifying to a large number of stocks, but a bottom-line risk cannot be altogether eliminated. Can you explain this in economic terms?

5.11 Consider the investment example from Section 5.2 in which a retiree invests \$100,000 in a fund in order to reap the benefits for 20 years. The rate of return on the fund for the past year was 14%, and the retiree hopes for a yearly profit of \$15,098 over the coming 20 years. If the rate of return remained at 14% for each year, then at the end of the xth year, the invested capital would be equal to $f(x) = (1+r)^x A - \sum_{k=0}^{x-1}(1+r)^k b$ for $x = 1, \ldots, 20$, where $A = 100{,}000$, $r = 0.14$, and $b = 15{,}098$. However, the yearly rate of return fluctuates with an average value of 14%. If last year the rate of return was $r\%$, then next year the rate of return will be $r\%$, $(1+f)r\%$ or $(1-f)r\%$ with respective probabilities p, $\frac{1}{2}(1-p)$, and $\frac{1}{2}(1-p)$. For each of the cases ($p = 0.8$, $f = 0.1$) and ($p = 0.5$, $f = 0.2$), simulate a histogram of the distribution of the number of years during the 20-year period that the invested capital at the end of the year will fall below or on the curve of the function $f(x)$.

5.12 The Argus Investment Fund's Spiderweb Plan is a 60-month-long contract according to which the customer agrees to deposit a fixed amount at the beginning of each month. The customer chooses beforehand for a fixed deposit of \$100, \$250, or \$500. Argus then immediately deposits 150 times that monthly amount, to remain in the fund over the five-year period (i.e., Argus deposits \$15,000 of capital in the fund if the customer opts for the \$100 fixed monthly deposit). The monthly amount deposited by the customer is actually the interest payment (8%) on the capital invested by the fund. Five years later, the customer receives the value of the investment minus the initial capital investment. Let's assume that the yearly rate of return on the Argus investment fund fluctuates according to the following probability model: if the return was $r\%$ for the previous year, then for the coming year the return will remain at $r\%$ with a probability of p_s, will change to $(1 - f_d)r\%$ with a probability of p_d, and will change to $(1 + f_u)r\%$ with a probability of p_u, where $p_u + p_d + p_s = 1$ and $0 < f_d, f_u < 1$. Choose reasonable values for the parameters p_s, p_d, p_u, f_d, and f_u. Simulate a histogram for the probability distribution of the customer's capital after five years. Also, use simulation to estimate the expected value and the standard deviation of the customer's rate of return on the monthly deposits.

5.13 An investor decides to place \$2,500 in an investment fund at the beginning of each year for a period of 20 years. The rate of return on the fund was 14% for the previous year. If the yearly rate of return remained at 14% for each year, then, at the end of 20 years, the investor will have an amount of $\sum_{k=1}^{20}(1 + 0.14)^k 2{,}500 = 259{,}421$ dollars. Suppose now that the yearly rate of return fluctuates according to the following probability model: if last year the rate of return was $r\%$, then during the coming year the rate of return will be $r\%$, $(1 + f)r\%$ or $(1 - f)r\%$ with respective probabilities p, $\frac{1}{2}(1 - p)$ and $\frac{1}{2}(1 - p)$. Use simulation to determine for several combinations of f and p a probability histogram for the investor's capital after 20 years. What are the expected value and the standard deviation of the investor's capital after 20 years?

5.14 Women spend on average about twice as much time in the restroom as men, but why is the queue for the women's restroom on average four or more times as long as the one for the men's? This intriguing question was answered in the article "Ladies in waiting" by Robert Matthews in *New Scientist* **167** (2000, July 29): 40. Explain the answer using the Pollaczek-Khintchine formula discussed in Section 5.2. Assume that there is one restroom for women only and one restroom for men only, the arrival processes of women and men are Poisson processes with equal intensities, and the coefficient of variation of the time people spend in the restroom is the same for women as for men.

5.15 What happens to the value of the probability of getting at least r sixes in one throw of $6r$ dice as $r \to \infty$? Explain your answer.

5.16 The owner of a casino in Las Vegas claims to have a perfectly balanced roulette wheel. A spin of a perfectly balanced wheel stops on red an average of 18 out of 38 times. A test consisting of 2,500 trials delivers 1,105 red finishes. If the wheel is perfectly balanced, is this result plausible? Use the normal distribution to answer this question.

5.17 Each year in Houndsville an average of 81 letter carriers are bitten by dogs. In the past year, 117 such incidents were reported. Is this number exceptionally high?

5.18 In a particular area, the number of traffic accidents hovers around an average of 1,050. Last year, however, the number of accidents plunged drastically to 920. Authorities suggest that the decrease is the result of new traffic safety measures that have been in effect for one year. Statistically speaking, is there cause to doubt this explanation? What would your answer be if, based on a yearly average of 105 traffic accidents, the record for the last year decreased to 92 accidents?

5.19 A national information line gets approximately 100 telephone calls per day. On a particular day, only 70 calls come in. Is this extraordinary?

5.20 A large table is marked with parallel and equidistant lines a distance D apart. A needle of length $L(\leq D)$ is tossed in the air and falls at random onto the table. The eighteenth century French scientist Georges-Louis Buffon proved that the probability of the needle falling across one of the lines is $\frac{2L}{\pi D}$. The Italian mathematician M. Lazzarini carried out an actual experiment in 1901, where the ratio L/D was taken equal to 5/6. He made 3,408 needle tosses and observed that 1,808 of them intersected one of the lines. This resulted in a remarkably accurate estimate of 3.14159292 for $\pi = 3.14159265\ldots$ (an error of about 2.7×10^{-7}). Do you believe that Lazzarini performed the experiment in a statistically sound way?

5.21 A gambler claims to have rolled an average of 3.25 points per roll in 1,000 rolls of a fair die. Do you believe this?

5.22 In the 52 drawings of Lotto 6/45 in Orange Country last year an even number was drawn 162 times and an odd number 150 times. Does this outcome cast doubts on the unbiased nature of the drawings? *Hint*: the number of even numbers obtained in a single drawing of Lotto 6/45 has a hypergeometric distribution with expected value 2.93333 and standard deviation 1.32889.

5.23 The Dutch lotto formerly consisted of drawing six numbers from the numbers $1, \ldots, 45$ but the rules were changed. In addition to six numbers from $1, \ldots, 45$, a colored ball is drawn from six distinct colored balls. A statistical analysis of the lotto drawings in the first two years of the new lotto revealed that the blue ball was drawn 33 times in the 107 drawings. The lottery officials hurriedly announced that the painted balls are all of the same weight and that this outcome must have been due to chance. What do you think about this statement?

5.24 In a particular small hospital, approximately 25 babies per week are born, while in a large hospital approximately 75 babies per week are born. Which hospital, do you think, has a higher percentage of weeks during which more than 60% of the newborn babies are boys? Argue your answer without making any calculations. Using the continuity correction, calculate an approximation for each hospital for the probability that in a given week more than 60% of the newborn babies will be boys and compare this approximation with the exact value of the binomial probability.

5.25 A damage claims insurance company has 20,000 policyholders. The amount claimed yearly by policyholders has an expected value of \$150 and a standard deviation of \$750. Give an approximation for the probability that the total amount claimed in the coming year will be larger than 3.3 million dollars.

5.26 The Nero Palace casino has a new, exciting gambling machine: the multiplying bandit. How does it work? The bandit has a lever or “arm” that the player may depress up to ten times. After each pull, an H (heads) or a T (tails) appears, each with probability $\frac{1}{2}$. The game is over when heads appears for the first time, or when the player has pulled the arm ten times. The player wins $\$2^k$ if heads appears after k pulls ($1 \leq k \leq 10$), and wins $\$2^{11} = \$2{,}048$ if after ten pulls heads has not appeared. In other words, the payoff doubles every time the arm is pulled and heads does not appear. The initial stake for this game is \$15. What is the house advantage? Assume there are 2,000 games played each day. Give an approximation for the probability that the casino will lose money on a given day.

5.27 The Dutch Ministry of Education has taken a random sampling of the student population of 400. The students in the sample group were asked if they were in favor of the introduction of a weekend pass for public transportation. Suppose that 208 students were in favor of the pass. Give a 95% confidence interval for the estimate of the percentage of students from the entire student population that would be in favor of the pass. How large must the sample be to ensure that the 95% confidence interval has a margin of no more than 2%?

5.28 Six million voters are expected to vote in the upcoming presidential election. There are two candidates, A and B. The voters cast their ballots independently of one another and each voter will vote for candidate A with probability p and for candidate B with probability $1 - p$. Calculate for both $p = 0.5$ and $p = 0.501$ the probability that the difference in number of votes cast for each of the two candidates will be less than 300.

5.29 In 1986, an article appeared on the front page of the *New York Times* about the results of a research project on the effect of a light dose of aspirin on the incidence of heart attacks. By means of a carefully selected randomization method, a group of 22,000 healthy middle-aged males was randomly sorted into two groups of the same size: an aspirin group and a placebo group. In the aspirin group, 104 heart attacks occurred, while 209 heart attacks occurred in the placebo group. How can you argue, on the grounds of these results, that it is beyond a reasonable doubt that aspirin contributes to the prevention of heart attacks?

5.30 You are interested in assembling a random sample of young people that occasionally use soft drugs. To prevent people from falsely claiming not to use soft drugs, you have thought of the following procedure. The interviewer asks each young person to toss a coin, keeping the result of the toss a secret. The young person is then instructed that if he/she tosses heads he/she must answer "yes" to the question asked even if the true answer to the question is "no" and that if he/she tosses tails, he/she must simply answer the question with the truth. Suppose that the random sample consists of n young people. Let X_i equal 1 if the ith person answers "yes" and otherwise let X_i be equal to 0. Verify that the unknown value of the fraction of young people that use soft drugs can be estimated by $2\overline{X}(n) - 1$ with the corresponding 95% confidence interval $2\,\overline{X}(n) - 1 \;\pm 1.96\; \times 2\sqrt{\overline{X}(n)[1 - \overline{X}(n)]}/\sqrt{n}$.

5.31 In order to test a new pseudo-random number generator, we let it generate 100,000 random numbers. From this result, we go on to form a binary sequence in which the ith element will be equal to 0 if the ith randomly generated number is smaller than $\frac{1}{2}$, and will otherwise be equal to 1. The binary sequence turns out to consist of 49,487 runs. A run begins each time a number in the binary sequence differs from its direct predecessor. Do you trust the new random-number generator on the basis of this test outcome?

5.32 Use the gambler's ruin formula from Problem 3.26 to make plausible that, for any $c, d > 0$,

$$P(\text{process hits } c \text{ before } -d) = \frac{1 - e^{-2d\mu/\sigma^2}}{1 - e^{-2(d+c)\mu/\sigma^2}}$$

for a Brownian motion process with drift parameter $\mu \neq 0$ and variance parameter σ^2 (the probability is $d/(d+c)$ if $\mu = 0$). *Remark*: use the fact that $\lim_{\Delta \to 0}(1 + a\Delta)^{1/\Delta} = e^a$ for any constant a.

5.33 Consider the stock price process $\{S_t\}$ from Section 5.8.6. Verify that the probability of the stock price increasing to aS_0 without falling down first to bS_0 equals $[1 - b^{2(\mu - \frac{1}{2}\sigma^2)/\sigma^2}]/[1 - (b/a)^{2(\mu - \frac{1}{2}\sigma^2)/\sigma^2}]$ for $0 < b < 1 < a$.

5.34 You have an economy with a risky asset and a riskless asset. Your strategy is to hold always a constant proportion α of your wealth in the risky asset and the remaining proportion of your wealth in the riskless asset, where $0 < \alpha < 1$. The initial value of your wealth is V_0. The rate of return on the risky asset is described by a Brownian motion with a drift of 15% and a standard deviation of 30%. The instantaneous rate of return on the riskless asset is 7%.

(a) Let V_t denote your wealth at time t. Use a random-walk discretization of the process of rate of return on the risky asset in order to give an intuitive explanation of the result that $\ln(V_t/V_0)$ is a Brownian motion with drift parameter $r + \alpha(\mu - r) - \frac{1}{2}\alpha^2\sigma^2$ and variance parameter σ^2, where $r = 0.07, \mu = 0.15$, and $\sigma = 0.3$. Argue that the long-run rate of growth of your wealth is maximal for the Kelly fraction $\alpha^* = (\mu - r)/\sigma^2 = 0.89$.

(b) How much time is required in order to double your initial wealth with a probability of 90%?

6

Chance trees and Bayes' rule

Chance trees provide a useful tool for a better understanding of uncertainty and risk. A lot of people have difficulties assessing risks. Many physicians, for example, when performing medical screening tests, overstate the risk of actually having the disease in question to patients testing positive for the disease. They underestimate the false-positives of the test. Likewise, prosecutors often misunderstand the uncertainties involved in DNA evidence. They confuse the not-guilty probability of a suspect matching the trace evidence with the probability of a person randomly selected from a population matching the trace evidence. Incorrect reasoning with conditional probabilities is often the source of erroneous conclusions. A chance tree is useful in such once-only decision situations containing a degree of uncertainty. It depicts the uncertainty in an insightful way and it clarifies conditional probabilities by decomposing a compound event into its simpler components. We begin our discussion of chance trees with some entertaining problems, such as the three-doors problem and the related three prisoners problem. A lot of time and energy have been expended in the solving of these two problems; numerous people have racked their brains in search of their solutions, alas to no avail. There are several productive ways of analyzing the two problems, but by using a chance tree we run the least amount of risk of falling into traps. This chapter also provides an illustration of how the concept of the chance tree is used for analyzing uncertainties in medical screening tests. Bayes' rule provides an alternative approach in the analysis of situations in which probabilities must be revised in light of new information. This rule will also be discussed in this chapter.

6.1 The Monty Hall dilemma

Seldom has a probability problem so captured the imagination as the one we refer to as the Monty Hall dilemma. This problem, named after the popular 1970s game show host, attracted worldwide attention in 1990 when American columnist Marilyn vos Savant took it on in her weekly column in the Sunday *Parade* magazine. It goes like this. The contestant in a television game show must choose between three doors. An expensive automobile awaits the contestant behind one of the three doors, and gag prizes await him behind the other two. The contestant must try to pick the door leading to the automobile. He chooses a door randomly, appealing to Lady Luck. Then, as promised beforehand, the host opens one of the other two doors concealing one of the gag prizes. With two doors remaining unopened, the host now asks the contestant whether he wants to remain with his choice of door, or whether he wishes to switch to the other remaining door. The candidate is faced with a dilemma. What to do? In her weekly *Parade* column, Marilyn vos Savant advised the contestant to switch to the other remaining door, thereby raising his odds of winning the automobile to a $\frac{2}{3}$ probability. In the weeks that followed, vos Savant was inundated with thousands of letters, some rather pointed to say the least, from readers who disagreed with her solution to the problem. Ninety percent of the letter writers, including some professional mathematicians, insisted that it made no difference whether the player switched doors or not. Their argument was that each of the two remaining unopened doors had a $\frac{1}{2}$ probability of concealing the automobile. The matter quickly transcended the borders of the United States, gathering emotional impact along the way. Note the reaction in this letter to the editor published in a Dutch newspaper: "The unmitigated gall! Only sheer insolence would allow someone who failed mathematics to make the claim that the win probability is raised to $\frac{2}{3}$ by switching doors. Allow me to expose the columnist's error: Suppose there are one hundred doors, and the contestant chooses for door number one. He then has a 1% probability of having chosen the correct door, and there is a 99% probability that the automobile is concealed behind one of the other ninety-nine doors. The host then proceeds to open all of the doors from 2 through 99. The automobile does not appear behind any of them, and it then becomes apparent that it must be behind either door number one or door number one hundred. According to the columnist's reasoning, door number one hundred now acquires a 99% probability of concealing the automobile. This, of course, is pure balderdash. What we actually have here is a new situation consisting of only two possibilities, each one being equally probable." Now, not only is this writer completely wrong, he also provides, unintentionally, an ironclad case in favor of changing doors. Another writer claims in his letter to vos Savant: "As a professional mathematician it concerns me to see a growing lack of mathematical proficiency among the general public. The probability in question must be $\frac{1}{2}$; I caution you in future to steer clear of issues of which you have no understanding."† Martin Gardner, the spiritual father of the Monty Hall problem, writes: "There is no other branch of mathematics in which experts can so easily blunder as in probability theory."

The fact that there was so much dissension over the correct solution to the Monty Hall dilemma can be explained by the psychological given that many people naturally tend to assign equal probabilities to the two remaining doors at the last stage of the game. Some readers of vos Savant's column may have thought that when the game show host promised to open a door, he meant that he would pick a door at random. Were this actually the case, it would not be to the contestant's advantage to switch doors later. But the quizmaster had promised to open a door concealing one of the gag prizes. This changes the situation and brings relevant, previously unknown information to light. At the beginning of the game when there are three doors to choose from, the contestant has a $\frac{1}{3}$ probability that the automobile will be hidden behind his chosen door, and a

† Many other reactions and a psychological analysis of those reactions can be found in Marilyn vos Savant's *The Power of Logical Thinking*, St. Martin's Press, New York, 1997.

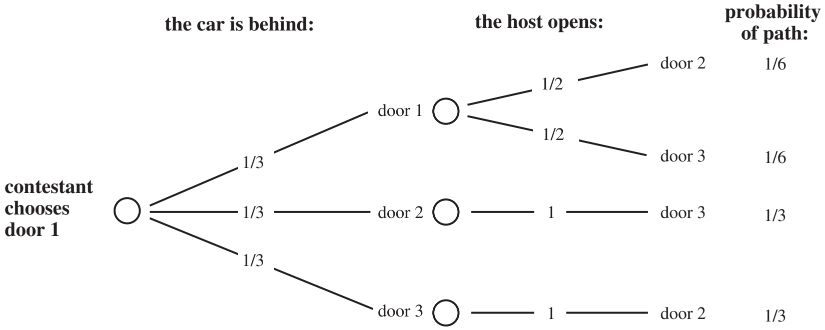


Fig. 6.1. Chance tree for the Monty Hall dilemma.

$\frac{2}{3}$ probability that it will not be behind his door. At this point, the host opens a door. The promise he makes at the outset is that after the contestant indicates his choice of door, and regardless of what that choice is (this is an essential given), the host will open one of the remaining doors without an automobile behind it. If the automobile is not behind the contestant's door, then it is in all certainty behind the door remaining after the host opens a door. In other words, there is a $\frac{2}{3}$ probability that switching doors at this stage will lead to the contestant's winning the automobile, while there is a $\frac{1}{3}$ probability that switching doors will not lead to the contestant's winning the automobile. Of course, it does rankle when contestants switch doors only to find that their original choice was the correct one.

6.1.1 Chance tree

The reasoning that leads to the correct answer of $\frac{2}{3}$ is simple, but you do have to get started along the right pathway. How can you reach the correct answer in a more systematic way without stumbling into a pattern of faulty intuitive thinking? One answer lies in computer simulation, another in the playing of a streamlined version of the game in which a ten-dollar bill is hidden under one of three coasters. A systematic approach using nothing but pencil and paper is also a possibility. This last option is carried out with the help of a chance tree. Chance trees make very clear that probabilities depend on available information. Figure 6.1 shows the chance tree for the Monty Hall problem. It shows all possible events with their corresponding probabilities. To make it as straightforward as possible, we have labeled the door first chosen by the contestant as door 1. The host's promise to open a door behind which there will be no automobile can also be seen in the chance tree. He will either open door

2 (if the automobile is behind door 3) or door 3 (if the automobile is behind door 2), and will open either door 2 or 3 randomly if the automobile is behind door 1. The numbers associated with the lines branching out from a node show the probability of the feasible events that may occur at that particular node. The probabilities of the possible pathways are calculated by multiplying the probabilities located at the various branches along the pathway. We can see by looking at the chance tree in Figure 6.1 that the two last pathways lead to the winning of the automobile. The probability of winning the automobile by switching doors is given by the sum of the probabilities of these two paths and is thus equal to $\frac{1}{3} + \frac{1}{3} = \frac{2}{3}$. And the correct answer is, indeed, $\frac{2}{3}$.

The Monty Hall dilemma clearly demonstrates how easy it is to succumb to faulty intuitive reasoning when trying to solve some probability problems. The same can be said of the following, closely related problem.

6.1.2 The problem of the three prisoners

Each of three prisoners A, B, and C is eligible for early release due to good behavior. The prison warden has decided to grant an early release to one of the three prisoners and is willing to let fate determine which of the three it will be. The three prisoners eventually learn that one of them is to be released, but do not know who the lucky one is. The prison guard does know. Arguing that it makes no difference to the odds of his being released, prisoner A asks the guard to tell him the name of one co-prisoner that will not be released. The guard refuses on the grounds that such information will raise prisoner A's release probability to $\frac{1}{2}$. Is the guard correct in his thinking, or is prisoner A correct? The answer is prisoner A when the guard names at random one of the prisoners B or C when both B and C are not released (if the guard names C only when he has no choice and if prisoner A knows this fact, then the situation becomes completely different). This is readily seen with a glance at the chance tree in Figure 6.2: both $P(A \text{ free} \mid \text{guard says } B)$ and $P(A \text{ free} \mid \text{guard says } C)$ are equal to $\frac{1}{6}/(\frac{1}{6} + \frac{1}{3}) = \frac{1}{3}$. Another way of arriving at the conclusion that the answer must be $\frac{1}{3}$ is to see this problem in the light of the Monty Hall problem. The to-be-freed prisoner is none other than the door with the automobile behind it. The essential difference between the two problems is that, in the prisoner's problem, there is no switching of doors/prisoners. If the contestant in the Monty Hall problem does not switch doors, the probability of his winning the automobile remains at $\frac{1}{3}$, even after the host has opened a door revealing no automobile!

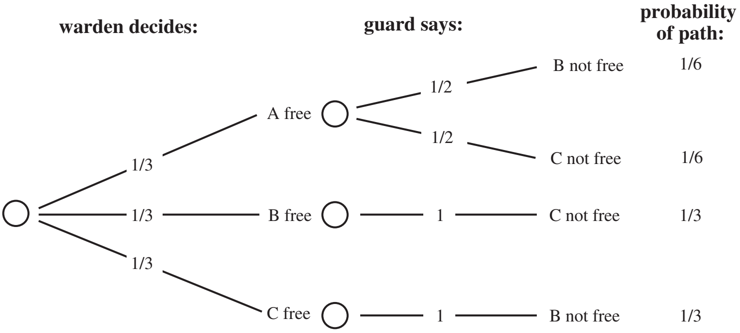


Fig. 6.2. Chance tree for the prisoner's problem.

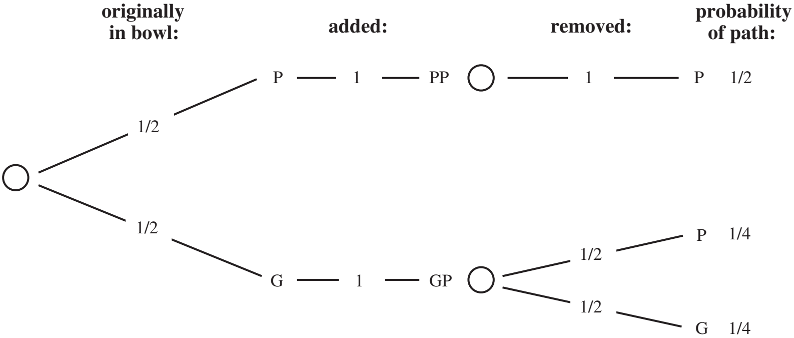


Fig. 6.3. Chance tree for the sushi delight problem.

6.1.3 Sushi delight

One fish is contained within the confines of an opaque fishbowl. The fish is equally likely to be a piranha or a goldfish. A sushi lover throws a piranha into the fish bowl alongside the other fish. Then, immediately, before either fish can devour the other, one of the fish is blindly removed from the fishbowl. The fish that has been removed from the bowl turns out to be a piranha. What is the probability that the fish that was originally in the bowl by itself was a piranha? This is another problem that can instigate heated discussions.

The correct answer to the question posed is $\frac{2}{3}$. This is easily seen from the chance tree in Figure 6.3. The first and second paths in the tree lead to the removal of a piranha from the bowl. The probability of occurrence of the first path is $\frac{1}{2} \times 1 \times 1 = \frac{1}{2}$, and the probability of occurrence of the second path is $\frac{1}{2} \times 1 \times \frac{1}{2} = \frac{1}{4}$. The desired probability of the fishbowl originally holding a piranha is the probability of occurrence of the first path given that the first or the second path has occurred. The definition of conditional probability gives

$$\begin{aligned} P(\text{path 1} \mid \text{path 1 or path 2}) &= \frac{P(\text{path 1})}{P(\text{path 1}) + P(\text{path 2})} \\ &= \frac{1/2}{1/2 + 1/4} = \frac{2}{3}. \end{aligned}$$

In the same way, it can be verified that the probability we are seeking is equal to

$$\frac{p}{p + \frac{1}{2}(1-p)}$$

if the fishbowl originally held a piranha with probability p and a goldfish with probability $1 - p$.

The sushi delight problem was originated by American scientist/writer Clifford Pickover and is a variant of the classic problem we will discuss at the end of this chapter in Problem 6.13.

6.1.4 Daughter-son problem

You are told that a family, completely unknown to you, has two children and that one of these children is a daughter. Is the chance of the other child being a daughter equal to $\frac{1}{2}$ or $\frac{1}{3}$? Are the chances altered if, aware of the fact that the family has two children only, you ring their doorbell and a daughter opens the door? In Section 2.9, computer simulation was used to obtain the answers $\frac{1}{3}$ and $\frac{1}{2}$ to the first and second questions. The assumption was made that each newborn child is equally likely to be a boy or a girl. In answering the second question, we also made the assumption that, randomly, one of the children will open the door. The answers $\frac{1}{3}$ and $\frac{1}{2}$ to the first and second questions can also be verified using a chance tree. We leave it to the reader to do so.

6.2 The test paradox

An inexpensive diagnostic test is available for a certain disease. Although the test is very reliable, it is not 100% reliable. If the test result for a given patient turns out to be positive, then further, more in-depth testing is called for to determine with absolute certainty whether or not the patient actually does suffer from the particular disease. Among persons who actually do have the disease, the test gives positive results in an average of 99% of the cases. For patients

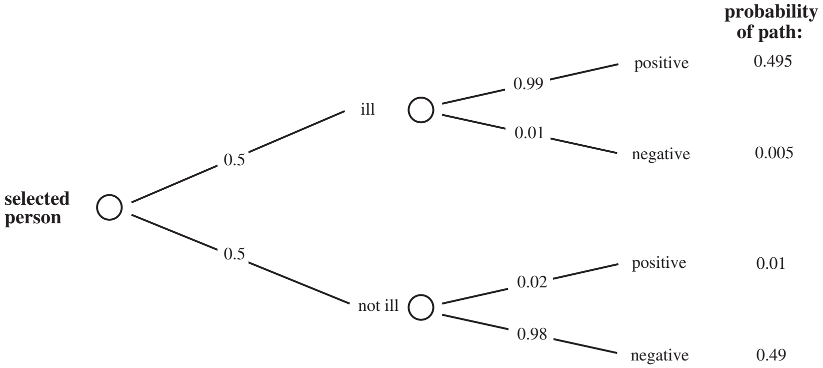


Fig. 6.4. Chance tree for the subgroup.

who do not have the disease, there is a 2% probability that the test will give a false-positive result. In one particular situation, the test is used at a polyclinic to test a subgroup of persons among whom it is known that one out of two has the disease. For a given person out of this subgroup, what is the probability that he will turn out to have the disease after having tested positively? To arrive at an answer, we must create a chance tree like the one shown in Figure 6.4. If we take the product of the probabilities along each pathway in the chance tree, we see that the first and third pathways lead to positive test results with probabilities 0.495 and 0.01, respectively. The probability that the person has the disease after testing positively is equal to the probability of the appearance of the first pathway given that the first or the third pathway has appeared. Next, the definition of conditional probability leads to

$$P(\text{path 1} \mid \text{path 1 or path 3}) = \frac{P(\text{path 1})}{P(\text{path 1}) + P(\text{path 3})}.$$

The probability we are seeking, then, is equal to

$$\frac{0.495}{0.495 + 0.01} = 0.9802.$$

In other words, in the subgroup, an average 98.0% of the positive test results are correct.

Let’s now suppose that, based on the success of the test, it is suggested that the entire population be tested for this disease on a yearly basis. Among the general population, an average of 1 out of one 1,000 persons has this disease. Is it a good idea to test everyone on a yearly basis? In order to answer this, we will calculate the probability of a randomly chosen person turning out to have the disease given that the person tests positively. To do this, we will refer to

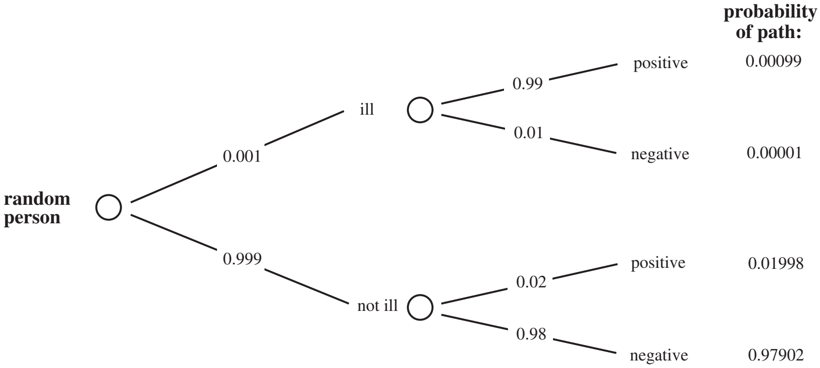


Fig. 6.5. Chance tree for the entire population.

the chance tree in Figure 6.5. This figure shows that the probability of a randomly chosen person having the disease given that the person tests positively is equal to

$$\frac{0.00099}{0.00099 + 0.01998} = 0.0472.$$

This leads to the seemingly paradoxical result that in an average of more than 95% of the cases that test positively, the persons in question do not actually have the disease. Considering this fact, many people would be unnecessarily distressed if the entire population were to be tested. An explanation for the fact that a reasonably reliable test works so unsatisfactorily for the entire population lies in the fact that the vast majority of the population does not have the disease. The result is that even though there is only a small probability of receiving a positive test result when one does not have the disease, the people among the general population who do not have the disease, by virtue of their sheer numbers, will nevertheless get a much larger number of positive results than the small group of people who are actually ill. In other words, the number of false-positives far outstrips the number of correct diagnoses when the entire population undergoes the test. This is underlined by the following reasoning. Suppose you test 10,000 randomly chosen people. There will be on average 9,990 people who do not have the illness, and ten people who do. This means, on average, $0.02 \times 9{,}990 = 199.8$ false-positives and $0.99 \times 10 = 9.9$ true-positives.

In the example above, we can see how important it is to keep an eye on the *basic proportions* between the various categories of people. If we ignore these proportions, we can end up coming to weird conclusions such as: "Statistics show that 10% of traffic accidents are caused by drunken drivers, which means that the other 90% are caused by sober drivers … is it then not sensible to allow only drunken drivers onto the roads?" This statement is attributed to M. Samford and should give politicians pause to refrain from making similar statements.

6.2.1 Bayes' rule[†]

The chance tree in Figure 6.5 describes uncertainties in a process that evolves over time. Initially, before the test is done, you have an estimate of 0.001 of the probability of the disease. After the test is done, you have a revised estimate of this probability. The former estimate is called the *prior* probability, and the latter estimate is called the *posterior* probability. An alternative method to calculate the posterior probability is Bayes' rule. The reasoning of this rule is based on a (subtle) use of conditional probabilities. Bayes' rule will be illustrated for the situation that the test is used for the whole population. In order to find the posterior probability of the disease given a positive test result, we first list the data

$$P(\text{disease}) = 0.001, \quad P(\text{no disease}) = 0.999,$$

$$P(\text{positive} \mid \text{disease}) = 0.99, \quad P(\text{negative} \mid \text{disease}) = 0.01,$$

$$P(\text{positive} \mid \text{no disease}) = 0.02, \quad P(\text{negative} \mid \text{no disease}) = 0.98.$$

The posterior probability $P(\text{disease} \mid \text{positive})$ satisfies the relation

$$P(\text{disease} \mid \text{positive}) = \frac{P(\text{positive and disease})}{P(\text{positive})}.$$

A repeated application of the definition of conditional probability gives

$$P(\text{positive and disease}) = P(\text{positive} \mid \text{disease})P(\text{disease})$$

and

$$\begin{aligned} P(\text{positive}) &= P(\text{positive and disease}) + P(\text{positive and no disease}) \\ &= P(\text{positive} \mid \text{disease})P(\text{disease}) \\ &\quad + P(\text{positive} \mid \text{no disease})P(\text{no disease}). \end{aligned}$$

[†] This rule is named after British parson Thomas Bayes (1702–1761), in whose posthumously published *Essay Toward Solving a Problem in the Doctrine of Chance*, an early attempt is made at establishing what we now refer to as Bayes' rule. However, it was Pierre Simon Laplace (1749–1827) who incorporated Bayes' work in the development of probability theory.

Consequently, the desired probability $P(\text{disease} \mid \text{positive})$ satisfies the formula

$$P(\text{disease} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{disease})P(\text{disease})}{P(\text{positive} \mid \text{disease})P(\text{disease}) + P(\text{positive} \mid \text{no disease})P(\text{no disease})}.$$

By filling in the above data, we find that

$$P(\text{disease} \mid \text{positive}) = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.02 \times 0.999} = 0.0472.$$

This is the same value as the one we found earlier.

The above derivation of the conditional probability $P(\text{disease} \mid \text{positive})$ is an illustration of Bayes' rule. It is possible to give a general mathematical formula for Bayes' rule. However, in specific applications, one better calculate the posterior probability according to Bayes' rule by using first principles as done in the above example.

Doctors should be more knowledgeable about chance trees and Bayes' formula. Consider the following situation. A doctor discovers a lump in a woman's breast during a routine physical exam. The lump could be a cancer. Without performing any further tests, the probability that the woman has breast cancer is 0.01. A mammogram is a test that, on average, is correctly able to establish whether a tumor is benign or cancerous 90% of the time. A positive test result indicates that a tumor is cancerous. What is the probability that the woman has breast cancer if the test result from a mammogram is positive? Results from a psychological study indicate that many doctors think that the probability $P(\text{cancer} \mid \text{positive})$ is slightly lower than the probability $P(\text{positive} \mid \text{cancer})$ and estimate the former probability as being about 80%. The actual value for the probability $P(\text{cancer} \mid \text{positive})$, however, is only 8.3% (verify)! A similar misconception sometimes occurs in court cases when the probability of innocence in an accused person with the same physical characteristics as the perpetrator is confused with the probability that a randomly selected member of the public looks like the perpetrator. Most such mistakes can be prevented by presenting the relevant information in terms of frequencies instead of probabilities. In the example of the mammogram test, the information might then consist of the fact that, of 1,000 women examined, there were ten who had cancer. Of these ten, eight had a positive mammogram, whereas of the 990 healthy women, 99 had a positive mammogram. Based on the information presented in this way, most doctors would then be able to correctly estimate the probability of breast cancer given a positive mammogram as being equal to $9/(9+99) \approx 8.3\%$.

Changing risk representations from probabilities into natural frequencies can turn the innumeracy of nonstatisticians into insight.[†]

6.3 Problems

6.1 The roads are safer at nonrush hour times than during rush hour because fewer accidents occur outside of rush hour than during the rush hour crunch. Do you agree or do you disagree?

6.2 On a table before you are two bowls containing red and white marbles; the first bowl contains seven red and three white marbles, and the second bowl contains 70 red and 30 white marbles. You are asked to select one of the two bowls, from which you will blindly draw two marbles (with no replacing of the marbles). You will receive a prize if at least one of the marbles you picked is white. In order to maximize your probability of winning the prize, do you choose the first bowl or the second bowl?

6.3 You are one of 50,000 spectators at a baseball game. Upon entering the ballpark, each spectator has received a ticket bearing an individual number. A winning number will be drawn from all of these 50,000 numbers. At a certain point, five numbers are called out over the loudspeaker. These numbers are randomly drawn and include the winning number. Your number is among the five numbers called. What is the probability of your ticket bearing the winning number?

6.4 Now consider the Monty Hall dilemma from Section 6.1 with the following difference: you learned beforehand that there is a 0.2 probability of the automobile being behind door 1, a 0.3 probability of its being behind door 2, and a 0.5 probability of its being behind door 3. Your strategy is to choose the door with the lowest probability (door 1) in the first round of the game, and then to switch doors to one with a higher probability after the host has opened a gag prize door. Set up a chance tree to determine your probability of winning the automobile.

6.5 Consider the Monty Hall dilemma with the following twists: there are five doors, and the host promises to open two of the gag prize doors after the contestant has chosen a door. Set up a chance tree to calculate the probability of the contestant winning the automobile by switching doors.

6.6 Consider the following variant of the Monty Hall dilemma. There are now four doors, behind one of which there is an automobile. You first indicate a door. Then the host opens another door behind which a gag prize is to be found. You are now given the opportunity to switch doors. Regardless of whether or not you switch, the host then opens another door (not the door of your current choice) behind which no automobile is to be found. You are now given a final opportunity to switch doors. What is the best strategy for playing this version of the game?

6.7 The final match of world championship soccer is to be played between England and the Netherlands. The star player for the Dutch team, Dennis Nightmare, has

[†] In his book *Calculated Risks* (Simon & Schuster, 2002) Gerd Gigerenzer advocates that doctors and lawyers be educated in more understandable representations of risk.

been injured. The probability of his being fit enough to play in the final is being estimated at 75%. Pre-game predictions have estimated that, without Nightmare, the probability of a Dutch win is 30% and with Nightmare, 50%. Later, you hear that the Dutch team has won the match. Without having any other information about events that occurred, what would you say was the probability that Dennis Nightmare played in the final?

6.8 Passers-by are invited to take part in the following sidewalk betting game. Three cards are placed into a hat. One card is red on both sides, one is black on both sides, and one is red on one side and black on the other side. A participant is asked to pick a card out of the hat at random, taking care to keep just one side of the card visible. After having picked the card and having seen the color of the visible side of the card, the owner of the hat bets the participant equal odds that the other side of the card will be the same color as the one shown. Is this a fair bet?

6.9 Alcohol checks are regularly conducted among drivers in a particular region. Drivers are first subjected to a breathtest. Only after a positive breathtest result is a driver taken for a blood test. This test will determine whether the driver has been driving under the influence of alcohol. The breathtest yields a positive result among 90% of drunken drivers and yields a positive result among only 5% of sober drivers. As it stands at present, a driver can only be required to do a breathtest after having exhibited suspicious driving behavior. It has been suggested that it might be a good idea to subject drivers to breathtests randomly. Current statistics show that one out of every 20 drivers on the roads in the region in question is driving under the influence. Calculate the probability of a randomly tested driver being unnecessarily subjected to a blood test after a positive breathtest.

6.10 You know that bowl A has three red and two white balls inside and that bowl B has four red and three white balls. Without your being aware of which one it is, one of the bowls is randomly chosen and presented to you. Blindfolded, you must pick two balls out of the bowl. You may proceed according to one of the following strategies:

(a) you will choose and replace (i.e., you will replace your first ball into the bowl before choosing your second ball).

(b) you will choose two balls without replacing any (i.e., you will not replace the first ball before choosing a second).

The blindfold is then removed and the colors of both of the balls you chose are revealed to you. Thereafter you must make a guess as to which bowl your two balls came from. For each of the two possible strategies, determine how you can make your guess depending on the colors you have been shown. Which strategy offers the higher probability for a correct guess as to which bowl the balls came from? Does the answer to this question contradict your intuitive thinking?

6.11 There are two taxicab companies in a particular city, "Yellow Cabs" and "White Cabs." Of all the cabs in the city, 85% are "Yellow Cabs" and 15% are "White Cabs." The issue of cab color has become relevant in a hit-and-run case before the courts in this city, in which witness testimony will be essential in determining the guilt or innocence of the cab driver in question. In order to test witness reliability, the courts have set up a test situation similar to the one occurring on the night of the hit-and-run accident. Results showed that 80% of the participants in the test case correctly identified the cab color, whereas 20% of the participants identified the wrong company. What is the probability that the accused hit-and-run cabbie is a "White Cabs" employee? (This problem is taken from the book of Kahneman et al.; see the footnote in Chapter 1.)

6.12 A doctor finds evidence of a serious illness in a particular patient and must make a determination about whether or not to advise the patient to undergo a dangerous operation. If the patient does suffer from the illness in question, there is a 95% probability that he will die if he does not undergo the operation. If he does undergo the operation, he has a 50% probability of survival. If the operation is conducted and it is discovered that the patient does not suffer from the illness, there is a 10% probability that the patient will die due to complications resulting from the operation. If it has been estimated that there is a 20% to 30% probability of the patient actually having the illness in question, how should the doctor advise her patient?

6.13 Consider the sushi delight problem from Section 6.1. Suppose now that both a piranha and a goldfish are added to the fishbowl alongside the original fish. What is the probability that the original fish is a piranha if a piranha is taken out of the bowl?

6.14 Suppose that there is a DNA test that determines with 100% accuracy whether or not a particular gene for a certain disease is present. A woman would like to do the DNA test, but wants to have the option of holding out hope that the gene is not present in her DNA even if it is determined that the gene for the illness is, indeed, present. She makes the following arrangement with her doctor. After the test, the doctor will toss a fair coin into the air, and will tell the woman the test results only if those results are negative and the coin has turned up heads. In every other case, the doctor will not tell her the test results. Suppose that there is a 1 out of 100 probability that the woman does have the gene for the disease in question before she is tested. What is the revised value of this probability if the woman's doctor does not inform her over the test results? (Marilyn vos Savant, *Parade* magazine, February 7, 1999).

6.15 At a particular airport, each passenger must pass through a special fire arms detector. An average of 1 out of every 100,000 passengers is carrying a fire arm. The detector is 100% accurate in the detection of fire arms, but in an average of 1 in 10,000 cases, it results in a false alarm while the passenger is not carrying a fire arm. In cases when the alarm goes off, what is the probability that the passenger in question is carrying a fire arm?

6.16 A sum of money is placed in each of two envelopes. The amounts differ from one another, but you do not know what the values of the two amounts are. You do know that the values lie between two boundaries m and M with $0 < m < M$. You choose an envelope randomly. After inspecting its contents, you may switch envelopes. Set up a chance tree to verify that the following procedure will give you a probability of greater than $\frac{1}{2}$ of winding up with the envelope holding the most cash.

(a) Choose an envelope and look to see how much cash is inside.

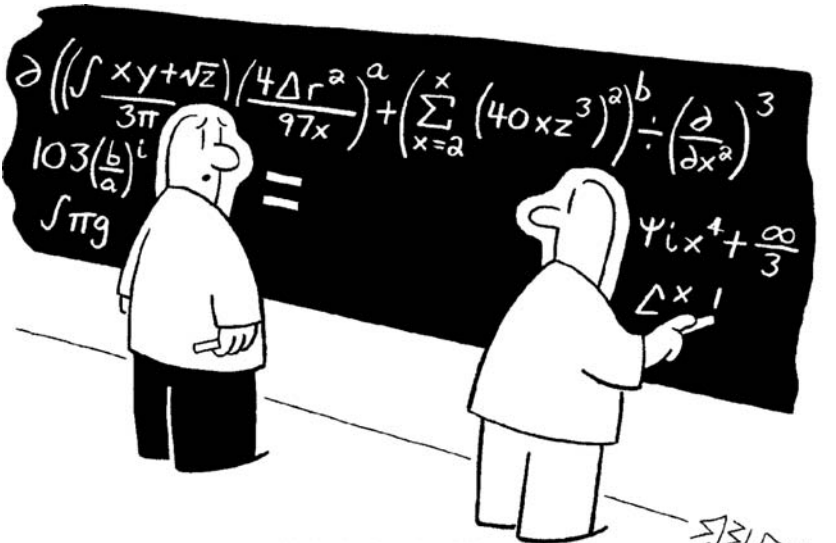
(b) Pick a random number between m and M.

(c) If the number you drew is greater than the amount of cash in your envelope, you exchange the envelope. Otherwise, you keep the envelope you have.

6.17 In a television game show, you can win 10,000 dollars by guessing the composition of red and white marbles contained in a nontransparent vase. The vase contains a very large number of marbles. You must guess whether the vase has twice as many red marbles as white ones, or whether it has twice as many white ones as red ones. Beforehand, both possibilities are equally likely to you. To help you guess, you are given a one-time opportunity of picking one, two, or three marbles out of the vase. This action, however, comes at the expense of the 10,000 dollar prize money. If you opt to choose one marble out of the vase, $750 will be subtracted from the $10,000 should you win. Two marbles will cost you $1,000 and three marbles will cost you $1,500. Set up a chance tree to determine which strategy will help you maximize your winnings.

PART TWO

Essentials of probability


∂((∫ xy+√z / 3π)(4Δr² / 97x)^a + (Σ_{x=2}^{x} (40xz³)²)^b ÷ (∂ / ∂x²)³
103(b/a)^i
∫πg
=
Ψix⁴ + ∞/3
Δ× 1
S.BURNS

"Little help?"

7
Foundations of probability theory

Constructing the mathematical foundations of probability theory has proven to be a long-lasting process of trial and error. The approach consisting of defining probabilities as relative frequencies in cases of repeatable experiments leads to an unsatisfactory theory. The frequency view of probability has a long history that goes back to Aristotle. It was not until 1933 that the great Russian mathematician Andrej Nikolajewitsch Kolmogorov (1903–1987) laid a satisfactory mathematical foundation of probability theory. He did this by taking a number of axioms as his starting point, as had been done in other fields of mathematics. Axioms state a number of minimal requirements that the mathematical objects in question (such as points and lines in geometry) must satisfy. In the axiomatic approach of Kolmogorov, probability figures as a function on subsets of a sample space. The axioms are the basis for the mathematical theory of probability. As a milestone, the law of large numbers can be deduced from the axioms by logical reasoning. The law of large numbers confirms our intuition that the probability of an event in a repeatable experiment can be estimated by the relative frequency of its occurrence in many repetitions of the experiment. This law, which has already been discussed in Chapter 2, is the fundamental link between theory and the real world. The purpose of this chapter is to discuss the axioms of probability theory in more detail and to derive from the axioms the most basic rules for the calculation of probabilities. These rules include the addition rule and the more general inclusion-exclusion rule. Various examples will be given to illustrate the rules.

7.1 Probabilistic foundations

A probability model for a chance experiment consists of a complete description of all possible outcomes of the experiment and an assignment of probability to these outcomes. The set of all possible outcomes of the experiment is called the *sample space*. A sample space is always such that one and only one of the possible outcomes occurs if the experiment is performed. The classic example is the experiment of tossing a coin. Then, the sample space consists of the two outcomes H and T, where H means that the outcome of the toss is a head and T that it is a tail. Each of the two outcomes gets assigned a probability of $\frac{1}{2}$ if the coin is fair. Another example is the experiment of rolling a die. The sample space is the set $\{1, 2, \ldots, 6\}$, where the outcome i means that i dots appear on the up face. Each of the six outcomes get assigned a probability of $\frac{1}{6}$, assuming that the die is unbiased. For the experiment of taking at random one coin from your pocket with two dimes and three quarters, the sample space consists of the two outcomes D and Q that get assigned the probabilities $\frac{2}{5}$ and $\frac{3}{5}$. Before we formulate the axioms of probability, we first introduce some concepts from set theory.

7.1.1 Countable and uncountable sets

The set of natural numbers (positive integers) is an infinite set and is the prototype of a countably infinite set. In general, a nonfinite set is called *countable* if a one to one function exists which maps the elements of the set to the set of natural numbers. In other words, every element of the set can be assigned to a unique natural number and conversely each natural number corresponds to a unique element of the set. For example, the set of squared numbers $1, 4, 9, 16, 25, \ldots$ is countable. Not all sets with an infinite number of elements are countably infinite. The set of all points on a line and the set of all real numbers between 0 and 1 are examples of infinite sets that are not countable. The German mathematician Georg Cantor (1845–1918) proved this result in the nineteenth century. This discovery represented an important milestone in the development of mathematics and logic (the concept of infinity, to which even scholars from ancient Greece had devoted considerable energy, obtained a solid theoretical basis for the first time through Cantor's work). Sets that are neither finite nor countably infinite are called *uncountable*. Chance experiments with either countably infinite or uncountable sample spaces are common.

Example 7.1 Consider the chance experiment consisting of the number of tosses of a fair coin needed to obtain three heads in a row. The sample space of this experiment is the set of integers $3, 4, 5, \ldots$, where the outcome k indicates that three heads in a row was first achieved after k coin tosses. This sample space is countably infinite. It will be seen in Example 7.8 that an outcome corresponding to a sequence of tosses in which three heads in a row never occurs need not be included in the sample space.

Example 7.2 Consider the experiment in which a random point in a circle with radius R is chosen by a blindfolded person throwing a dart at a dartboard. The sample space of this experiment consists of the set of pairs of real numbers (x, y) where $x^2 + y^2 \leq R^2$. This sample space is uncountable.

Example 7.3 The experiment in which the number of alpha-particles emitted by a radioactive source in a fixed time interval is counted has the countably infinite set of the nonnegative integers as sample space. The sample space of the experiment in which the time until the first emission of a particle is measured is the uncountable set of the positive real numbers.

7.1.2 Axioms of probability theory

The axioms of probability for an experiment with a finite or countably infinite sample space are the same as those for one with an uncountable sample space. A distinction must be made, however, between the sorts of subsets to which probabilities can be assigned, whether these subsets occur in countable or uncountable sample spaces. In the case of a finite or countably infinite sample space, probabilities can be assigned to each subset of the sample space. In the case of an uncountable sample space, weird subsets can be constructed to which we cannot associate a probability. These technical matters will not be discussed in this introductory book. The reader is asked to accept the fact that, for more fundamental mathematical reasons, probabilities can only be assigned to sufficiently well-behaved subsets of an uncountable sample space. In the case that the sample space is the set of real numbers, then essentially only those subsets consisting of a finite interval, the complement of each finite interval, and the union of each countable number of finite intervals are assigned a probability. These subsets suffice for practical purposes. If the probability measure on the sample space is denoted by P, then P must satisfy the following properties

Axiom 7.1. $P(A) \geq 0$ *for each subset A.*

Axiom 7.2. $P(A) = 1$ *when A is equal to the sample space.*

Axiom 7.3. $P\big(\bigcup_{i=1}^{\infty} A_i\big) = \sum_{i=1}^{\infty} P(A_i)$ *for every collection of pairwise disjoint subsets* $A_1, A_2, \ldots.$

The notation $\bigcup_{i=1}^{\infty} A_i$ indicates the set of all outcomes which belong to at least one of the subsets $A_1, A_2, \ldots.$ The subsets $A_1, A_2, \ldots$ are said to be *pairwise disjoint* when any two subsets have no element in common. The first two axioms simply express a probability as a number between 0 and 1. The crucial axiom 7.3 states that, for any sequence of *mutually exclusive* events, the probability of at least one of these events occurring is the sum of their individual probabilities. In probability terms, any subset of the sample space is called an *event*. If the outcome of the chance experiment belongs to A, the event A is said to *occur*. The events $A_1, A_2, \ldots$ are said to be *mutually exclusive* if the corresponding sets $A_1, A_2, \ldots$ are pairwise disjoint.

The entire theory of probability is built on the above three axioms. For a finite sample space the axioms are equivalent to those given in Section 2.2. An immediate consequence of the above axioms is that

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

for any finite sequence of pairwise disjoint sets $A_1, \ldots, A_n$. A formal proof of this (obvious) result will be postponed to Section 7.3.

The standard notation for the sample space is the symbol Ω. An element in Ω is denoted by ω. A sample space together with an assignment of probabilities to events is called a *probability space*. As has already been pointed out in Chapter 2, for a finite or countably infinite sample space Ω, it is sufficient to assign a probability $p\,(\omega)$ to each element $\omega \in \Omega$ where $p\,(\omega) \geq 0$ and $\sum_{\omega \in \Omega} p\,(\omega) = 1$. A probability measure P on Ω is then defined by specifying the probability of each subset A of Ω as

$$P(A) = \sum_{\omega \in A} p\,(\omega)$$

(the notation $\omega \in A$ means that ω belongs to the set A and $\sum_{\omega \in A} p(\omega)$ is the notation for the sum of all $p(\omega)$'s with $\omega \in A$). It is left to the reader to verify P satisfies the Axioms 7.1 to 7.3.

Example 7.4 John, Pedro and Rosita each roll one fair die. How do we calculate the probability that the score of Rosita is equal to the sum of the scores of John and Pedro?

Solution. The sample space of the chance experiment is taken as $\Omega = \{(i, j, k)$ $i, j, k = 1, \ldots, 6\}$, where the outcome (i, j, k) occurs if the score of John is i dots, the score of Pedro is j dots, and the score of Rosita is k dots. Each of the 216 possible outcomes is equally probable and thus gets assigned a probability mass of $\frac{1}{216}$. The score of Rosita is equal to the sum of the scores of John and Pedro if one of the 15 outcomes (1,1,2), (1,2,3), (2,1,3), (1,3,4), (3,1,4), (2,2,4), (1,4,5), (4,1,5), (2,3,5), (3,2,5), (1,5,6), (5,1,6), (2,4,6), (4,2,6), (3,3,6) occurs. The probability of this event is thus $\frac{15}{216}$.

Example 7.5 A certain family has four children. Does the family more likely consist of two children of each sex than of three children of one sex and one of the other?

Solution. At first thought one might imagine that it is most likely that the family consists of two children of each sex. After all, on average half of the newborn children in the population are boys and the other half are girls. But the solution is not that straightforward. By constructing a sample space for the experiment, you can think precisely about the events. The sample space consists of the 16 elements $BBBB$, $BBBG$, $BBGB$, $BGBB$, $GBBB$, $BBGG$, $GGBB$, $BGGB$, $GBBG$, $BGBG$, $GBGB$, $GGGB$, $GGBG$, $GBGG$, $BGGG$, and $GGGG$. Assuming that an unborn child has 50% probability of being a female and assuming independence of the sex of newborns, each element of the sample space is equally probable and gets assigned a probability of $\frac{1}{16}$. There are six possible outcomes for which the family has two children of each sex and so this event has probability $\frac{3}{8}$. There are eight possible outcomes for which the family has three children of one sex and one of the other, and so this event has probability $\frac{1}{2}$. Hence, it is more likely that the family has three children of one sex and one of the other than that it has two children of each sex. This conclusion has been reached under the assumption that the probability of a newborn child being a girl is the same as the probability of its being a boy. In reality, the probability of a child being born female is slightly under 50% (see Example 5.4 in Chapter 5), but the conclusion remains valid.

The next two examples illustrate the choice of a probability measure for an uncountable sample space.

Example 7.2 (continued). How do we calculate the probability of the dart hitting the bull's-eye?

Solution. The assumption of the dart hitting the dartboard at a random point is translated by assigning the probability

$$P(A) = \frac{\text{the area of the region } A}{\pi R^2}$$

to each subset A of the sample space. If the bull's-eye of the dartboard has radius b, the probability of the dart hitting the bull's-eye is $\pi b^2/(\pi R^2) = b^2/R^2$. The following observation is made. The probability that the dart will hit a *prespecified* point is zero. It only makes sense to speak of the probability of hitting a given region of the dartboard. This observation expresses a fundamental difference between a probability model with a finite or countably infinite sample space and a probability model with an uncountable sample space.

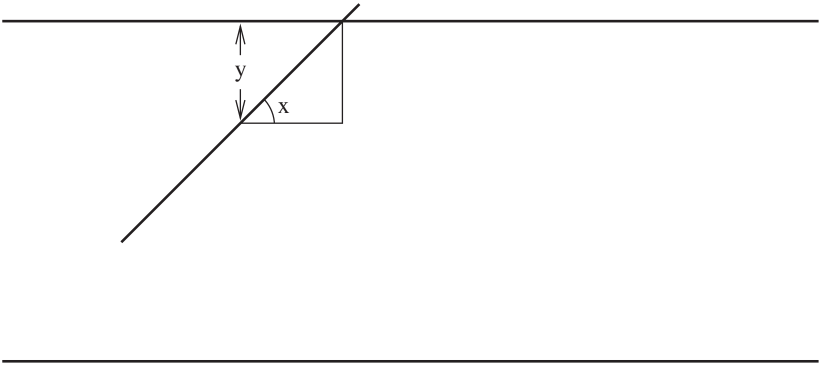


Fig. 7.1. The landing of Buffon's needle.

Example 7.6 A floor is ruled with equally spaced parallel lines a distance D apart. A needle of length L is dropped at random on the floor. It is assumed that $L \leq D$. What is the probability that the needle will intersect one of the lines? This problem is known as Buffon's needle problem.

Solution. This geometric probability problem can be translated into the picking of a random point in a certain region. Let y be the distance from the center of the needle to the closest line and let x be the angle at which the needle falls, where x is measured against a line parallel to the lines on the floor; see Figure 7.1. The sample space of the experiment can be taken as the rectangle R consisting of the points (x, y) with $0 \leq x \leq \pi$ and $0 \leq y \leq \frac{1}{2}D$. The needle will land on a line only if the hypotenuse of the right-angled triangle in Figure 7.1 is less than half of the length L of the needle. That is, we get an intersection only if $\frac{y}{\sin(x)} < \frac{1}{2}L$. Thus, the probability that the needle will intersect one of the lines equals the probability that a point (x, y) chosen at random in the rectangle R satisfies $y < \frac{1}{2}L\sin(x)$. In other words, the area under the curve $y = \frac{1}{2}L\sin(x)$ divided by the total area of the rectangle R gives the probability of an intersection. This ratio is

$$\frac{\int_0^\pi \frac{1}{2}L\sin(x)\,dx}{\frac{1}{2}\pi D} = \frac{-L\cos(x)}{\pi D}\bigg|_0^\pi$$

and so

$$P(\text{needle intersects one of the lines}) = \frac{2L}{\pi D}.$$

Problem 7.1 Two players A and B each roll one die. The absolute difference of the outcomes is computed. Player A wins if the difference is 0, 1, or 2; otherwise, player B wins. What is the probability that player A wins?

Problem 7.2 Independently of each other, two people think of a number between 1 and 10. What is the probability that five or more numbers will separate the two numbers chosen at random by the two people?

Problem 7.3 Sixteen bridge teams including the teams Johnson and Smith participate in a tournament. The tournament is organized as a knock-out tournament and has four rounds. The 16 teams are evenly matched. In each round the remaining teams are paired by drawing lots.

(a) What is the probability that the teams Johnson and Smith will meet in the first round?

(b) What is the probability that these two teams will meet in the final?[†]

Problem 7.4 Three friends go to the cinema together on a weekly basis. Before buying their tickets, all three friends toss a fair coin into the air once. If one of the three gets a different outcome than the other two, that one pays for all three tickets; otherwise, everyone pays his own way. Set up a probability model to calculate the probability that one of the three friends will have to pay for all three tickets. What is the probability that one of the three friends pays for all the tickets?

Problem 7.5 The game of franc-carreau was a popular game in eighteenth century France. In this game, a coin is tossed on a chessboard. The player wins if the coin does not fall on one of the lines of the board. Suppose now that a round coin with a diameter of d is blindly tossed on a large table. The surface of the table is divided into squares whose sides measure a in length, such that $a > d$. Define an appropriate probability space and calculate the probability of the coin falling entirely within the confines of a square. *Hint*: consider the position of the coin's middle point.

Problem 7.6 Two people have agreed to meet at the train station between 12.00 and 1.00 p.m. Independently of one another, each person is to appear at a completely random moment between the hours of 12.00 and 1.00. What is the probability that the two persons will meet within 10 minutes of one another?

Problem 7.7 The numbers B and C are chosen at random between –1 and 1, independently of each other. What is the probability that the quadratic equation $x^2 + Bx + C = 0$ has real roots? Also, derive a general expression for this probability when B and C are chosen at random from the interval $(-q, q)$ for any $q > 0$.

[†] The reader is assumed to be familiar with binomial coefficients. Combinatorics for probability is discussed in the Appendix.

Problem 7.8 A dart is thrown at random on a rectangular board. The board measures 20 cm by 50 cm. A hit occurs if the dart lands within 5 cm of any of the four corner points of the board. What is the probability of a hit?

Problem 7.9 A point is chosen at random inside a triangle with height h and base of length b. What is the probability that the perpendicular distance from the point to the base is larger than d? What is the probability that the randomly chosen point and the base of the triangle will form a triangle with an obtuse angle when the original triangle is equilateral?

7.1.3 Continuity property of probability

Probability is a continuous set function. To explain this property, consider a nondecreasing sequence of sets $E_1,\ E_2,\ldots$. The sequence $E_1,\ E_2,\ldots$ is said to be *nondecreasing* if the set E_{n+1} contains the set E_n for all $n \geq 1$. Let's define the set E by $E = \bigcup_{i=1}^{\infty} E_i$ and denote this set by $E = \lim_{n\to\infty} E_n$. Then, the continuity property states that

$$\lim_{n\to\infty} P(E_n) = P\big(\lim_{n\to\infty} E_n\big).$$

The proof is instructive. Define $F_1 = E_1$ and let the set F_{n+1} consist of the points of E_{n+1} that are not in E_n for $n \geq 1$. It is readily seen that the sets $F_1,\ F_2,\ldots$ are pairwise disjoint and satisfy $\bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i$ $(= E_n)$ for all $n \geq 1$ and $\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} E_i$. Thus,

$$\begin{aligned}
P\big(\lim_{n\to\infty} E_n\big) &= P\left(\bigcup_{i=1}^{\infty} E_i\right) = P\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} P(F_i) = \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} F_i\right) \\
&= \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} E_i\right) = \lim_{n\to\infty} P(E_n),
\end{aligned}$$

proving the continuity property. The proof uses Axiom 7.3 in the third and fifth equalities (Axiom 7.3 implies that $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$ for any finite sequence of pairwise disjoint sets $A_1,\ldots,A_n$; (see Rule 7.1 in Section 7.3).

The result $\lim_{n\to\infty} P(E_n) = P(\lim_{n\to\infty} E_n)$ holds also for a *nonincreasing* sequence of sets $E_1,\ E_2,\ldots$ (E_{n+1} is contained in E_n for all $n \geq 1$), in which case the set $\lim_{n\to\infty} E_n$ is defined as the intersection of all sets E_i for $i \geq 1$. The intersection is the set of all outcomes that belong to each of the sets E_i.

Problem 7.10 Use the axioms to prove the following results:

(a) $P(A) \leq P(B)$ if the set A is contained in the set B.

(b) $P\big(\cup_{k=1}^{\infty} A_k\big) \leq \sum_{k=1}^{\infty} P(A_k)$ for any sequence of subsets $A_1, A_2, \ldots$ (this result is known as *Boole's inequality*).

Problem 7.11 Let $A_1, A_2, \ldots$ be an infinite sequence of subsets with the property that $\sum_{k=1}^{\infty} P(A_k) < \infty$. Define the set C as $C = \{\omega\colon \omega \in A_k$ for infinitely many $k\}$. Use the continuity property of probabilities to prove that $P(C) = 0$ (this result is known as the *Borel-Cantelli lemma*).

7.2 Compound chance experiments

A chance experiment is called a *compound* experiment if it consists of several elementary chance experiments. In Section 2.2, several examples were given of compound experiments along with the corresponding probability spaces. The question arises as to how, in general, we define a probability space for a compound experiment in which the elementary experiments are physically independent of each other. By physically independent, we mean that the outcomes from any one of the elementary experiments have no influence on the functioning or outcomes of any of the other elementary experiments. We first answer the question for the case of a finite number of physically independent elementary experiments $\varepsilon_1, \ldots, \varepsilon_n$. Assume that each experiment ε_k has a finite or countable sample space Ω_k on which the probability measure P_k is defined such that the probability $p_k(\omega_k)$ is assigned to each element $\omega_k \in \Omega_k$. The sample space of the compound experiment is then given by the set Ω consisting of all $\omega = (\omega_1, \ldots, \omega_n)$, where $\omega_k \in \Omega_k$ for $k = 1, \ldots, n$. A natural choice for the probability measure P on Ω arises by assigning the probability $p(\omega)$ to each element $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ by using the *product rule*

$$p(\omega) = p_1(\omega_1) \times p_2(\omega_2) \times \cdots \times p_n(\omega_n).$$

This choice for the probability measure is not only intuitively the obvious one, but we can also prove that it is the only probability measure satisfying property $P(AB) = P(A)P(B)$ when the elementary experiments that generate event A are physically independent of those elementary experiments that give rise to event B. This important result of the uniqueness of the probability measure satisfying this property justifies the use of the product rule for compound chance experiments.

Example 7.7 In the "Reynard the Fox" café, it normally costs \$3.50 to buy a pint of beer. On Thursday nights, however, customers pay \$0.25, \$1.00, or \$2.50 for the first pint. In order to determine how much they will pay, customers must throw a dart at a dartboard that rotates at high speed. The dartboard is divided into eight segments of equal size. Two of the segments read \$0.25, four of the segments read \$1, and two more of the segments read \$2.50. You pay whatever you hit. Two friends, John and Peter, each throw a dart at the board and hope for the best. What is the probability that the two friends will have to pay no more than \$2 between them for their first pint?

Solution. The sample space of the experiment consists of the nine outcomes $(L, L), (L, M), (M, L), (L, H), (H, L), (M, M), (M, H), (H, M)$, and (H, H), where L stands for hitting a low-priced segment, M stands for hitting a medium-priced segment, H stands for hitting a high-priced segment and the first (second) component of each outcome refers to the throw of John (Peter). Assuming that, independently, the two darts hit the dartboard at a random point, the probability $\frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$ is assigned to each of the outcomes (L, L), (L, H), (H, L), and (H, H), the probability $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ to the outcome (M, M), and the probability $\frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$ to each of the outcomes (L, M), (M, L), (H, M), and (M, H). The two friends will have to pay no more than \$2 between them for their first pint if one of the four outcomes (L, L), (L, M), (M, L), (M, M) occurs. The probability of this event is thus $\frac{9}{16}$.

Example 7.8 Two desperados play Russian roulette in which they take turns pulling the trigger of a six-cylinder revolver loaded with one bullet (after each pull of the trigger, the magazine is spun to randomly select a new cylinder to fire). What is the probability that the desperado who begins will be the one to shoot himself dead?

Solution. The sample space we use for this chance experiment is the set

$$\Omega = \{F, MF, MMF, \ldots\} \cup \{MM \ldots\},$$

where the element $M \ldots MF$ with the first $n - 1$ letters all being M represents the event that the first $n - 1$ times that the trigger is pulled, no shot is fired, and that the fatal shot is fired on the nth attempt. The element $MM \ldots$ represents the event that no fatal shot is fired when the two desperados repeatedly pull the trigger without ever stopping. Formally, this element should be included in the sample space. In accordance with our intuition, it will be seen below that the element $MM \ldots$ is in fact superfluous. By representing the element $M \ldots MF$ with the first $n - 1$ letters all being M by the integer n, and the element $MM \ldots$ by the symbol ∞, we can also express the sample space Ω as

$$\Omega = \{1, 2, \ldots\} \cup \{\infty\}.$$

This representation of the sample space is the one we will use. Given the fact that the outcomes of the pulling of the trigger are independent of one another, and that with each pull of the trigger there is a probability of $\frac{1}{6}$ that the fatal shot will be fired, it is reasonable to assign the probability

$$p(n) = \left(\frac{5}{6}\right)^{n-1} \frac{1}{6} \qquad \text{for } n = 1, 2, \ldots$$

to the element n from the sample space for $n \in \{1, 2, \ldots\}$. To complete the specification of the probability measure P, we have to assign a probability to the element ∞. The probabilities $p(n)$ satisfy $\sum_{n=1}^{\infty} p(n) = 1$ because the geometric series $\sum_{k=1}^{\infty} x^k$ sums to $\frac{1}{1-x}$ for all $0 < x < 1$. Axiom 7.2 then implies that P must assign the value 0 the element ∞. If we now define A as the event that the fatal shot is fired by the desperado who begins, then $P(A)$ is given by

$$P(A) = \sum_{n=0}^{\infty} p(2n+1) = \sum_{n=0}^{\infty} \left(\frac{5}{6}\right)^{2n} \frac{1}{6}$$
$$= \frac{1}{6} \sum_{n=0}^{\infty} \left(\frac{25}{36}\right)^{n} = \frac{1}{6} \left(\frac{1}{1 - \frac{25}{36}}\right) = 0.5436.$$

Problem 7.12 In a tennis tournament between three players A, B, and C, each player plays the others once. The strengths of the player are as follows: $P(A \text{ beats } B) = 0.5$, $P(A \text{ beats } C) = 0.7$, and $P(B \text{ beats } C) = 0.4$. Assuming independence of the match results, calculate the probability that player A wins at least as many games as any other player.

Problem 7.13 Two people take turns selecting a ball at random from a bowl containing three white balls and seven red ones. The winner is the person who is the first to select a white ball. It is assumed that the balls are selected with replacement. Define an appropriate sample space and calculate the probability that the person who begins will win.

Problem 7.14 In repeatedly rolling two dice, what is the probability of getting a total of 6 before a total of 7? What about an 8 and a 7? What is the probability of getting a total of 6 and a total of 8 in any order before two 7's?

7.2.1 A coin-tossing experiment†

When a compound chance experiment consists of an infinite number of independent elementary chance experiments, it has an uncountable sample space and the choice of an appropriate probability measure is less obvious. We illustrate how we deal with such experiments by way of an illustration of a compound

† This section can be skipped without loss of continuity.

experiment consisting of an infinite number of tosses of a fair coin. We model the sample space of this experiment using all infinite sequences $\omega = (\omega_1, \omega_2, \ldots)$, where ω_i is equal to H when the ith coin toss comes up heads, and is equal to T otherwise. It can be proved that this sample space is uncountable. In order to be able to define a probability measure on this sample space, we must begin by restricting our attention to a class of appropriately chosen subsets. The so-called cylinder sets form the basis of this class of subsets. In the case of our chance experiment, a *cylinder set* is the set of all outcomes ω where the first n elements $\omega_1, \ldots, \omega_n$ have specified outcomes for finite values of n. A natural choice for the probability measure on the sample space is to assign the probability $P^{(\infty)}(A) = \left(\frac{1}{2}\right)^n$ to each cylinder set A with n specified elements. In this way, the event that heads first occurs at the kth toss can be represented by the cylinder set A_k whose ω's have the finite beginning $T, T, \ldots, T, H$, and can be assigned a probability of $\left(\frac{1}{2}\right)^k$. The collection $\bigcup_{k=1}^{\infty} A_k$ represents the event that at some point heads occurs. The probability measure on the class of cylinder sets can be extended to one defined on a sufficiently general class of subsets capable of representing all possible events of this chance experiment.

In Section 2.1, we stated that the fraction of coin-tosses in which heads occurs converges to $\frac{1}{2}$ with probability 1 when the number of tosses increases without limit. We are now in a position to state this claim more rigorously with the help of the probability measure $P^{(\infty)}$. To do so, we adopt the notation $K_n(\omega)$ to represent the number of heads occurring in the first n elements of ω. Furthermore let C be the collection of all outcomes ω for which $\lim_{n\to\infty} K_n(\omega)/n = \frac{1}{2}$. For very many sequences ω, the number $K_n(\omega)/n$ does not converge to $\frac{1}{2}$ as $n \to \infty$ (e.g., this is the case for any sequence ω with finitely many H's). However, "nature" chooses a sequence from the collection C according to $P^{(\infty)}$: the *theoretical* (*strong*) *law of large numbers* states that the probability $P^{(\infty)}$ measure assigns a probability of 1 to the collection C. In mathematical notation, the result is

$$P^{(\infty)}\left(\left\{\omega : \lim_{n\to\infty} \frac{K_n(\omega)}{n} = \frac{1}{2}\right\}\right) = 1.$$

This type of convergence is called *convergence with probability one*. The strong law of large numbers is of enormous importance: it provides a direct link between theory and practice. It was a milestone in probability theory when around 1930 A. N. Kolmogorov proved this law from the simple axioms of probability theory. In general, the proof of the strong law of large numbers requires advanced mathematics beyond the scope of this book. However, for the special case of the coin-tossing experiment, an elementary proof can be given by using the Borel-Cantelli lemma (see also Section 14.2.2 in Chapter 14).

7.3 Some basic rules

The axioms of probability theory directly imply a number of basic rules that are useful for calculating probabilities. We first repeat some basic notation. The event that at least one of events A or B occurs is called the *union* of A and B and is written $A \cup B$. The event that both A and B occur is called the *intersection* of A and B and is written $A \cap B$, or simply AB. The notation AB for the intersection of events A and B will be used throughout this book. The notation for union and intersection of two events extends to finite sequences of events. Given events $A_1, \ldots, A_n$, the event that at least one occurs is written $A_1 \cup A_2 \cup \cdots \cup A_n$, and the event that all occur is written $A_1 A_2 \cdots A_n$.

Rule 7.1. *For any finite number of mutually exclusive events $A_1, \ldots, A_n$*

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

Rule 7.2. *For any event A*

$$P(A) = 1 - P(A^c),$$

where the event A^c consists of all outcomes that are not in A.

Rule 7.3. *For any two events A and B*

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

Rule 7.4. *For any finite number of events $A_1, \ldots, A_n$*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{\substack{i,j:\\ i<j}} P(A_i A_j) + \sum_{\substack{i,j,k:\\ i<j<k}} P(A_i A_j A_k) - \cdots$$
$$+ (-1)^{n-1} P(A_1 A_2 \cdots A_n).$$

The proofs of these rules are simple and instructive and nicely demonstrate how useful propositions can be obtained from "minimal" axioms.

To prove Rule 7.1, denote by $\emptyset$ the empty set of outcomes (`null event`). We first show that

$$P(\emptyset) = 0.$$

Applying Axiom 7.3 with $A_i = \emptyset$ for $i = 1, 2, \ldots$ gives $P(\emptyset) = \sum_{i=1}^{\infty} a_i$, where $a_i = P(\emptyset)$ for each i. This implies that $P(\emptyset) = 0$. Let $A_1, \ldots, A_n$

be any finite sequence of pairwise disjoint sets. Augment this sequence with $A_{n+1} = \emptyset$, $A_{n+2} = \emptyset, \ldots$. Then, by Axiom 7.3

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right) = \sum_{i=1}^{n} P\left(A_i\right).$$

It is noted that Rule 7.1 and the property $P(\emptyset) = 0$ show that, for a finite sample space, the Axioms 7.1 to 7.3 are equivalent to the axioms in Section 2.2. The added generality of Axiom 7.3 is necessary when the sample space is infinite.

The proof of Rule 7.2 is as follows. The set $A \cup A^c$ is by definition equal to the sample space. Hence, by Axiom 2, $P(A \cup A^c) = 1$. The sets A and A^c are disjoint. It now follows from Rule 7.1 that $P(A \cup A^c) = P(A) + P(A^c)$. This gives the complement rule $P(A) = 1 - P(A^c)$.

To prove Rule 7.3, denote by A_1 the set of outcomes that belong to A but not to B. Let B_1 be the set of outcomes that are in B but not in A and let $C = AB$ be the set of outcomes that are both in A and B. The sets A_1, B_1, and C are pairwise disjoint. Moreover,

$$A \cup B = A_1 \cup B_1 \cup C, \qquad A = A_1 \cup C \qquad \text{and} \qquad B = B_1 \cup C.$$

Applying Rule 7.1 gives

$$P(A \cup B) = P(A_1) + P(B_1) + P(C).$$

Also, $P(A) = P(A_1) + P(C)$ and $P(B) = P(B_1) + P(C)$. By substituting the latter two relations into the expression for $P(A \cup B)$ and noting that $C = AB$, we find

$$\begin{aligned} P(A \cup B) &= P(A) - P(C) + P(B) - P(C) + P(C) \\ &= P(A) + P(B) - P(AB). \end{aligned}$$

Rule 7.4 will only be proved for the special case that the sample space is finite or countably infinite. In this case $P(A) = \sum_{\omega \in A} p(\omega)$, where $p(\omega)$ is the probability assigned to the individual element ω of the sample space. Fix ω. If $\omega \notin \cup_{i=1}^{n} A_i$, then ω does not belong to any of the sets A_i and $p(\omega)$ does not contribute to either the left-hand side or the right-hand side of the expression in Rule 7.4. Assume that $\omega \in \cup_{i=1}^{n} A_i$. Then, there is at least one set A_i to which ω belongs. Let s be the number of sets A_i to which ω belongs. In the left-hand side of the expression in Rule 7.4, $p(\omega)$ contributes only once. In the first term of the right-hand side of this expression, $p(\omega)$ contributes s times, in the second term $\binom{s}{2}$ times, in the third term $\binom{s}{3}$

times, and so on. Thus, the coefficient of $p(\omega)$ in the right-hand side is

$$s - \binom{s}{2} + \binom{s}{3} - \cdots + (-1)^{s-1}\binom{s}{s}.$$

Rule 7.4 follows by proving that this coefficient is equal to 1. Since $\binom{s}{1} = s$ and $\binom{s}{0} = 1$, we find

$$\begin{aligned} &s - \binom{s}{2} + \binom{s}{3} - \cdots + (-1)^{s-1}\binom{s}{s} \\ &\quad = 1 - \left[\binom{s}{0} - \binom{s}{1} + \binom{s}{2} - \binom{s}{3} + \cdots + (-1)^{s}\binom{s}{s}\right] \\ &\quad = 1 - (-1+1)^s = 1, \end{aligned}$$

where the second equality uses Newton's binomium $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}$.

Next, we give several illustrative applications of the above properties. We first illustrate Rule 7.2, which is known as the *complement rule*. This rule states that the probability of an event occurring is one minus the probability that it does not occur. This simple property is extremely useful. It is often easier to compute the complementary probability than the probability itself.

Example 7.9 What is the probability of getting at least one ace in a poker hand of five cards dealt from 52 cards?

Solution. Let A be the event that you get at least one ace. It is easier to compute the probability of the complementary event A^c that you get no ace in a poker hand of five cards. For the sample space of the chance experiment, we take all ordered five-tuples $(x_1, x_2, x_3, x_4, x_5)$, where x_i corresponds to the suit and value of the ith card you get dealt. The total number of possible outcomes equals $52 \times 51 \times 50 \times 49 \times 48$. The number of outcomes without ace equals $48 \times 47 \times 46 \times 45 \times 44$. Assuming that the cards are randomly dealt, all possible outcomes are equally likely. Then, the event A^c has the probability

$$P(A^c) = \frac{48 \times 47 \times 46 \times 45 \times 44}{52 \times 51 \times 50 \times 49 \times 48} = 0.6588.$$

Hence, the probability of getting at least one ace in a poker hand of five cards is $1 - P(A^c) = 0.3412$.

Rule 7.3 is often called the *addition rule* and is illustrated with the following example.

Example 7.10 A single card is randomly drawn from a thoroughly shuffled deck of 52 cards. What is the probability that the drawn card will be either a heart or an ace?

Solution. For the sample space of this chance experiment, we take the set consisting of the 52 elements

$$\spadesuit A, \ldots, \spadesuit 2, \quad \heartsuit A, \ldots \heartsuit 2, \quad \clubsuit A, \ldots, \clubsuit 2, \quad \diamondsuit A, \ldots, \diamondsuit 2,$$

where, for example, the outcome $\clubsuit 7$ means that the seven of clubs is drawn. All possible outcomes are equally likely and thus each outcome gets assigned the same probability $\frac{1}{52}$. Let A be the event that the drawn card is a heart and B the event that the drawn card is an ace. These two events are not mutually exclusive. We are looking for the probability $P(A \cup B)$ that at least one of the events A and B occurs. This probability can be calculated by applying Rule 7.3

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

In this case, $P(AB)$ stands for the probability that the drawn card is the ace of hearts. The events A and B correspond to sets that contain 13 and 4 elements, respectively, and thus have respective probabilities $\frac{13}{52}$ and $\frac{4}{52}$. The event AB corresponds to a set that is a singleton and thus has probability $\frac{1}{52}$. Hence, the probability that the drawn card is either a heart or an ace equals

$$P(A \cup B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}.$$

Example 7.11 In the Lotto 6/42, six different numbers are picked at random from the numbers $1, \ldots, 42$. What is the probability that number 10 is picked?

Solution. The answer is $\frac{6}{42}$. A formal derivation goes as follows. Take as sample space all possible permutations of the integers $1, \ldots, 42$. Imagine that the six numbers are picked by taking the numbers in the first six positions of a random permutation. For $i = 1, \ldots, 6$, let A_i be the event that number 10 is in the ith position. Then, $P(A_i) = 41!/42! = 1/42$ for $i = 1, \ldots, 6$. The events $A_1, \ldots, A_6$ are disjoint and so

$$\begin{aligned} P(\text{number 10 is picked}) &= P(A_1 \cup \ldots \cup A_6) \\ &= P(A_1) + \ldots + P(A_6) = \frac{6}{42}. \end{aligned}$$

Problem 7.15 The probability that the events A and B both occur is 0.3. The individual probabilities of the events A and B are 0.7 and 0.5. What is the probability that neither event A nor event B occurs?

Problem 7.16 The event A has probability $\frac{2}{3}$ and there is a probability of $\frac{3}{4}$ that at least one of the events A and B occurs. What are the smallest and largest possible values for the probability of event B?

Problem 7.17 A small transport company has two vehicles, a truck and a van. The truck is used 75% of the time. Both vehicles are used 30% of the time and neither of the vehicles is used for 10% of the time. What is the probability that the van is used on any given day?

Problem 7.18 In the casino game of Chuck-a-Luck, three dice are contained within an hourglass-shaped, rotating cage. You bet on one of the six possible numbers and the cage is rotated. You lose money only if your number does not come up on any of the three dice. Much to the pleasure of the casinos, people sometimes reason as follows: the probability of my number coming up on one die is 1/6 and so the probability of my number coming up on one of the three dice is $3 \times \frac{1}{6} = \frac{1}{2}$. Why is this reasoning false? How do you calculate the correct value of the probability that your number will come up on any of the three dice? *Hint*: use a formula for $P(A \cup B \cup C)$.

Problem 7.19 An integer is chosen at random from the integers $1, \ldots, 1000$. What is the probability that the integer chosen is divisible by 3 or 5? What is the probability that the integer chosen is divisible by 3, 5 or 7?

Problem 7.20 For the upcoming drawing of the Bingo Lottery, five extra prizes have been added to the pot. Each prize consists of an all-expenses paid vacation trip. Each prize winner may choose from among three possible destinations A, B, and C. The three destinations are equally popular. The prize winners choose their destinations independently of each other. Calculate the probability that at least one of the destinations A and B will be chosen. Also, calculate the probability that not each of the three destinations will be chosen.

Inclusion-exclusion rule

Rule 7.4 extends Rule 7.3 and is known as the *inclusion-exclusion rule*. This rule states that the probability of the union of n events equals the sum of the probabilities of these events taken one at a time, minus the sum of the probabilities of these events taken two at a time, plus the sum of the probabilities of these events taken three at a time, and so on. We illustrate this property with the following classic example.

Example 7.12 Letters to n different persons are randomly put into n pre-addressed envelopes. What is the probability that at least one person receives the correct letter?

Solution. For the formulation of the sample space for this chance experiment, it is convenient to give the label i to the envelope with the address of person i for $i = 1, \ldots, n$. Then, we take the set of all possible orderings $(e_1, \ldots, e_n)$ of the integers $1, \ldots, n$ as our sample space. In the outcome $\omega = (e_1, \ldots, e_n)$, the letter to person i is put into the envelope with label e_i for $i = 1, \ldots, n$. The total number of possible outcomes is $n \times (n-1) \times \cdots \times 1 = n!$. Since the letters are put randomly into the envelopes, all possible orderings are equally likely and thus each outcome $(e_1, \ldots, e_n)$ gets assigned the same probability $\frac{1}{n!}$. For fixed i, let A_i be the event that the letter for person i is put into the envelope with label i. The probability that at least one person receives the correct letter is given by $P(A_1 \cup A_2 \cup \cdots \cup A_n)$. The probabilities in the inclusion-exclusion formula in Rule 7.4 are easy to calculate. For fixed i, the total number of orderings $(e_1, \ldots, e_n)$ with $e_i = i$ is equal to $(n-1)!$. This gives

$$P(A_i) = \frac{(n-1)!}{n!} \qquad \text{for } i = 1, \ldots, n.$$

Next fix i and j with $i \neq j$. The number of orderings $(e_1, \ldots, e_n)$ with $e_i = i$ and $e_j = j$ is equal to $(n-2)!$. Hence

$$P(A_i A_j) = \frac{(n-2)!}{n!} \qquad \text{for all } i \text{ and } j \text{ with } i \neq j.$$

Continuing in this way, we find

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) = {} & \binom{n}{1}\frac{(n-1)!}{n!} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} \\ & - \ldots + (-1)^{n-1}\binom{n}{n}\frac{1}{n!}. \end{aligned}$$

Since $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, this expression simplifies to

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) &= \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!} \\ &= 1 - \sum_{k=0}^{n} \frac{(-1)^k}{k!}. \end{aligned}$$

A surprising conclusion can be drawn from this result. A basic result from calculus is that $\sum_{k=0}^{\infty} (-1)^k/k! = e^{-1}$ with $e = 2.718\ldots$ (see the Appendix). Thus, for large n, the probability that at least one person will receive the correct letter is approximately equal to $1 - e^{-1} = 0.632$, independently of how large n is.

Having obtained the probability of at least one person receiving a correct letter, it is not difficult to argue that

$$P(\text{exactly } j \text{ persons receive a correct letter}) = \frac{1}{j!}\sum_{k=0}^{n-j} \frac{(-1)^k}{k!}$$

for $j = 0, 1, \ldots, n$. To verify this, denote by N_m the number of permutations of the integers $1, \ldots, m$ so that no integer remains in its original position.

Since the probability that a random permutation of $1, \ldots, m$ has this property is $\sum_{k=0}^{m} (-1)^k/k!$, it follows that $N_m/m! = \sum_{k=0}^{m} (-1)^k/k!$. The number of permutations of the integers $1, \ldots, n$ so that exactly j integers remain in their original positions equals $\binom{n}{j} N_{n-j}$. Thus

$$P(\text{exactly } j \text{ persons receive a correct letter}) = \frac{\binom{n}{j} N_{n-j}}{n!}.$$

Noting that $\binom{n}{j}/n! = \frac{1}{j!(n-j)!}$ and inserting the expression for $N_{n-j}/(n-j)!$, the above formula for the probability of exactly j persons receiving a correct letter follows. This probability tends to the Poisson probability $e^{-1}/j!$ as the number of envelopes becomes large (see Section 4.2.3 for a discussion of the Poisson approximation).

Many probability problems of a combinatorial nature can be solved by using the inclusion-exclusion rule. We give two more examples.

Example 7.13 Fifteen tourists are stranded in a city with four hotels, all of which are located near each other in the city center. Each hotel has enough rooms available to accommodate all 15 tourists. Each tourist randomly chooses a hotel, independently of the choices made by the others. What is the probability that not all four hotels will be chosen?

Solution. We leave it to the reader to verify that the desired probability is given by

$$\sum_{k=1}^{4} (-1)^{k+1} \binom{4}{k} \frac{(4-k)^{15}}{4^{15}} = 0.0533.$$

Example 7.14 Suppose $n = 10$ married couples are invited to a bridge party. Bridge partners are chosen at random, without regard to gender. What is the probability that no one will be paired with his or her spouse?

Solution. Denote by A_i the event that couple i is paired as bridge partners. Take as sample space the set of all possible permutations of the integers $1, \ldots, 2n$, where the integers $2i - 1$ and $2i$ represent couple i. We leave it to the reader to verify that the complementary probability $P(A_1 \cup A_2 \cup \ldots \cup A_n)$ of at least one couple being paired as bridge partners is given by

$$\sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{n \times (n-1) \times \cdots \times (n-k+1) \times 2^k \times (2n-2k)!}{(2n)!}.$$

This probability has the value 0.4088 for $n = 10$ (the probability of at least one couple being paired as bridge partners tends to $1 - e^{-\frac{1}{2}} = 0.3935$ as the number of couples gets large).

Problem 7.21 What is the probability that in a player's hand of 13 cards at least one suit will be missing?

Problem 7.22 Consider the card game Jeu de Treize from Problem 3.32. Use the inclusion-exclusion rule to verify that the probability of the dealer winning the first round is 0.6431.

Problem 7.23 What is the probability that a hand of 13 cards contains four of a kind?

8

Conditional probability and Bayes

The concept of conditional probability is one of the most important concepts in probability theory. It is extremely useful in problem solving. Conditional probabilities express the fact that probabilities alter when the available information alters. In this chapter you will learn about the basics of conditional probability and the law of conditional probabilities. The concept of conditional probability is intuitive for most people. For example, most people reason as follows to find the probability of getting two aces when two cards are selected at random from an ordinary deck of cards. The probability of getting an ace on the first card is $\frac{4}{52}$. Given that one ace is gone from the deck, the probability of getting an ace on the second card is $\frac{3}{51}$. The desired probability is therefore $\frac{4}{52} \times \frac{3}{51}$.

In many problems, the law of conditional probabilities facilitates the calculation of the desired probabilities through an appropriate conditioning argument. Also, conditional probabilities are inextricably bound up with Bayes' rule. This rule can be seen as a way of understanding how the probability of an event is affected by a new piece of information. A particularly useful form of this rule is Bayes' rule in odds form.

8.1 Conditional probability

The starting point for the definition of conditional probability is a chance experiment for which a sample space and a probability measure P are defined. Let A be an event of the experiment. The probability $P(A)$ reflects our knowledge of the occurrence of event A *before* the experiment takes place. Therefore the probability $P(A)$ is sometimes referred to as the *a priori* probability of A or the *unconditional* probability of A. Suppose now we are told that an event B has occurred in the experiment, but we still do not know the precise outcome in the set B. In light of this added information, the set B replaces the sample space as the set of possible outcomes and consequently the probability of the occurrence of event A changes. A conditional probability now reflects our knowledge of the occurrence of the event A given that event B has occurred. The notation for this new probability is $P(A \mid B)$.

Definition 8.1 *For any two events A and B with $P(B) > 0$, the conditional probability $P(A \mid B)$ is defined as*

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

Here AB stands for the occurrence of both event A and event B. This is not an arbitrary definition. It can be intuitively reasoned through a comparable property of the relative frequency. Let's define the relative frequency $f_n(E)$ of the occurrence of event E as $\frac{n(E)}{n}$, where $n(E)$ represents the number of times that E occurs in n repetitions of the experiment. Assume, now, that in n independent repetitions of the experiment, event B occurs r times simultaneously with event A and s times without event A. We can then say that $f_n(AB) = \frac{r}{n}$ and $f_n(B) = \frac{r+s}{n}$. If we divide $f_n(AB)$ by $f_n(B)$, then we find that

$$\frac{f_n(AB)}{f_n(B)} = \frac{r}{r+s}.$$

Now define $f_n(A \mid B)$ as the relative frequency of event A in those repetitions of the experiment in which event B has occurred. From $f_n(A \mid B) = \frac{r}{r+s}$ we now get the following relationship

$$f_n(A \mid B) = \frac{f_n(AB)}{f_n(B)}.$$

This relationship accounts for the definition of conditional probability $P(A \mid B)$.

Example 8.1 Someone has rolled a fair die twice. You know that one of the rolls turned up a face value of six. What is the probability that the other roll turned up a six as well?

Solution. Take as sample space the set $\{(i, j) | i, j = 1, \ldots, 6\}$, where i and j denote the outcomes of the first and second rolls. A probability of $\frac{1}{36}$ is assigned to each element of the sample space. The event of two sixes is given by $A = \{(6, 6)\}$ and the event of at least one six is given by $B = \{(1, 6), \ldots, (5, 6), (6, 6), (6, 5), \ldots, (6, 1)\}$. Applying the definition of conditional probability gives

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{1/36}{11/36}.$$

Hence the desired probability is $\frac{1}{11}$ (not $\frac{1}{6}$).

The above example illustrates once again how careful you have to be when you are interpreting the information a problem is conveying. The wording of the problem is crucial: you know that one of the dice turned up a six but you do not know which one. In the case where one of the dice had dropped on the floor and you had seen the outcome six for that die, the probability of the other die turning up a six would have been $\frac{1}{6}$.

Example 8.2 John, Pedro, and Rosita are experienced dart players. The probability of John hitting the bull's eye in a single throw is $\frac{1}{3}$. This hitting probability is $\frac{1}{5}$ for Pedro and $\frac{1}{4}$ for Rosita. The three players each throw simultaneously one dart. Two of the darts hit the bull's eye and one of the darts misses the bull's eye. What is the probability that John is the one who missed?

Solution. The sample space of the chance experiment consists of the eight elements $(H, H, H), (H, H, M), (H, M, H), (H, M, M), (M, H, H), (M, H, M)$, (M, M, H), and (M, M, M), where M stands for "miss" and H stands for "hit." The first component of each element of the sample space refers to John's throw, the second component refers to Pedro's throw, and the third component refers to Rosita's throw. By the independence of the outcomes of the individual throws, we assign the probability $\frac{1}{3} \times \frac{1}{5} \times \frac{1}{4} = \frac{1}{60}$ to the outcome (H, H, H), the probability $\frac{1}{3} \times \frac{1}{5} \times \frac{3}{4} = \frac{3}{60}$ to the outcome (H, H, M), the probability $\frac{1}{3} \times \frac{4}{5} \times \frac{1}{4} = \frac{4}{60}$ to the outcome (H, M, H), the probability $\frac{1}{3} \times \frac{4}{5} \times \frac{3}{4} = \frac{12}{60}$ to the outcome (H, M, M), the probability $\frac{2}{3} \times \frac{1}{5} \times \frac{1}{4} = \frac{2}{60}$ to the outcome (M, H, H), the probability $\frac{2}{3} \times \frac{1}{5} \times \frac{3}{4} = \frac{6}{60}$ to the outcome (M, H, M), the probability $\frac{2}{3} \times \frac{4}{5} \times \frac{1}{4} = \frac{8}{60}$ to the outcome (M, M, H), and the probability $\frac{2}{3} \times \frac{4}{5} \times \frac{3}{4} = \frac{24}{60}$ to the outcome (M, M, M). We are now ready to determine the desired probability $P(A \mid B)$, where A is the event that John misses and B is the event that exactly two of the darts hit the target. The event AB occurs if the outcome (M, H, H) occurs and the event B occurs if one of the outcomes $(H, H, M), (H, M, H), (M, H, H)$ occurs. Thus, $P(AB) = \frac{2}{60}$ and $P(B) = \frac{3}{60} + \frac{4}{60} + \frac{2}{60} = \frac{9}{60}$. Applying the formula $P(A \mid B) = P(AB)/P(B)$, we can now conclude that

$$P(\text{John misses} \mid \text{exactly two darts hit the target}) = \frac{2}{9}.$$

Problem 8.1 Every evening, two weather stations issue a weather forecast for the next day. The weather forecasts of the two stations are independent of each other. On average, the weather forecast of station 1 is correct in 90% of the cases, irrespective of the weather type. This percentage is 80% for station 2. On a given day, station 1 predicts sunny weather for the next day, whereas station 2 predicts rain. What is the probability that the weather forecast of station 1 will be correct?

Problem 8.2 Someone has tossed a fair coin three times. You know that one of the tosses came up heads. What is the probability that at least one of the other two tosses came up heads as well?

Problem 8.3 Suppose a bridge player's hand of 13 cards contains an ace. What is the probability that the player has only one ace? What is the answer to this question if you know that the player had the ace of hearts?

8.1.1 Assigning probabilities through conditional probabilities

The formula for the conditional probability $P(A \mid B)$ can be rewritten as

$$P(AB) = P(A \mid B)P(B).$$

This phrasing lines up more naturally with the intuitive way people think about probabilities. In many cases, $P(AB) = P(A \mid B)P(B)$ is used in attributing probabilities to elements of the sample space. In illustration of this, consider the experiment in which two marbles are randomly chosen without replacements from a receptacle holding seven red and three white marbles. One possible choice for the sample space of this experiment is the set consisting of four elements (R, R), (R, W), (W, W), and (W, R), where R stands for red and W for white. The first component of each element indicates the color of the first marble chosen and the second component the color of the second marble chosen. On grounds of the reasoning that $P(\text{1st marble is red }) = \frac{7}{10}$ and $P(\text{2nd marble is white }|\text{1st marble is red}) = \frac{3}{9}$, we attribute the probability of $P(R, W) = \frac{7}{10} \times \frac{3}{9} = \frac{7}{30}$ to the element (R, W). In the same way we attribute the probabilities $P(R, R) = \frac{7}{10} \times \frac{6}{9} = \frac{7}{15}$, $P(W, W) = \frac{3}{10} \times \frac{2}{9} = \frac{1}{15}$, and $P(W, R) = \frac{3}{10} \times \frac{7}{9} = \frac{7}{30}$ to the remaining elements. It is common practice in this type of problem to assign probabilities to the elements of the sample space as a product of probabilities, one marginal and the others conditional. To do so, one uses the formula

$$\begin{aligned} &P(A_1A_2 \cdots A_n) \\ &\quad = P(A_1) \times P(A_2 \mid A_1) \times P(A_3 \mid A_1A_2) \times \cdots \times P(A_n \mid A_1A_2 \cdots A_{n-1}), \end{aligned}$$

this being an extension of the formula $P(A) = P(A \mid B)P(B)$.

Example 8.3 A group of 15 tourists is stranded in a city with four hotels of the same class. Each of the hotels has enough room available to accommodate the 15 tourists. The group's guide, who has a good working relationship with each of the four hotels, assigns the tourists to the hotels as follows. First, he randomly determines how many are to go to hotel A, then how many of the remaining tourists are to go to hotel B, and then how many are to go to hotel C.

All remaining tourists are sent to hotel D. Note that each stage of the assignment the guide draws at random a number between zero and the number of tourists left. What is the probability that all four hotels receive guests from the group?

Solution. Let the outcome (i_A, i_B, i_C, i_D) correspond with the situation in which i_A tourists are sent to hotel A, i_B tourists to hotel B, i_C tourists to hotel C, and i_D tourists to hotel D. The probability

$$\frac{1}{16} \times \frac{1}{16 - i_A} \times \frac{1}{16 - i_A - i_B}$$

is assigned to the outcome (i_A, i_B, i_C, i_D) for $0 \leq i_A, i_B, i_C, i_D \leq 15$ and $i_A + i_B + i_C + i_D = 15$. The probability that all four hotels will receive guests is given by

$$\sum_{i_A=1}^{12} \sum_{i_B=1}^{13-i_A} \sum_{i_C=1}^{14-i_A-i_B} \frac{1}{16} \times \frac{1}{16 - i_A} \times \frac{1}{16 - i_A - i_B} = 0.2856.$$

Problem 8.4 You travel from Amsterdam to Sidney with change of airplanes in Dubai and Singapore. You have one piece of luggage. At each stop your luggage is transferred from one airplane to another. At the airport in Amsterdam there is a probability of 5% that your luggage is not placed in the right plane. This probability is 3% at the airport in Dubai and 2% at the airport in Singapore. What is the probability that your luggage does not reach Sidney with you? If your luggage does not reach Sidney with you, what is the probability that it was lost at the airport of Dubai?

Problem 8.5 Seven individuals have reserved tickets at the opera. The seats they have been assigned are all in the same row of seven seats. The row of seats is accessible from either end. Assume that the seven individuals arrive and take their seats in a random order. What is the probability of all seven individuals taking their seats without having to squeeze past an already seated individual? Use conditional probabilities to answer this question. *Hint*: assume that the individuals get up from their seats one by one and in a random order, and calculate the probability of the individuals leaving without having to squeeze past others in their row.

Problem 8.6 Your favorite team participates in a knock-out system that consists of four rounds. If your team has reached round i, it will survive this round with a given probability of p_i for $i = 1, \ldots, 4$. After the competition, you are informed that your team is not the final winner. This is the only information you get about the tournament. What is the probability that your team was eliminated in round i for $i = 1, \ldots, 4$?

8.1.2 Independent events

In the special case of $P(A \mid B) = P(A)$, the occurrence of event A is not contingent on the occurrence or nonoccurrence of event B. Event A is then said to be independent of event B. In other words, if A is independent of B, then learning that event B has occurred does not change the probability that event A occurs. Since $P(A \mid B) = \frac{P(AB)}{P(B)}$, it follows that A is independent of B if the equation $P(AB) = P(A)P(B)$ holds true. This equation is symmetric in A and B: if A is independent of B, then B is also independent of A. Summarizing:

Definition 8.2 *Two events A and B are said to be independent if*

$$P(AB) = P(A)P(B).$$

The reader should be aware that independent events and disjoint events are completely different things. If events A and B are disjoint, you calculate the probability of the union $A \cup B$ by *adding* the probabilities of A and B. For independent events A and B you calculate the probability of the intersection AB by *multiplying* the probabilities of A and B. Since $P(AB) = 0$ for disjoint events A and B, independent events are typically not disjoint.

Example 8.4 Suppose two fair dice are thrown. Let A be the event that the number shown by the first die is even, and B the event that the sum of the dice is odd. Do you think the events A and B are independent?

Solution. The experiment has 36 possible outcomes (i, j), where i is the number shown by the first die and j the number shown by the second die. All possible outcomes are equally likely. Simply, by counting, $P(A) = 18/36$, $P(B) = 18/36$, and $P(AB) = 9/36$. Since $P(AB) = P(A)P(B)$, events A and B are independent.

Remark 8.1 In the case that events A, B, and C are pairwise independent, it is not necessarily true that $P(ABC) = P(A)P(B)P(C)$. This can be shown using Example 8.4. In addition to the events A and B from Example 8.4, let C be the event that the number shown by the second die is even. Events A, B, and C are pairwise independent, but $P(ABC)(= 0)$ is not equal to $P(A)P(B)P(C)$. In general, events $A_1, \ldots, A_n$ are said to be independent if $P(A_{i_1} \ldots A_{i_k}) = P(A_{i_1}) \times \cdots \times P(A_{i_k})$ for every collection $A_{i_1}, \ldots, A_{i_k}$ and $2 \leq k \leq n$.

In practical problems it is rarely needed to check independence in such detail as in Example 8.4, but independence of events can be usually be concluded directly from the physical setup of the underlying chance experiment. Independent events A and B typically arise in a compound experiment consisting of physically independent subexperiments, where one subexperiment alone determines whether event A occurs and another subexperiment alone determines whether event B occurs.

Problem 8.7 Two fair coins are tossed. Let A be the event that heads appears on the first coin and let B be the event that the coins display the same outcome. Are the events A and B independent?

8.1.3 The law of conditional probabilities

It is often the case that the unconditional probability $P(A)$ of an event A is found most easily by expressing it in terms of conditional probabilities. The idea is to choose an appropriate sequence of *mutually exclusive* events $B_1, \ldots, B_n$ such that the event A can only occur when one of the disjoint events $B_1, \ldots, B_n$ occurs. Next, the probability $P(A)$ can be obtained by applying the following rule:

Rule 8.1 *Let A be an event that can only occur when one of mutually exclusive events $B_1, \ldots, B_n$ occurs. Then*

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n).$$

This rule is called the *law of conditional probabilities*. The proof of this law is simple and instructive. The assumption that event A can only occur if one of the events $B_1, \ldots, B_n$ also occurs means, in terms of sets, that the subset A of the sample space is contained in the union $B_1 \cup \cdots \cup B_n$ of the subsets $B_1, \ldots, B_n$. This implies

$$A = AB_1 \cup AB_2 \cup \cdots \cup AB_n,$$

where AB_i stands for the set of outcomes belonging both to set A and set B_i. The assumption that the sets $B_1, \ldots, B_n$ are pairwise disjoint implies that the sets $AB_1, \ldots, AB_n$ are also pairwise disjoint. By Rule 7.1 in Chapter 7, we then have

$$P(A) = P(AB_1) + P(AB_2) + \cdots + P(AB_n).$$

This relationship and the definition $P(A \mid B) = P(AB)/P(B)$ lead to the law of conditional probabilities. This law is naturally also applicable when the sample space is divided by a countably infinite number of disjoint subsets $B_1, B_2, \ldots$ instead of by a finite number.

A nice illustration of the law of conditional probabilities is provided by the craps example in Section 3.3 of Chapter 3. Another nice illustrative example is the following one.

Example 8.5 The upcoming Tour de France bicycle tournament will take place from July 1 through July 23. One hundred eighty cyclists will participate in the event. What is the probability that two or more participating cyclists will have birthdays on the same day during the tournament?

Solution. Denoting by A the event that two or more participating cyclists will have birthdays on the same day during the tournament, event A can occur only if one of the mutually exclusive events $B_2, \ldots, B_{180}$ occurs. Event B_i occurs when exactly i participating cyclists have birthdays during the tournament. The conditional probability $P(A \mid B_i)$ is easy to calculate. It refers to the birthday problem with i persons coming from a "planet" where the year has 23 days. The birthday problem was studied in detail in Chapter 3. The reader may easily verify that

$$P(A \mid B_i) = \begin{cases} 1 - \frac{23\times 22\times\cdots\times(23-i+1)}{(23)^i}, & 2 \leq i \leq 23 \\ 1, & i \geq 24 \end{cases}$$

with $P(A \mid B_0) = P(A \mid B_1) = 0$. The probability $P(B_i)$ is given by

$$P(B_i) = \binom{180}{i}\left(\frac{23}{365}\right)^i\left(1 - \frac{23}{365}\right)^{180-i}, \qquad 0 \leq i \leq 180.$$

Putting the pieces together, we find

$$\begin{aligned} P(A) &= \sum_{i=2}^{180} P(A \mid B_i)P(B_i) \\ &= 1 - P(B_0) - P(B_1) - \sum_{i=2}^{23} \frac{23 \times 22 \times \cdots \times (23 - i + 1)}{(23)^i} P(B_i). \end{aligned}$$

This yields the value 0.8841 for the probability $P(A)$.

Problem 8.8 A die is rolled to yield a number between 1 and 6, and then a coin is tossed that many times. What is the probability that heads will not appear?

Problem 8.9 Three friends travel by plane for the first time. They got assigned the seats A (window), B (middle), and C (aisle) in the same row. On the seats A and C passengers cannot wrongly fasten their seat belts, but an inexperienced traveler on the middle seat B fastening the seat belt first has a 50-50 chance of wrongly fastening the seat belt by picking the buckle from seat A and the belt from seat C. Assuming that the three friends take their seats one by one in random order, calculate the probability that the seat belts are fastened correctly.

Problem 8.10 Let's return to the casino game Red Dog from Problem 3.25. Using the law of conditional probabilities, calculate the probability of the player winning. *Hint*: argue first that the probability of a spread of i points is given by $\frac{1}{52!}[(12 - i) \times 4 \times 4 \times 2]$.

Problem 8.11 Consider the scratch-lottery problem from Section 4.2.3. Each week one million scratch-lottery tickets are printed. Assume that in a particular week only one-half of the tickets printed are sold. What is the probability of at least one winner in that week? *Hint*: use results from Example 7.12.

Problem 8.12 A fair die is rolled repeatedly. Let p_n be the probability that the sum of scores will ever be n. Use the law of conditional probabilities to find a recursion equation for p_n. Verify numerically that p_n tends to $\frac{1}{3.5} = 0.2857$ as n gets large. Can you explain this result?

Problem 8.13 A fair coin is tossed k times. Let a_k denote the probability of having no two heads in a row in the sequence of tosses. Use the law of conditional probabilities to obtain a recurrence relation for a_k. Calculate a_k for $k = 5$, 10, 25, and 50.

Problem 8.14 It is believed that a sought-after wreck will be in a certain sea area with probability $p = 0.4$. A search in that area will detect the wreck with probability $d = 0.9$ if it is there. What is the revised probability of the wreck being in the area when the area is searched and no wreck is found?

Problem 8.15 Consider the lost boarding pass puzzle that was stated in Section 2.9.2. Assume now that N people are lining up to board an airplane with N seats. Verify that the probability of the last passenger getting his/her own seat equals $\frac{1}{2}$, regardless of the value of $N \geq 2$.

8.2 Bayes' rule in odds form

Bayes' rule specifies how probabilities must be updated in the light of new information. The Bayesian approach to probable inference is remarkably straightforward and intuitive. In Chapter 6 we discussed the standard form of Bayes' rule. However, the essence of Bayesian reasoning is best understood by considering the odds form of Bayes' rule for the situation where there is question of a hypothesis being either true or false. An example of such a situation is a court case where the defendant is either guilty or not guilty. Let H represent the event that the hypothesis is true, and $\overline{H}$ the event that the hypothesis is false. Before examining the evidence, a Bayesian analysis begins with assigning prior subjective probabilities $P(H)$ and $P(\overline{H}) = 1 - P(H)$ to the mutually exclusive events H and $\overline{H}$. How do the prior probabilities change once evidence in the form of the knowledge that the event E has occurred becomes available? In our example of the court case, event E could be the evidence that the accused has the same blood type as the perpetrator's, whose blood has been found at the scene of the crime. The updated value of the probability that the hypothesis is true given the fact that event E has occurred is denoted by $P(H \mid E)$. To calculate the posterior probability $P(H \mid E)$, we use Bayes' rule. This rule can be expressed in several ways. A convenient form uses odds. Odds are often used to represent probabilities. Gamblers usually think in terms of "odds" instead of probabilities. For an event with probability of $\frac{2}{3}$, the odds are 2 to 1 (written 2:1), while for an event with a probability of $\frac{3}{10}$, the odds are 3:7. The odds form of Bayes' rule reads as follows:

Rule 8.2 *The posterior probability $P(H \mid E)$ satisfies*

$$\frac{P(H \mid E)}{P(\overline{H} \mid E)} = \frac{P(H)}{P(\overline{H})}\frac{P(E \mid H)}{P(E \mid \overline{H})}.$$

In words, Bayes' rule in odds form states that

$$\textit{posterior odds} = \textit{prior odds} \times \textit{likelihood ratio.}$$

This insightful formula follows by twice applying the definition of conditional probability. By doing so, we obtain

$$P(H \mid E) = \frac{P(HE)}{P(E)} = P(E \mid H)\frac{P(H)}{P(E)}.$$

The same expression holds for $P(\overline{H} \mid E)$ with H replaced by $\overline{H}$. Dividing the expression for $P(H \mid E)$ by the expression for $P(\overline{H} \mid E)$ results in the odds form of Bayes' rule.

The factor $\frac{P(H)}{P(\overline{H})}$ gives the prior odds and represents the odds in favor of the hypothesis H before the evidence has been presented. The ratio of $P(E \mid H)$ and $P(E \mid \overline{H})$ is called the likelihood ratio or the Bayes' factor. The likelihood ratio gives the odds of obtaining the evidence when the hypothesis under consideration is true. It represents the impact the evidence will have on the belief in the hypothesis. If it is likely that the evidence will be observed when the hypothesis under consideration is true, then the Bayes' factor will be large. Bayes' rule updates the prior odds of the hypothesis H by multiplying them with the likelihood ratio and thus measures how much new evidence should alter a belief in a hypothesis. With two independent pieces of evidence E_1 and E_2, Bayes' rule can be applied iteratively. You could use the first piece of evidence to calculate initial posterior odds, and then use that posterior odds as new prior odds to calculate second posterior odds given the second piece of evidence. In practical situations such as in judicial decision making, the likelihood ratio of the evidence is typically determined by an expert.[†] However it is not the expert's task to tell the court what the prior odds are. The prior probability $P(H)$ represents the personal opinion of the court before the evidence is taken into account.

Example 8.6 A murder is committed. The perpetrator is either one or the other of the two persons X and Y. Both persons are on the run from authorities, and, after an initial investigation, both fugitives appear equally likely to be the perpetrator. Further investigation reveals that the actual perpetrator has blood type A. Ten percent of the population belongs to the group having this blood type. Additional inquiry reveals that person X has blood type A, but offers no information concerning the blood type of person Y. In light of this new information, what is the probability that person X is the perpetrator?

Solution. In answering this question, use H to denote the event that person X is the perpetrator. Let E represent the new evidence that person X has blood type A. The prior probabilities of H and $\overline{H}$ before the appearance of the new evidence E are given by

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

In addition, it is also true that

$$P(E \mid H) = 1 \quad \text{and} \quad P(E \mid \overline{H}) = \frac{1}{10}.$$

Applying Bayes' rule in odds form at this point, we find that

$$\frac{P(H \mid E)}{P(\overline{H} \mid E)} = \frac{1/2}{1/2} \times \frac{1}{1/10} = 10.$$

The odds in favor, then, are 10 to 1 that person X is the perpetrator given that this person has blood type A. Otherwise stated, from $P(H \mid E)/[1 - P(H \mid E)]$

[†] In both legal and medical cases, the conditional probabilities $P(H \mid E)$ and $P(E \mid H)$ are sometimes confused with each other. A classic example is the famous court case of People vs. Collins in Los Angeles in 1964. In this case, a couple matching the description of a couple that had committed an armed robbery was arrested. Based on expert testimony, the district attorney claimed that the frequency of couples matching the description was roughly 1 in 12 million. Although this was the estimate for $P(E \mid \overline{H})$, the district attorney treated this estimate as if it was $P(\overline{H} \mid E)$ and incorrectly concluded that the couple was guilty beyond reasonable doubt. The prosecutor's fallacy had dramatic consequences in the case of Regina vs. Sally Clark in UK in 1999. Sally Clark was convicted for murder because of the cot deaths of two of her newborn children within a period of one year. A revision of her process benefited from Bayesian arguments and led to her release in 2001.

$= 10$, it follows that

$$P(H \mid E) = \frac{10}{11}.$$

The probability of Y being the perpetrator is $1 - \frac{10}{11} = \frac{1}{11}$ and not, as may be thought, $\frac{1}{10} \times \frac{1}{2} = \frac{1}{20}$. The error in this reasoning is that the probability of person Y having blood type A is not $\frac{1}{10}$ because Y is not a randomly chosen person; rather, Y is first of all a person having a 50% probability of being the perpetrator, whether or not he is found at a later time to have blood type A. Bayesian analysis sharpens our intuition in a natural way.

Another nice illustration of Bayes' rule in odds form is provided by legal arguments used in the discussion of the O.J. Simpson trial.[†]

Example 8.7 Nicole Brown was murdered at her home in Los Angeles on the night of June 12, 1994. The prime suspect was her husband O.J. Simpson, at the time a well-known celebrity famous both as a TV actor and as a retired professional football star. This murder led to one of the most heavily publicized murder trials in the United States during the last century. The fact that the murder suspect had previously physically abused his wife played an important role in the trial. The famous defense lawyer Alan Dershowitz, a member of the team of lawyers defending the accused, tried to belittle the relevance of this fact by stating that only 0.1% of the men who physically abuse their wives actually end up murdering them. Was the fact that O.J. Simpson had previously physically abused his wife irrelevant to the case?

Solution. The answer to the question is no. In this particular court case it is important to make use of the crucial fact that Nicole Brown was murdered. The question, therefore, is not what the probability is that abuse leads to murder, but the probability that the husband is guilty in light of the fact that he had previously abused his wife. This probability can be estimated with the help of Bayes' formula and a few facts based on crime statistics. Define the following

E = the event that the husband has physically abused his wife in the past
M = the event that the wife has been murdered
G = the event that the husband is guilty of the murder of his wife.

The probability in question is the conditional probability $P(G \mid EM)$. We can use Bayes' formula expressed in terms of the posterior odds to calculate this

[†] This example is based on the article J.F. Merz and J.P. Caulkins, Propensity to abuse-propensity to murder?, *Chance Magazine* **8** (1995): 14.

probability. In this example, Bayes' formula is given by

$$\frac{P(G \mid EM)}{P(\overline{G} \mid EM)} = \frac{P(G \mid M)}{P(\overline{G} \mid M)} \frac{P(E \mid GM)}{P(E \mid \overline{G}M)},$$

where $\overline{G}$ represents the event that the husband is not guilty of the murder of his wife. How do we estimate the conditional probabilities on the right-hand side of this formula? In 1992, 4,936 women were murdered in the United States, of which roughly 1,430 were murdered by their (ex)husbands or boyfriends. This results in an estimate of $\frac{1{,}430}{4{,}936} = 0.29$ for the prior probability $P(G \mid M)$ and an estimate of 0.71 for the prior probability $P(\overline{G} \mid M)$. Furthermore, it is also known that roughly 5% of married women in the United States have at some point been physically abused by their husbands. If we assume that a woman who has been murdered by someone other than her husband had the same chance of being abused by her husband as a randomly selected woman, then the probability $P(E \mid \overline{G}M)$ is equal to 5%. The remaining probability on the right-hand side is $P(E \mid GM)$. We can base our estimate of this probability on the reported remarks made by Simpson's famous defense attorney, Alan Dershowitz, in a newspaper article. In the newspaper article, Dershowitz admitted that a substantial percentage of the husbands who murder their wives have, previous to the murders, also physically abused their wives. Given this statement, the probability $P(E \mid GM)$ will be taken to be 0.5. By substituting the various estimated values for the probabilities into Bayes' formula in odds form, we find that

$$\frac{P(G \mid EM)}{P(\overline{G} \mid EM)} = \frac{0.29}{0.71} \frac{0.5}{0.05} = 4.08.$$

We can translate the odds into probabilities using the fact that $P(\overline{G} \mid EM) = 1 - P(G \mid EM)$. This results in a value for $P(G \mid EM)$ of 0.81. In other words, there is an estimated probability of 81% that the husband is the murderer of his wife in light of the knowledge that he had previously physically abused her. The fact that O.J. Simpson had physically abused his wife in the past was therefore certainly very relevant to the case.

Problem 8.16 In a certain region, it rains on average once in every ten days during the summer. Rain is predicted on average for 85% of the days when rainfall actually occurs, while rain is predicted on average for 25% of the days when it does not rain. Assume that rain is predicted for tomorrow. What is the probability of rainfall actually occurring on that day?

Problem 8.17 You have five coins colored red, blue, white, green, and yellow. Apart from the variation in color, the coins look identical. One of the coins is unfair and when tossed comes up heads with a probability of $\frac{3}{4}$; the other four are fair coins. You have no further information about the coins apart from having observed that the blue coin, tossed three times, came up heads on all three tosses. On the grounds of this observation, you indicate that the blue coin is the unfair one. What is the probability of your being correct in this assumption?

Problem 8.18 A friendly couple tells you that they did a 100% reliable sonogram test and found out that they are going to have twin boys. They asked the doctor about the probability of identical twins rather than fraternal twins. The doctor could only give them the information that the population proportion of identical twins is one-third (identical twins are always of the same sex but fraternal twins are random). Can you give the probability the couple asked for? *Remark*: This problem is taken from the paper "Bayesians, frequentists and scientists," by B. Efron, in *Journal of the American Statistical Association* **100** (2005): 1–5.

8.3 Bayesian statistics[†]

In addition to its application in court cases, Bayesian statistics is often used by accountants and tax inspectors to perform audits. Bayesian statistics is also often used to predict election results based on the results of new opinion polls and to update the degree of belief in the effectiveness of medical treatments given new clinical data. Bayesian statistics has also been used for spam filtering. By seeing which words and combination of words appear most often in spam, but rarely in nonspam, the Bayesian filter can determine which e-mails have a higher probability of being spam than others. One of the principal advantages of Bayesian statistics is the ability to perform the analysis sequentially, where new information can be incorporated into the analysis as soon as it becomes available.

Example 8.8 On January 1, 2002, the euro was introduced as the new coin in many European countries. Belgian students made the papers at the beginning of January 2002, with an experiment in which a one-euro coin with the image of King Albert was tossed 250 times and came up heads 140 times. What can be said about this coin?

Solution. In classical statistics, the null hypothesis for this experiment would be that the coin is fair. The null hypothesis would then be tested by calculating the probability of 140 or more heads out of 250 tosses with a fair coin. This

[†] This section can be skipped at first reading.

probability is 0.0332. The approach of classical statistics thus calculates the probability of the data occurring under the null hypothesis. In Bayesian statistics, by contrast, one computes the probability that the null hypothesis is true given the data. More precisely, in the Bayesian approach, one assumes a prior distribution of the probability that a toss of the coin comes up heads. This distribution is revised when data become available. To show how this process works, imagine that there are nine possible values 0.1, 0.2, . . . , 0.9 for the probability θ that the toss of a coin will land heads up. To start with, you have assigned a prior probability $p_0(\theta_i)$ to each possible value $\theta_i = \frac{i}{10}$:

θ_i	$p_0(\theta_i)$	θ_i	$p_0(\theta_i)$	θ_i	$p_0(\theta_i)$
0.1	0.05	0.4	0.15	0.7	0.10
0.2	0.05	0.5	0.30	0.8	0.05
0.3	0.10	0.6	0.15	0.9	0.05

That is, before running an experiment, you believe that with probability 0.05 you have a coin coming up heads on average once in ten tosses, with probability 0.05 you have a coin coming up heads on average twice in ten tosses, etc. Next, after having observed 140 heads in 250 tosses, you calculate for every value θ_i the probability

$$\begin{aligned} &P\,(140 \text{ times heads in 250 tosses} \mid \theta = \theta_i) \\ &\quad = \binom{250}{140} \theta_i^{140} (1 - \theta_i)^{250-140}. \end{aligned}$$

If we denote this probability as $L(\theta_i)$, we find that:

θ_i	$L(\theta_i)$	θ_i	$L(\theta_i)$	θ_i	$L(\theta_i)$
0.1	0	0.4	1.16×10^{-7}	0.7	9.48×10^{-7}
0.2	0	0.5	0.008357	0.8	0
0.3	0	0.6	0.022250	0.9	0

Thereafter, you calculate the posterior probability

$$p(\theta_i) = P(\theta = \theta_i \mid 140 \text{ times heads in 250 tosses}).$$

This is done by applying Bayes' rule

$$p(\theta_i) = \frac{L(\theta_i) p_0(\theta_i)}{\sum_{k=1}^{9} L(\theta_k) p_0(\theta_k)} \qquad \text{for } i = 1, \ldots, 9.$$

This rule follows from arguments that are familiar by now

$$\begin{aligned} &P(\theta = \theta_i \mid 140 \text{ times heads in 250 tosses}) \\ &\quad = \frac{P(\theta = \theta_i \text{ and 140 times heads in 250 tosses})}{P(140 \text{ heads in 250 tosses})} \\ &\quad = \frac{P(140 \text{ heads in 250 tosses} \mid \theta = \theta_i) p_0(\theta_i)}{\sum_{k=1}^{9} P(140 \text{ heads in 250 tosses} \mid \theta = \theta_k) p_0(\theta_k)}. \end{aligned}$$

Applying Bayes' rule gives the following probabilities $p(\theta_i)$:

θ_i	$p(\theta_i)$	θ_i	$p(\theta_i)$	θ_i	$p(\theta_i)$
0.1	0	0.4	2.98×10^{-6}	0.7	1.62×10^{-5}
0.2	0	0.5	0.42895	0.8	0
0.3	0	0.6	0.57102	0.9	0

Comparing the posterior distribution with the prior distribution, one can see the effect of the data. An attractive property of the Bayesian approach is that when extra data become available, previous data do not lose their value. When new data become available, one takes the current posterior distribution as the new prior distribution and adjusts it as shown above. For example, imagine that 250 additional tosses of a one-euro coin land heads up 127 times. The above noted posterior distribution would then be adjusted as follows (verify!):

θ_i	$p(\theta_i)$	θ_i	$p(\theta_i)$	θ_i	$p(\theta_i)$
0.1	0	0.4	0	0.7	0
0.2	0	0.5	0.9821	0.8	0
0.3	0	0.6	0.0179	0.9	0

One could also have arrived at this posterior distribution by adjusting the original prior distribution $p_0(\theta_i)$ on the basis of $140 + 127 = 267$ times heads in $250 + 250 = 500$ tosses of the coin! Finally, it is notable that the posterior distribution becomes increasingly less sensitive to the originally chosen prior distribution as the available data increase. For example, the prior distribution $p_0(\theta_i) = \frac{1}{9}$ for $i = 1, \ldots, 9$ leads to the following posterior distribution when 500 tosses turn up heads 267 times:

θ_i	$p(\theta_i)$	θ_i	$p(\theta_i)$	θ_i	$p(\theta_i)$
0.1	0	0.4	0	0.7	0
0.2	0	0.5	0.9649	0.8	0
0.3	0	0.6	0.0351	0.9	0

The posterior distributions in the last two tables are quite similar, even though the priors are far apart.

Example 8.9 Two candidates A and B are contesting the election of governor in a given state. The candidate who wins the popular vote becomes governor. A random sample of the voting population is undertaken to find out the preference of the voters. The sample size of the poll is 1,000 and 517 of the polled voters favor candidate A. What can be said about the probability of candidate A winning the election?

Solution. The number of respondents in the poll who favor candidate A has a binomial distribution whose success probability represents the fraction of the voting population in favor of candidate A. Let's assume that, prior to polling, this success probability has the following prior distribution $p_0(\theta_i)$ on the possible values $\theta = 0.30, 0.31, \ldots, 0.69, 0.70$

$$p_0(\theta) = \begin{cases} \frac{\theta - 0.29}{4.41} & \text{for } \theta = 0.30, \ldots, 0.50, \\ \frac{0.71-\theta}{4.41} & \text{for } \theta = 0.51, \ldots, 0.70. \end{cases}$$

Hence, the prior probability of candidate A getting the majority of the votes at the election is $p_0(0.51) + \cdots + p_0(0.70) = 0.476$. However, 517 of the 1,000 polled voters favor candidate A. In light of this new information, what is the probability of candidate A getting the majority of the votes at the time of election? This probability is given by $p(0.51) + \cdots + p(0.70)$, where $p(\theta)$ is the posterior probability that the fraction of the voting population in favor of candidate A equals θ. This posterior probability is easily calculated from

$$p(\theta) = \frac{\binom{1000}{517}\theta^{517}(1-\theta)^{1000-517}p_0(\theta)}{\sum_{a=30}^{70}\binom{1000}{517}\left(\frac{a}{100}\right)^{517}\left(1-\frac{a}{100}\right)^{1000-517}p_0\left(\frac{a}{100}\right)}.$$

Performing the numerical calculations, we find that the posterior probability of candidate A getting the majority of the votes at the election equals

$$p(0.51) + \cdots + p(0.70) = 0.7632.$$

The posterior probability of a tie at the election equals $p(0.50) = 0.1558$.

Reasoning with conditional probabilities can be quite subtle, as is shown by the next example.

Example 8.10 A diamond merchant has lost a case with a very expensive diamond somewhere in a large city in an isolated area. The case has been found again but the diamond has vanished. However, the empty case contains DNA of the person who took the diamond. The city has 150,000 inhabitants who are each eligible for taking the diamond. An expert declares that the probability of a randomly chosen person matching the DNA profile is 10^{-6}. The police search a database with 5,120 DNA profiles and find one person matching the DNA from the case. Apart from the DNA evidence, there is no additional background evidence related to the suspect. On the basis of the extreme infrequency of the DNA profile and the fact that the population of potential penetrators is only 150,000 people, the prosecutor jumps to the conclusion that the odds of the suspect not being the thief are practically nil and calls for a tough sentence. What do you think of this conclusion?

Solution. The conclusion made by the prosecutor could not be more wrong. The prosecutor argues: "The probability that a person chosen at random would match the DNA profile found on the diamond case is negligible and the number of inhabitants of the city is not very large. The suspect matches this DNA profile, thus it is nearly one hundred percent certain that he is the perpetrator." This is a textbook example of the faulty use of probabilities. The probability that the suspect is innocent of the crime is altogether different from the probability that a randomly chosen person matches the DNA profile in question. What we are actually looking for is the probability that among all persons matching the DNA profile in question the arrested person is the perpetrator. Counsel for defense could reason as follows to estimate this probability: "We know that the suspect matches, but among the other $150{,}000 - 5{,}120 = 144{,}880$ individuals the expected number of people matching the DNA profile is $144{,}880 \times 10^{-6} = 0.14488$. So the probability that the suspect is guilty is $1/(1 + 0.14488) = 0.8735$. It is not beyond reasonable doubt that the suspect is guilty and thus the suspect must be released." The intuitive reasoning of the counsel of the defense leads to an exact estimate for the true probability of guilt, as will be shown next by Bayesian calculation.

Let us call a person matching the DNA profile on the empty case a D-person. In the situation before the diamond was lost, the probability distribution of the number of D-persons among the population of the potential perpetrators can be accurately modeled by a Poisson distribution with expected value $\lambda = 150{,}000 \times 10^{-6} = 0.15$ (see Section 4.2.1). That is, letting B_m be the event that the population of the potential perpetrators contains m people matching the

DNA profile in question, the prior probabilities

$$p_m^{(0)} = P(B_m) \qquad \text{for } m = 0, 1, \ldots$$

are modeled by the Poisson probabilities

$$p_m^{(0)} = e^{-\lambda}\frac{\lambda^m}{m!} \qquad \text{for } m = 0, 1, \ldots.$$

First we adjust the prior probabilities for the information that the diamond was taken by a type-D person. This leads to the posterior probabilities

$$p_m^{(1)} = P(B_m \mid A) \qquad \text{for } m = 1, 2, \ldots,$$

where A is the event that the diamond was taken by a type-D person. Taking $P(A \mid B_m) = m/150{,}000$, it follows that $p_m^{(1)} = P(B_m \mid A)$ is given by

$$\begin{aligned} p_m^{(1)} &= \frac{P(AB_m)}{P(A)} = \frac{P(A \mid B_m)P(B_m)}{\sum_{k=0}^{\infty} P(A \mid B_k)P(B_k)} \\ &= \frac{(m/150{,}000)p_m^{(0)}}{\sum_{k=0}^{\infty}(k/150{,}000)p_k^{(0)}} = \frac{1}{\lambda}mp_m^{(0)} \qquad \text{for } m = 1, 2, \ldots. \end{aligned}$$

Next we adjust the posterior probabilities $p_m^{(1)}$ for the information that the search through the database with 5,120 DNA profiles yielded exactly one person matching the DNA profile in question. The adjusted posterior probabilities $p_m^{(2)}$ are given by

$$p_m^{(2)} = P(B_m \mid AC) \qquad \text{for } m = 1, 2, \ldots,$$

where C denotes the event that a search among 5,120 randomly chosen people yields exactly one D-person (it is reasonable to consider the group of 5,120 people in the database of the police as a randomly composed group for the situation considered). Using the fact that

$$P(AC) = \sum_{k=0}^{\infty} P(ACB_k) = \sum_{k=0}^{\infty} P(C \mid AB_k)P(AB_k),$$

we find that

$$\begin{aligned} p_m^{(2)} = P(B_m \mid AC) &= \frac{P(C \mid AB_m)P(AB_m)}{\sum_{k=0}^{\infty} P(C \mid AB_k)P(AB_k)} \\ &= \frac{P(C \mid AB_m)P(B_m \mid A)\,P(A)}{\sum_{k=0}^{\infty} P(C \mid AB_k)P(B_k \mid A)\,P(A)}. \end{aligned}$$

We now observe that $P(C \mid AB_k) = P(C \mid B_k)$ for all k. The probability $P(C \mid B_k)$ is nothing else than the probability of getting one red ball when

5,120 balls are drawn at random without replacement from an urn containing k red balls and 150,000–k white balls. Abbreviating $P(C \mid B_k)$ as c_k, we have (see the hypergeometric distribution in Section 4.3 of Chapter 4)

$$c_k = \frac{\binom{k}{1}\binom{150{,}000-k}{5{,}119}}{\binom{150{,}000}{5{,}120}} \qquad \text{for } k = 1, 2, \ldots .$$

Putting the pieces together, we find

$$p_m^{(2)} = \frac{c_m\, p_m^{(1)}}{\sum_{k=0}^{\infty} c_k\, p_k^{(1)}} \qquad \text{for } m = 1, 2, \ldots .$$

The conditional probability of the suspect being the perpetrator is $\frac{1}{m}$ when there are in total m people matching the DNA profile in question. Hence, by averaging $\frac{1}{m}$ over the posterior probabilities $p_m^{(2)}$, we find that

$$P(\text{the suspect is guilty}) = \sum_{m=1}^{\infty} \frac{1}{m} \times p_m^{(2)}.$$

The value of this probability is calculated as 0.8735. This means that there is a rather troubling probability of more than 12% that the suspect is not the thief of the diamond!

The Bayesian calculation needs only a minor modification when it is assumed that the suspect from the database is twice as likely the thief as any other person with the DNA profile. Then, $P(\text{the suspect is guilty}) = \sum_{m=1}^{\infty} \frac{2}{m+1} p_m^{(2)} =$ 0.9142.

9
Basic rules for discrete random variables

In the first part of this book, we worked many times with models of random variables. In performing a chance experiment, one is often not interested in the particular outcome that occurs but in a specific numerical value associated with that outcome. Any function that assigns a real number to each outcome in the sample space of the experiment is called a *random variable*. The purpose of this chapter is to familiarize the reader with a number of basic rules for calculating characteristics of random variables such as the expected value and the variance. These rules are easiest explained and understood in the context of discrete random variables. Therefore, the discussion in this chapter is restricted to the case of discrete random variables. However, the rules for discrete random variables apply with obvious modifications to other types of random variables as well. In Chapter 10, we discuss so-called continuous random variables. Such random variables have a continuous interval as the range of possible values.

9.1 Random variables

Intuitively, a random variable is a variable that takes on its values by chance. The convention is to use capital letters such as X, Y, Z to denote random variables. Formally, a random variable is defined as a real-valued function on the sample space of a chance experiment. A random variable X assigns a numerical value $X(\omega)$ to each element ω of the sample space. For example, if X is the sum of the dots when rolling twice one fair die, the random variable X assigns the numerical value $i + j$ to the outcome (i, j) of the chance experiment. As said before, a random variable X takes on its values by chance. A random variable X gets its value only after the underlying chance experiment has been performed. Before the experiment, we can only describe the set of possible values of X. The probabilities associated with these possible values are determined by the probability measure P on the sample space of the chance experiment. In the above example, the possible values of the random variable X are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12 with associated probabilities $\frac{1}{36}, \frac{2}{36}, \frac{3}{36}, \frac{4}{36}, \frac{5}{36}, \frac{6}{36}, \frac{5}{36}, \frac{4}{36}, \frac{3}{36}, \frac{2}{36}$, and $\frac{1}{36}$. A random variable X is said to be *discrete* if its set of possible values is finite or countably infinite. The set of possible values of X is called the range of X and is denoted by I. The *probability mass function* of a discrete random variable X is defined by $P(X = x)$ for $x \in I$, where the notation $P(X = x)$ is shorthand for the probability mass assigned by the probability measure P to the set of all outcomes ω for which $X(\omega) = x$. Many examples of discrete random variables can be found in the Chapters 2–4. In particular, the reader may wish to read Section 2.2 of Chapter 2. We give here one other example.

Example 9.1 In your pocket you have three dimes (coins of 10 cents) and two quarters (coins of 25 cents). You grab at random two coins from your pocket. What is the probability mass function of the amount you grabbed?

Solution. The sample space of the chance experiment is chosen as $\Omega = \{(D, D), (D, Q), (Q, D), (Q, Q)\}$. The outcome (D, D) occurs if the first coin taken is a dime and the second one is also a dime, the outcome (D, Q) occurs if the first coin taken is a dime and the second one is a quarter, etc. The probability $\frac{3}{5} \times \frac{2}{4} = \frac{3}{10}$ is assigned to the outcome (D, D), the probability $\frac{3}{5} \times \frac{2}{4} = \frac{3}{10}$ to the outcome (D, Q), the probability $\frac{2}{5} \times \frac{3}{4} = \frac{3}{10}$ to the outcome (Q, D), and the probability $\frac{2}{5} \times \frac{1}{4} = \frac{1}{10}$ to the outcome (Q, Q). Let the random variable X denote the total number of cents you have grabbed. The random variable X has 20, 35, and 50 as possible values. The random variable X takes on the value 20 if the outcome (D, D) occurs, the value 35 if either the outcome (D, Q) or (Q, D) occurs, and the value 50 if the outcome (Q, Q) occurs. Thus, the probability mass function of X is given by $P(X = 20) = \frac{3}{10}$, $P(X = 35) = \frac{3}{10} + \frac{3}{10} = \frac{3}{5}$, and $P(X = 50) = \frac{1}{10}$.

9.2 Expected value

The most important characteristic of a random variable is its *expected value*. Synonyms for expected value are *expectation, mean*, and *first moment*. In Chapter 2 we informally introduced the concept of expected value. The expected value of a discrete random variable is a weighted mean of the values the random variable can take on, the weights being furnished by the probability mass function of the random variable. The nomenclature of expected value may be misleading. The expected value is in general not a typical value that the random variable can take on. It is often helpful to interpret the expected value of a random variable as the long-run average value of the variable over many independent repetitions of an experiment (see also Section 2.3 of Chapter 2).

Definition 9.1 *The expected value of the discrete random variable X having I as its set of possible values is defined by*

$$E(X) = \sum_{x \in I} x \, P(X = x).$$

Before we give some examples, note the following remarks. Definition 9.1 is only meaningful if the sum is well defined. The sum is always well defined if the range I is finite. However, the sum over countably infinite many terms is not always well defined when both positive and negative terms are involved. For example, the infinite series $1 - 1 + 1 - 1 + \ldots$ has the sum 0 when you sum the terms according to $(1 - 1) + (1 - 1) + \ldots$, whereas you get the sum 1 when you sum the terms according to $1 + (-1 + 1) + (-1 + 1) + (-1 + 1) + \cdots$. Such abnormalities cannot happen when all terms in the infinite summation are nonnegative. The sum of infinitely many *nonnegative* terms is always well defined, with ∞ as a possible value for the sum. For a sequence $a_1, a_2, \ldots$ consisting of both positive and negative terms, a basic result from the theory of series states that the infinite series $\sum_{k=1}^{\infty} a_k$ is always well defined with a finite sum if the series is absolutely convergent, where absolute convergence means that $\sum_{k=1}^{\infty} |a_k| < \infty$. In case the series $\sum_{k=1}^{\infty} a_k$ is absolutely convergent, the sum is uniquely determined and does not depend on the order in which the individual terms are added. For a discrete random variable X with range I, it is said that the expected value $E(X)$ *exists* if X is nonnegative or if $\sum_{x \in I} |x| \, P(X = x) < \infty$. An example of a random variable X for which $E(X)$ does not exist is the random variable X with probability mass function $P(X = k) = \frac{3}{\pi^2 k^2}$ for $k = \pm 1, \pm 2, \ldots$ (a celebrated result from calculus is that $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$). The reason that $E(X)$ does not exist is the well-known fact from calculus that $\sum_{k=1}^{\infty} 1/k = \infty$.

Example 9.1 (continued) What is the the expected value of the amount you grabbed from your pocket?

Solution. Since the probability mass function of the number of cents you grabbed from your pocket is given by $P(X = 20) = \frac{3}{10}$, $P(X = 35) = \frac{3}{5}$, and $P(X = 50) = \frac{1}{10}$, the expected value of the amount you grabbed is equal to

$$E(X) = 20 \times \frac{3}{10} + 35 \times \frac{3}{5} + 50 \times \frac{1}{10} = 32 \text{ cents.}$$

Example 9.2 Joe and his friend make a guess every week whether the Dow Jones index will have risen at the end of the week or not. Both put $10 in the pot. Joe observes that his friend is just guessing and is making his choice by the toss of a fair coin. Joe asks his friend if he could contribute \$20 to the pot and submit his guess together with that of his brother. The friend agrees. In each week, however, Joe's brother submits a prediction opposite to that of Joe. The person having a correct prediction wins the entire pot. If more than one person has a correct prediction, the pot is split evenly. How favorable is the game to Joe and his brother?

Solution. Let the random variable X denote the winnings of Joe and his brother in any given week. Either Joe or his brother will have a correct prediction. If Joe's friend is wrong he wins nothing, and if he is correct he shares the \$30 pot with either Joe or his brother. Thus, X takes on the values 30 and 15 with equal chances. This gives $E(X) = \frac{1}{2} \times 30 + \frac{1}{2} \times 15 = 22.5$ dollars. Joe and his brother have an expected profit of \$2.5 every week.

Example 9.3 Three friends go to the cinema together every week. Each week, in order to decide which friend will pay for the other two, they all toss a fair coin into the air simultaneously. They continue to toss coins until one of the three gets a different outcome from the other two. What is the expected value of the number of trials required?

Solution. Let the random variable X denote the number of trials until one of the three friends gets a different outcome from the other two. The probability that any given trial does not lead to three equal outcomes is $p = 1 - \frac{1}{8} - \frac{1}{8} = \frac{3}{4}$. Thus

$$P(X = j) = (1-p)^{j-1}p \qquad \text{for } j = 1, 2, \ldots$$

with $p = \frac{3}{4}$. The expected value of X is given by

$$E(X) = \sum_{j=1}^{\infty} j(1-p)^{j-1}p = p\sum_{j=1}^{\infty} j(1-p)^{j-1} = \frac{p}{[1-(1-p)]^2} = \frac{1}{p},$$

using the fact that $\sum_{j=1}^{\infty} jx^{j-1} = 1/(1-x)^2$ for all $0 < x < 1$ (see the Appendix). Hence the expected value of the number of trials required is $\frac{4}{3}$.

The last two examples illustrate the fact that an explicit listing of the sample space is not always needed for the solution of a probability problem. In most problems, you will perform probability calculations without explicitly specifying a sample space; an assignment of probabilities to properly chosen events usually suffices.

Problem 9.1 You are playing a game in which four fair dice are rolled. A \$1 stake is required. The payoff is \$100 if all four dice show the same number and \$10 if two dice show the same even or odd number. What is an appropriate probability space for this experiment, and what is the expected value of the payoff?

Problem 9.2 Calculate the expected value of the greater of two numbers when two different numbers are picked at random from the numbers $1, \ldots, n$. What is the expected value of the absolute difference between the two numbers?

Problem 9.3 You spin a game board spinner with 1,000 equal sections numbered as $1, 2, \ldots,$ 1,000. After your first spin, you have to decide whether to spin the spinner for a second time. Your payoff is the total score of your spins as long as this score does not exceed 1,000; otherwise, your payoff is zero. What strategy maximizes the expected value of your payoff?

Problem 9.4 In a lottery, one thousand tickets numbered as $000, 001, \ldots, 999$ are sold. Each contestant buys only one ticket. The prize winners of the lottery are determined by drawing at random one number from the numbers $000, 001, \ldots, 999$. You are a prize winner when the number on your ticket is the same as the number drawn or is a random permutation of the number drawn. What is the probability mass function of the number of prize winners? What is the expected value of the number of prize winners and what is the probability that a randomly picked contestant will be a prize winner?

Problem 9.5 A stick is broken at random into two pieces. You bet on the ratio of the length of the longer piece to the length of the smaller piece. You receive \$$k$ if the ratio is between k and $k+1$ for some k with $1 \leq k \leq m-1$, while you receive \$$m$ if the ratio is larger than m. Here m is a given positive integer. What should be your stake to make this a fair bet? Verify that your stake should be $\$2[1 + \frac{1}{2} + \cdots + \frac{1}{m+1} - 1]$ (this amount is approximately equal to $\$2[\ln(m+1) + \gamma - 1 + \frac{1}{2(m+1)}]$ for m large, where $\gamma = 0.57722\ldots$ is Euler's constant).

Problem 9.6 Mary and Peter play the following game. They toss a fair coin until heads appears for the first time or m tosses are done, whichever occurs first. Here m is fixed in advance. If heads appears at the kth toss, then Peter pays Mary 2^k dollars when k is odd and otherwise Mary pays Peter 2^k dollars. Denote by E_m the expected value of Mary's net gain. Give an expression for E_m and calculate E_m for m=5, 10, and 20. What is $\lim_{m\to\infty} E_m$? Next consider the game without limit on the number of tosses. Let the random variable X^+ be the amount Mary will receive and let X^- the amount Mary will have to pay. What are the values of $E(X^+)$ and $E(X^-)$? Does the expected value of $X^+ - X^-$, Mary's net gain, exist?

Problem 9.7 Suppose that the random variable X is nonnegative and integer-valued. Verify that $E(X) = \sum_{k=0}^{\infty} P(X > k)$.

9.3 Expected value of sums of random variables

Let X and Y be two random variables that are defined on the same sample space with probability measure P. For example, for the experiment of rolling two dice, X is the smallest of the two outcomes and Y is the sum of the two outcomes. The following basic rule is of utmost importance.

Rule 9.1 *For any two random variables X and Y*

$$E(X + Y) = E(X) + E(Y),$$

provided that $E(X)$ and $E(Y)$ exist.

The proof is simple for the discrete case. Letting $Z = X + Y$, a key observation is

$$P(Z = z) = \sum_{x,y:\ x+y=z} P(X = x, Y = y),$$

where $P(X = x, Y = y)$ is the notation for the probability of the joint event that X takes on the value x and Y the value y. Also, we need the relations $\sum_y P(X = x, Y = y) = P(X = x)$ and $\sum_x P(X = x, Y = y) = P(Y = y)$. Thus

$$\begin{aligned} E(Z) &= \sum_z zP(Z = z) = \sum_z z \sum_{x,y:\ x+y=z} P(X = x, Y = y) \\ &= \sum_z \sum_{x,y:\ x+y=z} (x + y)P(X{=}x, Y = y) = \sum_{x,y} (x + y)P(X = x, Y = y) \end{aligned}$$

and so

$$\begin{aligned} E(Z) &= \sum_{x,y} xP(X = x, Y = y) + \sum_{x,y} yP(X = x, Y = y) \\ &= \sum_x x \sum_y P(X = x, Y = y) + \sum_y y \sum_x P(X = x, Y = y) \\ &= \sum_x xP(X = x) + \sum_y yP(Y = y), \end{aligned}$$

which proves the desired result $E(Z) = E(X) + E(Y)$. The same result holds for any finite number of random variables, each having a finite expected value. That is

$$E\left(X_1 + \ldots + X_n\right) = E\left(X_1\right) + \ldots + E\left(X_n\right)$$

if $E(X_i)$ exists for all $i = 1, \dots, n$. The result that the expected value of a finite sum of random variables equals the sum of the expected values is extremely useful. It is only required that the relevant expected values exist, but dependencies between the random variables are allowed. The utility of this result has already been demonstrated by several examples in Chapters 2 and 3. A trick that is often applicable to calculate the expected value of a random variable is to represent the random variable as the sum of random variables that can take on only values 0 and 1.

Example 9.4 Suppose that n children of differing heights are placed in line at random. You then select the first child from the line and walk with her/him along the line until you encounter a child who is taller or until you have reached the end of the line. If you do encounter a taller child, you also have her/him accompany you further along the line until you encounter yet again a taller child or reach the end of the line, and so on. What is the expected value of the number of children selected from the line?

Solution. Letting the random variable X denote the number of children selected from the line, we can most easily compute $E(X)$ by writing

$$X = X_1 + \cdots + X_n,$$

where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th child is selected from the line} \\ 0 & \text{otherwise.} \end{cases}$$

The probability that the ith child is the tallest among the first i children equals $\frac{1}{i}$. Hence

$$E(X_i) = 0 \times \left(1 - \frac{1}{i}\right) + 1 \times \frac{1}{i} = \frac{1}{i}, \qquad i = 1, \dots, n.$$

This gives

$$E(X) = 1 + \frac{1}{2} + \dots + \frac{1}{n}.$$

An insightful approximation can be given to this expected value. It is known from calculus that $1 + \frac{1}{2} + \dots + \frac{1}{n}$ can very accurately be approximated by $\ln(n) + \gamma + \frac{1}{2n}$, where $\gamma = 0.57722\dots$ is Euler's constant.

Example 9.5 What is the expected value of the number of times that in a thoroughly shuffled deck of 52 cards two adjacent cards are of the same rank (two aces, two kings, etc.)?

Solution. Let the random variable X_i be equal to 1 if the cards in the positions i and $i + 1$ are of the same kind and 0 otherwise. Then, $P(X_i = 1) = \frac{3}{52}$ and so $E(X_i) = \frac{3}{52}$ for $i = 1, \ldots, 51$. The expected value of the number of times that two adjacent cards are of the same rank is given by $E(X_1 + \ldots + X_{51}) = 51 \times \frac{3}{52} = 2.942$.

The linearity property of the expected value is a special case of a general result in calculus for sums and integrals. This property holds not only for discrete random variables, but also for any type of random variables. Another important type of random variable is the continuous random variable with a continuous interval as its range of possible values. Continuous random variables are to be discussed in Chapter 10 and subsequent chapters. The models of discrete and continuous random variables are the most important ones, but are not exhaustive. Also, there are so-called *mixed* random variables having properties of both discrete and continuous random variables. Think of your delay in a queue at a counter in a supermarket or the amount paid on an automobile insurance policy in a given year. These random variables take on either the discrete value zero with positive probability or a value in a continuous interval.

Problem 9.8 Consider Example 7.13 again. Calculate the expected number of hotels that remain empty. *Hint*: define the random variable X_i as equal to 1 if the ith hotel remains empty and 0 otherwise.

Problem 9.9 What is the expected number of distinct birthdays within a randomly formed group of 100 persons?

Problem 9.10 What is the expected value of the number of times that two adjacent letters are the same in a random permutation of the word Mississippi?

Problem 9.11 What is the expected value of the number of combinations of two consecutive numbers in a lottto drawing of six different numbers from the numbers $1, 2, \ldots, 45$?

9.4 Substitution rule and variance

Suppose X is a discrete random variable with a given probability mass function. In many applications, we wish to compute the expected value of some function of X. Note that any function of X (e.g., X^2 or $\sin(X)$) is also a random variable. Let $g(x)$ be a given real-valued function. Then the quantity $g(X)$ is a discrete random variable as well. The expected value of $g(X)$ can directly be calculated from the probability distribution of X.

Rule 9.2 *For any function g of the random variable X*

$$E\left[g(X)\right] = \sum_{x \in I} g(x)\, P(X = x)$$

provided that $\sum_{x \in I} |g(x)|\ P(X = x) < \infty$.

This rule is called the *substitution rule*. The proof of the rule is simple. If X takes on the values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$ and it is assumed that $g(x_i) \neq g(x_j)$ for $x_i \neq x_j$, then the random variable $Z = g(X)$ takes on the values $z_1 = g(x_1), z_2 = g(x_2), \ldots$ with the same probabilities $p_1, p_2, \ldots$. Next apply the definition $E(Z) = \sum_k z_k P(Z = z_k)$ and substitute $z_k = g(x_k)$ and $P(Z = z_k) = P(X = x_k)$. The proof needs an obvious modification when the assumption $g(x_i) \neq g(x_j)$ for $x_i \neq x_j$ is dropped.

A frequently made mistake of beginning students is to set $E\left[g(X)\right]$ equal to $g\left(E(X)\right)$. In general, $E\left[g(X)\right] \neq g\left(E(X)\right)$! Stated differently, the average value of the input X does not determine in general the average value of the output $g(X)$. As a counterexample, take the random variable X with $P(X = 1) = P(X = -1) = 0.5$ and take the function $g(x) = x^2$. An exception is the case of a linear function $g(x) = ax + b$. An immediate consequence of Rule 9.2 is:

Rule 9.3 *For any constants a and b*

$$E(aX + b) = aE(X) + b.$$

9.4.1 Variance

An important case of a function of X is the random variable $g(X) = (X - \mu)^2$, where $\mu = E(X)$ denotes the expected value of X. The expected value of $(X - \mu)^2$ is called the *variance* of X and is denoted by

$$\text{var}(X) = E[(X - \mu)^2].$$

It is a measure of the spread of the possible values of X. Often one uses the *standard deviation*, which is defined as the square root of the variance. It is useful to work with the standard deviation since it has the same units (e.g., dollar or cm) as $E(X)$. The standard deviation of a random variable X is usually denoted by $\sigma(X)$ and thus is defined by

$$\sigma(X) = \sqrt{\text{var}(X)}.$$

The formula for var(X) allows for another useful representation. Since $(X - \mu)^2 = X^2 - 2\mu X + \mu^2$, it follows from the linearity of the expectation operator and Rule 9.3 that $E[(X - \mu)^2] = E(X^2) - 2\mu E(X) + \mu^2$. Hence var($X$) is also given by

$$\text{var}(X) = E(X^2) - \mu^2.$$

Rule 9.3 for the expectation operator has the following analog for the variance operator:

Rule 9.4 *For any constants a and b*

$$\text{var}(aX + b) = a^2\text{var}(X).$$

The proof is left as an exercise to the reader.

Example 9.6 What is the variance of the sum of the dots when rolling two dice?

Solution. Let the random variable X denote the total score. Using the fact that $E(X) = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + 5 \times \frac{4}{36} + 6 \times \frac{5}{36} + 7 \times \frac{6}{36} + 8 \times \frac{5}{36} + 9 \times \frac{4}{36} + 10 \times \frac{3}{36} + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} = 7$, we find that

$$\begin{aligned}\text{var(X)} = {} & 2^2 \times \frac{1}{36} + 3^2 \times \frac{2}{36} + 4^2 \times \frac{3}{36} + 5^2 \times \frac{4}{36} + 6^2 \times \frac{5}{36} + 7^2 \times \frac{6}{36} \\ & + 8^2 \times \frac{5}{36} + 9^2 \times \frac{4}{36} + 10^2 \times \frac{3}{36} + 11^2 \times \frac{2}{36} + 12^2 \times \frac{1}{36} - 7^2 \\ = {} & 5\frac{5}{6}.\end{aligned}$$

The standard deviation of X is $\sqrt{\text{var}(X)} = 2.415$ dots.

Example 9.7 Suppose the random variable X has the Poisson distribution $P(X = k) = e^{-\lambda}\lambda^k/k!$ for $k = 0, 1, \ldots$. What are the expected value and the variance of X?

Solution. A remarkable property of the Poisson distribution is that its variance has the same value as its mean. That is

$$\text{var}(X) = E(X) = \lambda.$$

We only verify that var(X)=λ. In Section 4.2.1 it was shown that $E(X) = \lambda$. To evaluate $E(X^2)$, use the identity $k^2 = k(k-1) + k$. This gives

$$\begin{aligned}E(X^2) &= \sum_{k=0}^{\infty} k^2 P(X = k) \\ &= \sum_{k=1}^{\infty} k(k-1)P(X = k) + \sum_{k=1}^{\infty} kP(X = k) \\ &= \sum_{k=1}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} + E(X) = \lambda^2 \sum_{k=2}^{\infty} e^{-\lambda}\frac{\lambda^{k-2}}{(k-2)!} + \lambda.\end{aligned}$$

Since $\sum_{k=2}^{\infty} e^{-\lambda}\lambda^{k-2}/(k-2)! = \sum_{n=0}^{\infty} e^{-\lambda}\lambda^{n}/n! = 1$, we obtain $E(X^2) = \lambda^2 + \lambda$. Next, by $\text{var}(X) = E(X^2) - (E(X))^2$, the desired result follows.

The next example deals with the famous newsboy problem.

Example 9.8 Every morning, rain or shine, young Billy Gates can be found at the entrance to the metro, hawking copies of "The Morningstar." Demand for newspapers varies from day to day, but Billy's regular early morning haul yields him 200 copies. He purchases these copies for \$1 per paper, and sells them for \$1.50 apiece. Billy goes home at the end of the morning, or earlier if he sells out. He can return unsold papers to the distributor for \$0.50 apiece. From experience, Billy knows that demand for papers on any given morning is uniformly distributed between 150 and 250, where each of the possible values 150, ..., 250 is equally likely. What are the expected value and the standard deviation of Billy's net earnings on any given morning?

Solution. Denote by the random variable X the number of copies Billy would have sold on a given morning if he had ample supply. The actual number of copies sold by Billy is X if $X \leq 200$ and 200 otherwise. The probability mass function of X is given by $P(X = k) = \frac{1}{101}$ for $k = 150, \ldots, 250$. Billy's net earnings on any given morning is a random variable $g(X)$, where the function $g(x)$ is given by

$$g(x) = \begin{cases} -200 + 1.5x + 0.5(200 - x), & x \leq 200 \\ -200 + 1.5 \times 200, & x > 200. \end{cases}$$

Applying the substitution rule, we find that $E[g(X)]$ is given by

$$\sum_{k=150}^{250} g(k)P(X = k) = \frac{1}{101}\sum_{k=150}^{200}(-100 + k) + \frac{1}{101}\sum_{k=201}^{250} 100$$

and so

$$E[g(X)] = \frac{3{,}825}{101} + \frac{5{,}000}{101} = 87.3762.$$

To find the standard deviation of $g(X)$, we apply the formula $\text{var}(Z) = E(Z^2) - (E(Z))^2$ with $Z = g(X)$. This gives

$$\text{var}[g(X)] = E[(g(X))^2] - (E[g(X)])^2.$$

Letting $h(x) = (g(x))^2$, then $h(x) = (-100 + x)^2$ for $x \leq 200$ and $h(x) = 100^2$ for $x > 200$. By applying the substitution rule again, we find that $E[h(X)]$ equals

$$\sum_{k=150}^{250} h(k)P(X = k) = \frac{1}{101}\sum_{k=150}^{200}(-100 + k)^2 + \frac{1}{101}\sum_{k=201}^{250} 100^2$$

and so

$$E[(g(X))^2] = E[h(X)] = \frac{297{,}925}{101} + \frac{500{,}000}{101} = 7900.2475.$$

Hence, the variance of Billy's net earnings on any given morning is

$$\text{var}[g(X)] = 7900.2475 - (87.3762)^2 = 265.64.$$

Concluding, Billy's net earnings on any given morning has an expected value of 87.378 dollars and a standard deviation of $\sqrt{265.64} = 16.30$ dollars.

Problem 9.12 Calculate the standard deviation of the random variables appearing in the Examples 9.1 and 9.2.

Problem 9.13 Consider Example 9.3 again. What is the standard deviation of the number of trials required?

Problem 9.14 At the beginning of every month, a pharmacist orders an amount of a certain costly medicine that comes in strips of individually packed tablets. The wholesale price per strip is \$100, and the retail price per strip is \$400. The medicine has a limited shelf life. Strips not purchased by month's end will have reached their expiration date and must be discarded. When it so happens that demand for the item exceeds the pharmacist's supply, he may place an emergency order for \$350 per strip. The monthly demand for this medicine takes on the possible values 3, 4, 5, 6, 7, 8, 9, and 10 with respective probabilities 0.3, 0.1, 0.2, 0.2, 0.05, 0.05, 0.05, and 0.05. The pharmacist decides to order eight strips at the start of each month. What are the expected value and the standard deviation of the net profit made by the pharmacist on this medicine in any given month?

Problem 9.15 The University of Gotham City renegotiates its maintenance contract with a particular copy machine distributor on a yearly basis. For the coming year, the distributor has come up with the following offer. For a prepaid cost of \$50 per repair call, the university can opt for a fixed number of calls. For each visit beyond that fixed number, the university will pay \$100. If the actual number of calls made by a repairman remains below the fixed number, no money will be refunded. Based on previous experience, the university estimates that the number of repairs that will be necessary in the coming year will have a Poisson distribution with an expected value of 150. The university signs a contract with a fixed number of 155 repair calls. What are the expected value and the standard deviation of the maintenance costs in excess of the prepaid costs?

9.5 Independence of random variables

In Chapter 8, we dealt with the concept of independent events. It makes intuitive sense to say that random variables are independent when the underlying events are independent. Let X and Y be two random variables that are defined on the same sample space with probability measure P. The following definition does not require that X and Y are discrete random variables but applies to the general case of two random variables X and Y.

Definition 9.2 *The random variables X and Y are said to be independent if*

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

for any two real numbers x and y, where $P(X \leq x, Y \leq y)$ represents the probability of occurrence of both event $\{X \leq x\}$ and event $\{Y \leq y\}$.[†]

In words, the random variables X and Y are independent if the event of the random variable X taking on a value smaller than or equal to x and the event of the random variable Y taking on a value smaller than or equal to y are independent for all real numbers x, y. Using the axioms of probability theory it can be shown that Definition 9.2 is equivalent to

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for any two sets A and B of real numbers. The technical proof is omitted. It is not difficult to verify the following two rules from the alternative definition of independence.

Rule 9.5 *If X and Y are independent random variables, then the random variables $f(X)$ and $g(Y)$ are independent for any two functions f and g.*

In the case that X and Y are discrete random variables, another representation of independence can be given.

Rule 9.6 *Discrete random variables X and Y are independent if and only if*

$$P(X = x, Y = y) = P(X = x)P(Y = y) \qquad \textit{for all } x, y.$$

A very useful rule applies to the calculation of the expected value of the product of two independent random variables.

[†] In general, the n random variables $X_1, \ldots, X_n$ are said to be independent if they satisfy $P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n)$ for each n-tuple of real numbers $x_1, \ldots, x_n$. An infinite collection of random variables is said to be independent if every finite subcollection of them is independent. In applications the independence or otherwise of random variables is usually obvious from the physical construction of the process.

Rule 9.7 *If the random variables X and Y are independent, then*

$$E(XY) = E(X)E(Y),$$

assuming that $E(X)$ and $E(Y)$ exist.

We prove this important result for the case of discrete random variables X and Y. Let I and J denote the sets of possible values of the random variables X and Y. Define the random variable Z by $Z = XY$, then

$$\begin{aligned}
E(Z) &= \sum_z zP(Z = z) = \sum_z z \sum_{\substack{x\in I, y\in J:\\ xy=z}} P(X = x, Y = y)\\
&= \sum_z \sum_{\substack{x\in I, y\in J:\\ xy=z}} xyP(X = x, Y = y)\\
&= \sum_{x\in I, y\in J} xyP(X = x, Y = y) = \sum_{x\in I, y\in J} xyP(X = x)P(Y = y)\\
&= \sum_{x\in I} xP(X = x)\sum_{y\in J} yP(Y = y) = E(X)E(Y).
\end{aligned}$$

The converse of the above result is not true. It is possible that $E(XY) = E(X)E(Y)$, while X and Y are not independent. A simple example is as follows. Suppose two fair dice are tossed. Denote by the random variable V_1 the number appearing on the first die and by the random variable V_2 the number appearing on the second die. Let $X = V_1 + V_2$ and $Y = V_1 - V_2$. It is readily seen that the random variables X and Y are not independent. We leave it to the reader to verify that $E(X) = 7$, $E(Y) = 0$, and $E(XY) = E(V_1^2 - V_2^2) = 0$ and so $E(XY) = E(X)E(Y)$.

In Rule 9.1 we proved that the expectation operator has the linearity property. This property holds for the variance operator only under an independence assumption.

Rule 9.8 *If the random variables X and Y are independent, then*

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

The proof is as follows. Putting $\mu_X = E(X)$ and $\mu_Y = E(Y)$, it follows that $\text{var}(X + Y) = E[(X + Y)^2] - (\mu_X + \mu_Y)^2$ can be worked out as

$$\begin{aligned}
&E[X^2 + 2XY + Y^2] - \mu_X^2 - 2\mu_X\mu_Y - \mu_Y^2\\
&\quad = E(X^2) + 2\mu_X\mu_Y + E(Y^2) - \mu_X^2 - 2\mu_X\mu_Y - \mu_Y^2,
\end{aligned}$$

where the latter equality uses the fact that $E(XY) = E(X)E(Y)$, by the independence of X and Y. This gives $\text{var}(X + Y) = E(X^2) - \mu_X^2 + E(Y^2) - \mu_Y^2 =$

var(X) + var(Y). The extension of Rule 9.8 to the case of $n > 2$ independent random variables is obvious.

Problem 9.16 Let X and Y be two independent random variables. What are the expected value and the variance of $X - Y$?

Problem 9.17 Two fair dice are tossed. Let the random variable X denote the sum of the two numbers shown by the dice and let Y be the largest of these two numbers. Are the random variables X and Y independent? What are the values of $E(XY)$ and $E(X)E(Y)$?

Problem 9.18 A drunkard is standing in the middle of a very large town square. He begins to walk. Each step is a unit distance in one of the four directions East, West, North, and South. All four possible directions are equally probable. The direction for each step is chosen independently of the direction of the others. The drunkard takes a total of n steps.

(a) Verify that the quadratic distance of the drunkard to his starting point after n steps has expected value n, irrespective of the value of n. *Hint*: the squared distance of the drunkard to his starting point after n steps can be written as $(\sum_{i=1}^{n} X_i)^2 + (\sum_{i=1}^{n} Y_i)^2$, where the random variables X_i and Y_i denote the changes in the x-coordinate and the y-coordinate of the position of the drunkard caused by his ith step.

(b) Use the definition of variance to explain why the expected value of the distance of the drunkard to his starting point after n steps cannot be equal to $\sqrt{n}$. *Hint*: use the fact that $P(X = c) = 1$ for some constant c if var$(X) = 0$.

Problem 9.19 Let X_i denote the number of integers smaller than i that precede i in a random permutation of the integers $1, \ldots, 10$. What are the expected value and the variance of the sum $X_2 + \cdots + X_{10}$?

Convolution formula

Suppose X and Y are two discrete random variables each having the set of nonnegative integers as the range of possible values. A useful rule is

Rule 9.9 *If the nonnegative random variables X and Y are independent, then*

$$P(X + Y = k) = \sum_{j=0}^{k} P(X = j)P(Y = k - j) \qquad \text{for } k = 0, 1, \ldots$$

This rule is known as the *convolution rule*. The proof is as follows. Fix k. Let A be the event that $X+Y=k$ and let B_j be the event that $X=j$ for $j=0,1,\ldots$. The events $AB_0, AB_1, \ldots$ are mutually exclusive and so, by Axiom 3 in Chapter 7, $P(A)=\sum_{j=0}^{\infty} P(AB_j)$. In other words

$$P(X+Y=k)=\sum_{j=0}^{\infty} P(X+Y=k, X=j).$$

Obviously, $P(X+Y=k, X=j)=P(X=j, Y=k-j)$ and so

$$P(X+Y=k, X=j)=P(X=j)P(Y=k-j) \quad \text{for all } j, k,$$

by the the independence of X and Y. Thus

$$P(X+Y=k)=\sum_{j=0}^{\infty} P(X=j)P(Y=k-j).$$

Since $P(Y=k-j)=0$ for $j>k$, the convolution formula next follows.

Example 9.9 Suppose the random variables X and Y are independent and have Poisson distributions with respective means λ and μ. What is the probability distribution of $X+Y$?

Solution. To answer this question, we apply the convolution formula. This gives

$$\begin{aligned} P(X+Y=k) &= \sum_{j=0}^{k} e^{-\lambda}\frac{\lambda^j}{j!}e^{-\mu}\frac{\mu^{k-j}}{(k-j)!} \\ &= \frac{e^{-(\lambda+\mu)}}{k!}\sum_{j=0}^{k}\binom{k}{j}\lambda^j\mu^{k-j}, \end{aligned}$$

where the second equality uses the fact that $\binom{k}{j}=\frac{k!}{j!(k-j)!}$. Next, by Newton's binomial $(a+b)^k=\sum_{j=0}^{k}\binom{k}{j}a^j b^{k-j}$, we find

$$P(X+Y=k)=e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^k}{k!} \qquad \text{for } k=0,1,\ldots.$$

Hence, $X+Y$ is Poisson distributed with mean $\lambda+\mu$.

Problem 9.20 Modify the convolution formula in Rule 9.9 when the random variables X and Y are integer-valued but not necessarily nonnegative.

Problem 9.21 You repeatedly draw a random integer from the integers $1,\ldots,10$ until you have three different integers. What is the probability that you need r draws?

9.6 Special discrete distributions

This section summarizes several discrete distributions which appear frequently in applications. Most of these distributions have already been discussed in Chapter 4 in the context of real-world problems. The material in Sections 4.1, 4.1.2 and 4.3 links up seamlessly with Chapter 9.

1. Bernoulli distribution A random variable X is said to have a Bernoulli distribution with parameter p if the random variable can only assume the values 1 or 0 with

$$P(X = 1) = p \quad \text{and} \quad P(X = 0) = 1 - p,$$

where $0 < p < 1$. A Bernoulli random variable X can be thought of as the outcome of an experiment that can only result in "success" or "failure." Such an experiment is called a Bernoulli trial. It is easily verified that the mean and variance of X are given by

$$E(X) = p \quad \text{and} \quad \text{var}(X) = p(1 - p).$$

2. Binomial distribution A random variable X is said to have a binomial distribution with parameters n and p if

$$P(X = k) = \begin{cases} \binom{n}{k} p^k (1 - p)^{n-k} & \text{for } k = 0, 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

The random variable X can be thought of as the total number of successes in n independent Bernoulli trials with probability p of success on each trial (see Section 4.1 for examples). The explanation is simple. The probability of getting k "successes" and $n - k$ "failures" in a specific order is $p^k(1 - p)^{n-k}$ and the number of ways in which we can choose the k trials on which there is to be a "success" is $\binom{n}{k}$. This gives the formula for $P(X = k)$. The random variable X can be written as $Y_1 + \cdots + Y_n$, where $Y_1, \ldots, Y_n$ are independent random variables each having a Bernoulli distribution with parameter p. Using the fact that $E(Y_i) = p$ and $\text{var}(Y_i) = p(1 - p)$, an application of Rules 9.1 and 9.8 gives

$$E(X) = np \quad \text{and} \quad \text{var}(X) = np(1 - p).$$

3. Hypergeometric distribution A random variable X is said to have a hypergeometric distribution with parameters R, W, and n if

$$P(X = k) = \frac{\binom{R}{k}\binom{W}{n-k}}{\binom{R+W}{n}} \qquad \text{for } k = 0, 1, \ldots, n$$

and $P(X = k) = 0$ otherwise. By the convention $\binom{a}{b} = 0$ for $b > a$, we also have that $P(X = k) = 0$ for those k with $k > R$ or $n - k > W$. The random variable X can be thought of as the number of red balls drawn when n balls are drawn at random without replacement from an urn containing R red balls and W white balls (see Section 4.3 for examples). The explanation is as follows. The number of ways in which k red balls and $n-k$ white balls can be chosen from the urn is $\binom{R}{k}\binom{W}{n-k}$ and the total number of ways in which n balls can be chosen from the urn is $\binom{R+W}{n}$. The ratio of these two expressions gives $P(X=k)$. It is a matter of straightforward but tedious algebra to verify that (see also Problem 11.17)

$$E(X)=n\frac{R}{R+W} \quad \text{and} \quad \operatorname{var}(X)=n\frac{R}{R+W}\left(1-\frac{R}{R+W}\right)\frac{R+W-n}{R+W-1}.$$

4. Discrete uniform distribution A random variable X is said to have a discrete uniform distribution on the integers $a, a+1, \ldots, b$ if

$$P(X=k)=\frac{1}{b-a+1} \qquad \text{for } k=a, a+1, \ldots, b.$$

The random variable X can be thought of as the result of an experiment with finitely many outcomes, each of which is equally likely. It is a matter of straightforward algebra to verify that

$$E(X)=\frac{a+b}{2} \quad \text{and} \quad \operatorname{var}(X)=\frac{(b-a+1)^2-1}{12}.$$

5. Geometric distribution A random variable X is said to have a geometric distribution with parameter p if

$$P(X=k)=\begin{cases} p(1-p)^{k-1} & \text{for } k=1,2,\ldots \\ 0 & \text{otherwise.} \end{cases}$$

The random variable X can be interpreted as the number of trials in an experiment in which independent Bernoulli trials with success probability p are performed until the first success occurs. Using the basic relations $\sum_{k=1}^{\infty} kx^{k-1}=(1-x)^{-2}$ and $\sum_{k=1}^{\infty} k(k-1)x^{k-2}=2(1-x)^{-3}$ for $|x|<1$ (see the Appendix), it is easily verified that

$$E(X)=\frac{1}{p} \quad \text{and} \quad \operatorname{var}(X)=\frac{1-p}{p^2}.$$

6. Negative binomial distribution A random variable X is said to have a negative binomial distribution with parameters r and p if

$$P(X=k)=\begin{cases} \binom{k-1}{r-1}p^r(1-p)^{k-r} & \text{for } k=r, r+1, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

The random variable X can be interpreted as the number of trials in an experiment in which independent Bernoulli trials with success probability p are performed until the rth success occurs. The explanation is as follows. The probability of having the rth success at the kth trial equals the binomial probability $\binom{k-1}{r-1}p^{r-1}(1-p)^{k-1-(r-1)}$ of having $r-1$ successes among the first $k-1$ trials multiplied with the probability p of having a success at the kth trial. The random variable X can be written as $Y_1 + \cdots + Y_n$, where Y_i is the number of trials needed in order to go from $i-1$ to i successes and $Y_1, \ldots, Y_n$ are independent random variables each having a geometric distribution with parameter p. Using the fact that $E(Y_i) = p$ and $\text{var}(Y_i) = (1-p)/p^2$, an application of Rules 9.1 and 9.8 gives

$$E(X) = r\frac{1}{p} \quad \text{and} \quad \text{var}(X) = r\frac{1-p}{p^2}.$$

7. Poisson distribution A random variable X is said to have a Poisson distribution with parameter $\lambda > 0$ if

$$P(X = k) = \begin{cases} e^{-\lambda}\frac{\lambda^k}{k!} & \text{for } k = 0, 1, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

The Poisson distribution can be seen to be an approximation to the binomial distribution with parameters (n, p) when n is very large and p is very small so that $\lambda = np$ is of moderate size (see Section 4.2.1). In Example 9.7 it was verified that

$$E(X) = \lambda \quad \text{and} \quad \text{var}(X) = \lambda.$$

Problem 9.22 Daily Airlines flies every day from Amsterdam to London. The price for a ticket on this popular route is \$75. The aircraft has a capacity of 150 passengers. Demand for tickets is greater than capacity, and tickets are sold out well in advance of flight departures. The airline company sells 160 tickets for each flight to protect itself against no-show passengers. The probability of a passenger being a no-show is $q = 0.1$. No-show passengers are refunded half the price of their tickets. Passengers that do show up and are not able to board the flight due to the overbooking are refunded the full amount of their tickets plus an extra \$425 compensation. What is the probability that more passengers will turn up for a flight than the aircraft has the seating capacity for? What are the expected value and standard deviation of the daily return for the airline?

Problem 9.23 What is the fewest number of dice one can roll such that, when they are rolled simultaneously, there will be at least a 50% probability of rolling two or more sixes?

Problem 9.24 On bridge night, the cards are dealt round seven times. Only two times do you receive an ace. From the beginning, you had your doubts as to whether the cards were being shuffled thoroughly. Are these doubts confirmed?

Problem 9.25 In the World Series Baseball, the final two teams play a series consisting of a possible seven games until such time that one of the two teams has won four games. In one such final, two unevenly matched teams are pitted against each other and the probability that the weaker team will win any given game is equal to 0.45. Assuming that the results of the various games are independent from each other, calculate the probability of the weaker team winning the final. What are the expected value and the standard deviation of the number of games the final will take?

Problem 9.26 A die is rolled until a six appears for the third time. What is the probability distribution of the number of rolls required?

Problem 9.27 In the famous problem of Chevalier de Méré, players bet first on the probability that a six will turn up at least one time in four rolls of a fair die; subsequently, players bet on the probability that a double six will turn up in 24 rolls of a pair of fair dice. In a generalized version of the de Méré problem, the dice are rolled a total of $4 \times 6^{r-1}$ times; each individual roll consists of r fair dice being rolled simultaneously. A king's roll results in all of the r dice rolled turning up sixes. Argue that the probability of at least one king's roll converges to $1 - e^{-2/3} = 0.4866$ if $r \to \infty$.

Problem 9.28 Ten identical pairs of shoes are jumbled together in one large box. Without looking, someone picks four shoes out of the box. What is the probability that, among the four shoes chosen, there will be both a left and a right shoe?

Problem 9.29 There is a concert and 2,500 tickets are to be raffled off. You have sent in 100 applications. The total number of applications is 125,000. What are your chances of getting a ticket? Can you explain why this probability is approximately equal to $1 - e^{-2}$?

Problem 9.30 For a final exam, your professor gives you a list of 15 items to study. He indicates that he will choose eight for the actual exam. You will be required to answer five of those. You decide to study 10 of the 15 items. What is the probability that you will pass for the exam?

Problem 9.31 A psychologist claims that he can determine from a person's handwriting whether the person is left-handed or not. You do not believe the psychologist and therefore present him with 50 handwriting samples, of which 25 were written by left-handed people and 25 were written by right-handed people. You ask the psychologist to say which 25 were written by left-handed people. Will you change your opinion of him if the psychologist correctly identifies 18 of the 25 left-handers?

Problem 9.32 In European roulette the ball lands on one of the numbers $0, 1, \ldots, 36$ in every spin of the wheel. A gambler offers at even odds the bet that the house number 0 will come up once in every 25 spins of the wheel. What is the gambler's expected profit per dollar bet?

Problem 9.33 An absent-minded professor has m matches in his right pocket and m matches in his left pocket. Each time he needs a match, he reaches for a match in his left pocket with probability p and in his right pocket with probability $1 - p$. When the professor first discovers that one of his pockets is empty, what is the probability that the other pocket has exactly k matches for $k = 0, 1, \ldots, m$? This problem is known as the Banach match problem.

Problem 9.34 In the Lotto 6/45 six different numbers are drawn at random from the numbers $1, 2, \ldots, 45$. What are the probability mass functions of the largest number drawn and the smallest number drawn?

10
Continuous random variables

In many practical applications of probability, physical situations are better described by random variables that can take on a *continuum* of possible values rather than a *discrete* number of values. Examples are the decay time of a radioactive particle, the time until the occurrence of the next earthquake in a certain region, the lifetime of a battery, the annual rainfall in London, and so on. These examples make clear what the fundamental difference is between discrete random variables and continuous random variables. Whereas a discrete random variable associates *positive* probabilities to its individual values, any individual value has probability *zero* for a continuous random variable. It is only meaningful to speak of the probability of a continuous random variable taking on a value in some interval. Taking the lifetime of a battery as an example, it will be intuitively clear that the probability of this lifetime taking on a specific value becomes zero when a finer and finer unit of time is used. If you can measure the heights of people with infinite precision, the height of a randomly chosen person is a continuous random variable. In reality, heights cannot be measured with infinite precision, but the mathematical analysis of the distribution of heights of people is greatly simplified when using a mathematical model in which the height of a randomly chosen person is modeled as a continuous random variable. Integral calculus is required to formulate the continuous analog of a probability mass function. An initial impetus to this was given in Section 5.1.1. The purpose of this chapter is to familiarize the reader with the concept of the probability density function of a continuous random variable. This is always a difficult concept for the beginning student. However, integral calculus enables us to give an enlightening interpretation of a probability density. Also, this chapter summarizes the most important probability densities used in practice. Finally, the inverse-transformation method for generating a random observation from a continuous random variable and the concept of failure rate function will be discussed.

10.1 Concept of probability density

The most simple example of a continuous random variable is the random choice of a number from the interval (0,1). The probability that the randomly chosen number will take on a prespecified value is zero. It only makes sense to speak of the probability of the randomly chosen number falling in a given subinterval of (0,1). This probability is equal to the length of that subinterval. For example, if a dart is thrown at random to the interval (0,1), the probability of the dart hitting exactly the point 0.25 is zero, but the probability of the dart landing somewhere in the interval between 0.2 and 0.3 is 0.1 (assuming that the dart has an infinitely thin point). No matter how small Δx is, any subinterval of the length Δx has probability Δx of containing the point at which the dart will land. You might say that the probability mass associated with the landing point of the dart is smeared out over the interval (0, 1) in such a way that the density is the same everywhere. For the random variable X denoting the point at which the dart will land, we have that the cumulative probability $P(X \leq a) = a$ for $0 \leq a \leq 1$ can be represented as $P(X \leq a) = \int_0^a f(x)dx$ with the density $f(x)$ identically equal to 1 on the interval (0, 1). In order to introduce the concept of probability density within a general framework, it is instructive to consider the following example.

Example 10.1 A stick of unit length is broken at random into two pieces. What is the probability that the ratio of the length of the shorter piece to that of the longer piece is smaller than or equal to a for any $0 < a < 1$?

Solution. The sample space of the chance experiment is the interval (0,1), where the outcome $\omega = u$ means that the point at which the stick is broken is a distance u from the beginning of the stick. Let the random variable X denote the ratio of length of the shorter piece to that of the longer piece of the broken stick. Denote by $F(a)$ the probability that the random variable X takes on a value smaller than or equal to a. Fix $0 < a < 1$. The probability that the ratio of the length of the shorter piece to that of the longer piece is smaller than or equal to a is nothing else than the probability that a random number from the interval (0,1) falls either in $(\frac{1}{1+a}, 1)$ or in $(0, 1 - \frac{1}{1+a})$. The latter probability is equal to $2(1 - \frac{1}{1+a}) = \frac{2a}{1+a}$. Thus,

$$F(a) = \frac{2a}{1+a} \quad \text{for } 0 < a < 1.$$

Obviously, $F(a) = 0$ for $a \leq 0$ and $F(a) = 1$ for $a \geq 1$. Denoting by $f(a) = \frac{2}{(1+a)^2}$ the derivative of $F(a)$ for $0 < a < 1$ and letting $f(a) = 0$ outside the interval (0,1), it follows that

$$F(a) = \int_{-\infty}^{a} f(x)dx \qquad \text{for all } a.$$

In this specific example, we have a continuous analog of the cumulative probability $F(a)$ in the discrete case: if X is a discrete random variable having possible values $a_1, a_2, \ldots$ with associated probabilities $p_1, p_2, \ldots$, then the probability that X takes on a value smaller than or equal to a is represented by

$$F(a) = \sum_{i:a_i \leq a} p_i \qquad \text{for all } a.$$

We now come to the definition of a continuous random variable. Let X be a random variable that is defined on a sample space with probability measure P. It is assumed that the set of possible values of X is uncountable and is a finite or infinite interval on the real line.

Definition 10.1 *The random variable X is said to be (absolutely) continuously distributed if a function $f(x)$ exists such that*

$$P(X \leq a) = \int_{-\infty}^{a} f(x)\,dx \qquad \textit{for each real number a,}$$

where the function $f(x)$ satisfies

$$f(x) \geq 0 \qquad \textit{for all } x \quad \textit{and} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

The notation $P(X \leq a)$ stands for the probability that is assigned by the probability measure P to the set of all outcomes ω for which $X(\omega) \leq a$. The function $P(X \leq x)$ is called the *(cumulative) probability distribution function* of the random variable X, and the function $f(x)$ is called the *probability density function* of X. Unlike the probability distribution function of a discrete random variable, the probability distribution function of a continuous random variable has no jumps and is continuous everywhere.

Beginning students often misinterpret the nonnegative number $f(a)$ as a probability, namely as the probability $P(X = a)$. This interpretation is wrong. Nevertheless, it is possible to give an intuitive interpretation of the nonnegative number $f(a)$ in terms of probabilities. Before doing this, we present an example of a continuous random variable with a probability density function.

Example 10.2 Suppose that the lifetime X of a battery has the cumulative probability distribution function

$$P(X \leq x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{1}{4}x^2 & \text{for } 0 \leq x \leq 2, \\ 1 & \text{for } x > 2. \end{cases}$$

The probability distribution function $P(X \leq x)$ is continuous and is differentiable at each point x except for the two points $x = 0$ and $x = 2$. Also, the derivative is integrable. We can now conclude from the fundamental theorem of integral calculus that the random variable X has a probability density function. This probability density function is obtained by differentiation of the probability distribution function and is given by

$$f(x) = \begin{cases} \frac{1}{2}x & \text{for } 0 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

In each of the finite number of points x at which $P(X \leq x)$ has no derivative, it does not matter what value we give $f(x)$. These values do not affect $\int_{-\infty}^{a} f(x)\,dx$. Usually, we give $f(x)$ the value 0 at any of these exceptional points.

10.1.1 Interpretation of the probability density

The use of the word "density" originated with the analogy to the distribution of matter in space. In physics, any finite volume, no matter how small, has a positive mass, but there is no mass at a single point. A similar description applies to continuous random variables. To make this more precise, we first express $P(a < X \leq b)$ in terms of the density $f(x)$ for any constants a and b with $a < b$. Noting that the event $\{X \leq b\}$ is the union of the two disjoint events $\{a < X \leq b\}$ and $\{X \leq a\}$, it follows that $P(X \leq b) = P(a < X \leq b) + P(X \leq a)$. Hence

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx \qquad \text{for } a < b \end{aligned}$$

and so

$$P(a < X \leq b) = \int_{a}^{b} f(x)\,dx \qquad \text{for } a < b.$$

In other words, the area under the graph of $f(x)$ between the points a and b gives the probability $P(a < X \leq b)$. Next, we find that

$$\begin{aligned} P(X = a) &= \lim_{n\to\infty} P\left(a - \frac{1}{n} < X \leq a\right) \\ &= \lim_{n\to\infty} \int_{a-\frac{1}{n}}^{a} f(x)\,dx = \int_{a}^{a} f(x)dx, \end{aligned}$$

using the continuity property of the probability measure P stating that $\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n)$ for any nonincreasing sequence of events A_n

(see Section 7.1.3). Hence, we arrive at the conclusion

$$P(X = a) = 0 \qquad \text{for each real number } a.$$

This formally proves that, for a continuous random variable X, it makes no sense to speak of the probability that the random variable X will take on a *prespecified* value. This probability is always zero. It only makes sense to speak of the probability that the continuous random variable X will take on a value in some interval. Incidentally, since $P(X = c) = 0$ for any number c, the probability that X takes on a value in an interval with endpoints a and b is not influenced by whether or not the endpoints are included. In other words, for any two real numbers a and b with $a < b$, we have

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

The fact that the area under the graph of $f(x)$ can be interpreted as a probability leads to an intuitive interpretation of $f(a)$. Let a be a given continuity point of $f(x)$. Consider now a small interval of length Δa around the point a, say $[a - \frac{1}{2}\Delta a, a + \frac{1}{2}\Delta a]$. Since

$$P\left(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a\right) = \int_{a-\frac{1}{2}\Delta a}^{a+\frac{1}{2}\Delta a} f(x)\,dx$$

and

$$\int_{a-\frac{1}{2}\Delta a}^{a+\frac{1}{2}\Delta a} f(x)\,dx \approx f(a)\Delta a \qquad \text{for } \Delta a \text{ small},$$

we obtain that

$$P\left(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a\right) \approx f(a)\Delta a \qquad \text{for } \Delta a \text{ small}.$$

In other words, the probability of random variable X taking on a value in a *small* interval around point a is approximately equal to $f(a)\Delta a$ when Δa is the length of the interval. You see that the number $f(a)$ itself is *not* a probability, but is a relative measure for the likelihood that random variable X will take on a value in the immediate neighborhood of point a. Stated differently, the probability density function $f(x)$ expresses how densely the probability mass of random variable X is smeared out in the neighborhood of point x. Hence, the name of density function. The probability density function provides the most useful description of a continuous random variable. The graph of the density function provides a good picture of the likelihood of the possible values of the random variable.

10.1.2 Verification of a probability density

In general, how can we verify whether a random variable X has a probability density? In concrete situations, we first determine the cumulative distribution function $F(a) = P(X \leq a)$ and next we verify whether $F(a)$ can be written in the form of $F(a) = \int_{-\infty}^{a} f(x)\,dx$. A sufficient condition is that $F(x)$ is continuous at every point x and is differentiable except for a finite number of points x. The following two examples are given in illustration of this point.

Example 10.3 Let the random variable be given by $X = -\frac{1}{\lambda}\ln(U)$, where U is a random number between 0 and 1 and λ is a given positive number. What is the probability density function of X?

Solution. To answer the question, note first that X is a positive random variable. For any $x > 0$

$$\begin{aligned} P(X \leq x) &= P\left(-\frac{1}{\lambda}\ln(U) \leq x\right) = P(\ln(U) \geq -\lambda x) \\ &= P(U \geq e^{-\lambda x}) = 1 - P(U \leq e^{-\lambda x}), \end{aligned}$$

where the last equality uses the fact that $P(U < u) = P(U \leq u)$ for the continuous random variable U. Since $P(U \leq u) = u$ for $0 < u < 1$, it follows that

$$P(X \leq x) = 1 - e^{-\lambda x}, \qquad x > 0.$$

Obviously, $P(X \leq x) = 0$ for $x \leq 0$. Noting that the expression for $P(X \leq x)$ is continuous at every point x and is differentiable except at $x = 0$, we obtain by differentiation that X has a probability density function $f(x)$ with $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$. This density function is the so-called exponential density function. In many situations, it describes adequately the density function of the waiting time until a *rare* event occurs.

Example 10.4 A point is picked at random in the inside of a circular disk with radius r. Let the random variable X denote the distance from the center of the disk to this point. Does the random variable X have a probability density function and, if so, what is its form?

Solution. To answer the question, we first define a sample space with an appropriate probability measure P for the chance experiment. The sample space is taken as the set of all points (x, y) in the two-dimensional plane with $x^2 + y^2 \leq r^2$. Since the point inside the circular disk is chosen at random, we assign to each well-defined subset A of the sample space the probability

$$P(A) = \frac{\text{area of region } A}{\pi r^2}.$$

The cumulative probability distribution function $P(X \leq x)$ is easily calculated. The event $X \leq a$ occurs if and only if the randomly picked point falls in the disk of radius a with area πa^2. Therefore

$$P(X \leq a) = \frac{\pi a^2}{\pi r^2} = \frac{a^2}{r^2} \qquad \text{for } 0 \leq a \leq r.$$

Obviously, $P(X \leq a) = 0$ for $a < 0$ and $P(X \leq a) = 1$ for $a > r$. Since the expression for $P(X \leq x)$ is continuous at every point x and is differentiable except at the points $x = 0$ and $x = a$, it follows that X has a probability density function which is given by

$$f(x) = \begin{cases} \frac{2x}{r^2} & \text{for } 0 < x < r, \\ 0 & \text{otherwise.} \end{cases}$$

All of the foregoing examples follow the same procedure in order to find the probability density function of a random variable X. The cumulative probability distribution function $P(X \leq x)$ is determined first and this distribution function is differentiated next.

Problem 10.1 Let X be a positive random variable with probability density function $f(x)$. Define the random variable Y by $Y = X^2$. What is the probability density function of Y? Also, find the density function of the random variable $W = V^2$ if V is a number chosen at random from the interval $(-a, a)$ with $a > 0$.

Problem 10.2 A point Q is chosen at random inside the unit square. What is the density function of the sum of the coordinates of the point Q? What is the density function of the product of the coordinates of the point Q? Use geometry to find these densities.

Problem 10.3 The number X is chosen at random between 0 and 1. Determine the probability density function of each of the random variables $V = X/(1 - X)$ and $W = X(1 - X)$.

Problem 10.4 A stick of unit length is broken at random into two pieces. Let the random variable X represent the length of the shorter piece. What is the probability density of X? Also, use the probability distribution function of X to give an alternative derivation of the probability density of the random variable $X/(1 - X)$ from Example 10.1.

Problem 10.5 The numbers U_1 and U_2 are chosen at random from the interval $(0, 1)$, independently of each other. Let the random variables V and W be defined by $V = \min(U_1, U_2)$ and $W = \max(U_1, U_2)$. What are the probability density functions of the random variables V and W?

Problem 10.6 Suppose you decide to take a ride on the ferris wheel at an amusement park. The ferris wheel has a diameter of 30 meters. After several turns, the ferris wheel suddenly stops due to a power outage. What random variable determines your height above the ground when the ferris wheel stops? What is the probability that this height is not more than 22.5 meters? And the probability of no more than 7.5 meters? What is the probability density function of the random variable governing the height above the ground?

10.1.3 Expected value

The expected value of a continuous random variable X with probability density function $f(x)$ is defined by

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

provided that the integral $\int_{-\infty}^{\infty} |x| f(x)\, dx$ is finite (the latter integral is always well defined by the nonnegativity of the integrand). It is then said that $E(X)$ exists. In the case that X is a nonnegative random variable, the integral $\int_0^{\infty} x f(x)\, dx$ is always well defined when allowing ∞ as a possible value. The definition of expected value in the continuous case parallels the definition $E(X) = \sum x_i p(x_i)$ for a discrete random variable X with $x_1, x_2, \ldots$ as possible values and $p(x_i) = P(X = x_i)$. For dx small, the quantity $f(x)\, dx$ in a discrete approximation of the continuous case corresponds with $p(x)$ in the discrete case. The summation becomes an integral when dx approaches zero. Results for discrete random variables are typically expressed as sums. The corresponding results for continuous random variables are expressed as integrals.

As an illustration, consider the random variable X from Example 10.3. The expected value of the distance X equals

$$E(X) = \int_0^r x \frac{2x}{r^2}\, dx = \frac{2}{3} \frac{x^3}{r^2}\Big|_0^r = \frac{2}{3} r.$$

Example 10.1 (continued) A stick of unit length is broken at random into two pieces. What is the expected value of the ratio of the length of the shorter piece to that of the longer piece? What is the expected value of the ratio of the length of the longer piece to that of the shorter piece?

Solution. Denote by the random variable X the ratio of the length of the shorter piece to that of the longer piece and by the random variable Y the ratio of the length of the longer piece to that of the shorter piece. In Example 10.1

we showed that X has the probability distribution function $F(x) = \frac{2x}{x+1}$ with probability density $f(x) = \frac{2}{(x+1)^2}$ for $0 < x < 1$. Hence

$$E(X) = \int_0^1 x \frac{2}{(x+1)^2}\,dx = 2\int_0^1 \frac{1}{x+1}\,dx - 2\int_0^1 \frac{1}{(x+1)^2}\,dx$$
$$= 2\ln(x+1)\Big|_0^1 + 2\frac{1}{x+1}\Big|_0^1 = 2\ln(2) - 1.$$

In order to calculate $E(Y)$, note that $Y = \frac{1}{X}$. Hence, $P(Y \le y) = P(X \ge \frac{1}{y})$ for $y > 1$. This leads to $P(Y \le y) = 1 - \frac{2}{y+1}$ for $y > 1$. Thus, the random variable Y has the probability density function $\frac{2}{(y+1)^2}$ for $y > 1$ and so

$$E(Y) = \int_1^\infty y \frac{2}{(y+1)^2}\,dy = 2\ln(y+1)\Big|_1^\infty + 2\frac{1}{y+1}\Big|_1^\infty = \infty.$$

This finding is in agreement with the result of Problem 9.5 in Section 9.2. A little calculus was enough to find a result that otherwise is difficult to obtain from a simulation study.

Problem 10.7 The javelin thrower Big John throws the javelin more than x meters with probability $P(x)$, where $P(x) = 1$ for $0 \le x < 50$, $P(x) = \frac{1{,}200-(x-50)^2}{1{,}200}$ for $50 \le x < 80$, $P(x) = \frac{(90-x)^2}{400}$ for $80 \le x < 90$, and $P(x) = 0$ for $x \ge 90$. What is the expected value of the distance thrown in his next shot?

Problem 10.8 In an Internet auction of a collector's item ten bids are done. The bids are independent of each other and are uniformly distributed on (0, 1). The person with the largest bid gets the item for the price of the second largest bid (a so-called Vickrey auction). Argue that the probability of the second largest bid exceeding the value x is equal to $\sum_{k=2}^{10} \binom{10}{k}(1-x)^k x^{10-k}$ for $0 < x < 1$ and use this result to obtain the expected value of this bid. *Hint*: $\int_0^1 x^a(1-x)^b\,dx = a!b!/(a+b+1)!$ for any integers $a, b \ge 0$.

Problem 10.9 A point is chosen at random inside the unit square $\{(x, y) : 0 \le x, y \le 1\}$. What is the expected value of the distance from this point to the point (0,0)?

Problem 10.10 A point is chosen at random inside the unit circle. Let the random variable V denote the absolute value of the x-coordinate of the point. What is the expected value of V?

Problem 10.11 A point is chosen at random inside a triangle with height h and base of length b. What is the expected value of the perpendicular distance of the point to the base?

Problem 10.12 Let X be a nonnegative continuous random variable with density function $f(x)$. Use an interchange of the order of integration to verify that $E(X) = \int_0^\infty P(X > u)\,du$.

10.1.4 Variance

The substitution rule and concept of variance of a random variable were discussed in the Sections 9.4 and 9.5 for the case of a discrete random variable. The same results apply to the case of a continuous random variable. Let X be a continuous random variable with density $f(x)$. For any given function $g(x)$, the expected value of the random variable $g(X)$ can be calculated from

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

provided that the integral exists. An illustrative example is as follows.

Example 10.1 (continued) A stick of unit length is broken at random into two pieces. The random variable V represents the ratio of the length of the shorter piece to that of the longer piece. In the previous section we calculated $E(V)$ by determining the density function of V and applying the definition of $E(V)$. However, the substitution rule provides a simpler way to calculate $E(V)$ by using the fact that $V = g(U)$ when U is a random number from the interval (0.1) and function $g(u)$ is defined by $g(u) = u/(1-u)$ for $0 < u \leq \frac{1}{2}$ and $g(u) = (1-u)/u$ for $\frac{1}{2} < u < 1$. This gives

$$\begin{aligned} E(V) &= \int_0^{1/2} \frac{u}{1-u} du + \int_{1/2}^1 \frac{1-u}{u} du = 2\int_{1/2}^1 \frac{1-u}{u} du \\ &= 2\ln(u) - 2u \Big|_{1/2}^{1} = 2\ln(2) - 1. \end{aligned}$$

Next we concentrate on the standard deviation. Letting $\mu = E(X)$, the variance of the random variable X, which is defined by $\text{var}(X) = E[(X-\mu)^2]$, can be calculated from

$$\text{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx.$$

Using the alternative representation $\text{var}(X) = E(X^2) - \mu^2$, the variance of X is usually calculated from

$$\text{var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2.$$

The variance of X does not have the same dimension as the values of X. Therefore, one often uses the standard deviation of the random variable X, which is defined by

$$\sigma(X) = \sqrt{\text{var}(X)}.$$

As an illustration, we calculate the variance of the random variable X from Example 10.4

$$\text{var}(X) = \int_0^r x^2 \frac{2x}{r^2}\,dx - \left(\frac{2}{3}r\right)^2 = \frac{2r^2}{4} - \frac{4}{9}r^2 = \frac{1}{18}r^2.$$

The standard deviation of the distance from the randomly selected point inside the circle to the origin is $\sigma(X) = \sqrt{\text{var}(X)} = 0.2357r$.

Example 10.5 Let the random variable X represent a number drawn at random from the interval (a, b). What are the expected value and the variance of X?

Solution. The probability that X will fall into a subinterval of width w is $\frac{w}{b-a}$. Hence, $P(X \leq x) = \frac{x-a}{b-a}$ for $a \leq x \leq b$ and so the density function $f(x)$ of X is given by $f(x) = \frac{1}{b-a}$ for $a < x < b$ and $f(x) = 0$ otherwise. This gives

$$E(X) = \int_a^b x \frac{1}{b-a}\,dx = \frac{1}{2}\frac{x^2}{b-a}\bigg|_a^b = \frac{1}{2}\frac{b^2-a^2}{b-a} = \frac{a+b}{2},$$

using the fact that $b^2 - a^2 = (b-a)(b+a)$. Similarly, we find

$$E(X^2) = \int_a^b bx^2 \frac{1}{b-a}\,dx = \frac{1}{3}\frac{x^3}{b-a}\bigg|_a^b = \frac{1}{3}\frac{b^3-a^3}{b-a} = \frac{a^2+ab+b^2}{3},$$

using the fact that $b^3 - a^3 = (b^2 + ab + a^2)(b-a)$. Thus

$$\text{var}(X) = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

Problem 10.13 A point Q is chosen at random inside a sphere with radius r. What are the expected value and the standard deviation of the distance from the center of the sphere to the point Q?

Problem 10.14 The lifetime (in months) of a battery is a random variable X satisfying $P(X \leq x) = 0$ for $x < 5$, $P(X \leq x) = [(x-5)^3 + 2(x-5)]/12$ for $5 \leq x < 7$ and $P(X \leq x) = 1$ for $x \geq 7$. What are the expected value and the standard deviation of X?

Problem 10.15 Let X be a continuous random variable with probability density $f(x)$ and finite expected value $E(X)$.

(a) What constant c minimizes $E[(X-c)^2]$ and what is the minimal value of $E[(X-c)^2]$?

(b) Prove that $E(|X-c|)$ is minimal if c is chosen equal to the median of X, where the *median* of X is any value m for which $P(X \leq m) = P(X \geq m) = \frac{1}{2}$.[†]

Problem 10.16 Consider Problem 10.6 again. Calculate the expected value and standard deviation of the height above the ground when the ferris wheel stops.

Problem 10.17 In an inventory system, a replenishment order is placed when the stock on hand of a certain product drops to the level s, where the reorder point s is a given positive number. The total demand for the product during the lead time of the replenishment order has the probability density $f(x) = \lambda e^{-\lambda x}$ for $x > 0$. What are the expected value and standard deviation of the shortage (if any) when the replenishment order arrives?

Problem 10.18 Suppose that the continuous random variable X has the probability density function $f(x) = (\alpha/\beta)(\beta/x)^{\alpha+1}$ for $x > \beta$ and $f(x) = 0$ for $x \leq \beta$ for given values of the parameters $\alpha > 0$ and $\beta > 0$. This density is called the *Pareto* density, which provides a useful probability model for income distributions among others.

(a) Calculate the expected value, the variance and the median of X.

(b) Assume that the annual income of employed measured in thousands of dollars in a given country follows a Pareto distribution with $\alpha = 2.25$ and $\beta = 2.5$. What percentage of the working population has an annual income of between 25 and 40 thousand dollars?

(c) Why do you think the Pareto distribution is a good model for income distributions? *Hint*: use the probabilistic interpretation of the density function $f(x)$.

Problem 10.19 A stick of unit length is broken at random into two pieces. Let the random variable X represent the length of the shorter piece. What is the median of the random variable $(1-X)/X$?

Problem 10.20 Let the random variables V and W be defined by $V = \sqrt{U}$ and $W = U^2$ when U is a number chosen at random between 0 and 1. What are the expected values and the standard deviations of V and W?

[†] The median is sometimes a better measure for a random variable than the expected value. For example, this is the case for income distributions.

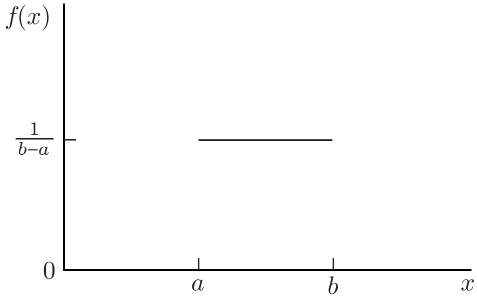


Fig. 10.1. Uniform density.

10.2 Important probability densities

Any nonnegative function $f(x)$ whose integral over the interval $(-\infty, \infty)$ equals 1 can be regarded as a probability density function of a random variable. In real-world applications, however, special mathematical forms naturally show up. In this section, we introduce several families of continuous random variables that frequently appear in practical applications. The probability densities of the members of each family all have the same mathematical form but differ only in one or more parameters. Uses of the densities in practical applications are indicated. Also, the expected values and the variances of the densities are listed without proof. A convenient method to obtain the expected values and the variances of special probability densities is the moment-generating function method to be discussed in Chapter 14.

10.2.1 Uniform density

A continuous random variable X is said to have a *uniform* density over the interval (a, b) if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

This density has two parameters a and b with $b > a$. Figure 10.1 gives the graph of the uniform density function. The uniform distribution provides a probability model for selecting a point at random from the interval (a, b). It is also used as a model for a quantity that is known to vary randomly between a and b but about which little else is known. Since $f(x) = 0$ outside the interval (a, b), the random variable X must assume a value in (a, b). Also, since $f(x)$ is constant over the interval (a, b), the random variable X is just as likely to be near any value in (a, b) as any other value. This property is also expressed by

$$P\left(c - \frac{1}{2}\Delta \leq X \leq c + \frac{1}{2}\Delta\right) = \int_{c-\frac{1}{2}\Delta}^{c+\frac{1}{2}\Delta} \frac{1}{b-a}\,dx = \frac{\Delta}{b-a},$$

regardless of c provided that the points $c - \frac{1}{2}\Delta$ and $c + \frac{1}{2}\Delta$ belong to the interval (a, b). The expected value and the variance of the random variable X are given by

$$E(X) = \frac{1}{2}(a+b) \quad \text{and} \quad \text{var}(X) = \frac{1}{12}(b-a)^2.$$

Also, an explicit expression can be given for the cumulative probability distribution function $F(x) = \int_{-\infty}^{x} f(y)\,dy$. This function satisfies $F(x) = 0$ for $x < a$, $F(x) = 1$ for $x \geq b$, and

$$F(x) = \frac{b-x}{b-a} \qquad \text{for } a \leq x < b.$$

10.2.2 Triangular density

A continuous random variable X is said to have a *triangular* density over the interval (a, b) if its probability density function is given by

$$f(x) = \begin{cases} h\frac{x-a}{m-a} & \text{for } a < x \leq m \\ h\frac{b-x}{b-m} & \text{for } m \leq x < b \\ 0 & \text{otherwise.} \end{cases}$$

This density has three parameters a, b, and m with $a < m < b$. The constant $h > 0$ is determined by $\int_a^b f(x)dx = 1$, and so

$$h = \frac{2}{b-a}.$$

Figure 10.2 gives the graph of the triangular density function. The density function increases linearly on the interval $[a, m]$ and decreases linearly on the interval $[m, b]$. The triangular distribution is often used as the probability model when little information is available about the quantity of interest but one knows its lowest possible value a, its most likely value m, and its highest possible value b. The expected value and the variance of the random variable X are given by

$$E(X) = \frac{1}{3}(a+b+m), \ \text{var}(X) = \frac{1}{18}(a^2+b^2+m^2-ab-am-bm).$$

Also, an explicit expression can be given for the cumulative probability distribution function $F(x) = \int_{-\infty}^{x} f(y)\,dy$. This function satisfies $F(x) = 0$ for

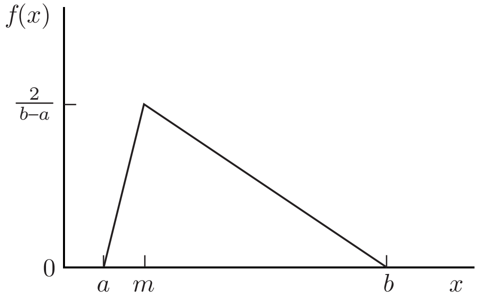


Fig. 10.2. Triangular density.

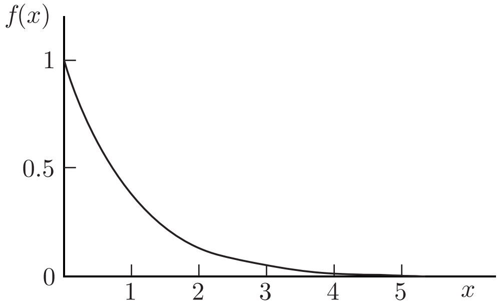


Fig. 10.3. Exponential density ($\lambda = 1$).

$x < a$, $F(x) = 1$ for $x \geq b$, and

$$F(x) = \begin{cases} \frac{(x-a)^2}{(b-a)(m-a)} & \text{for } a \leq x < m \\ 1 - \frac{(b-x)^2}{(b-a)(b-m)} & \text{for } m \leq x < b. \end{cases}$$

10.2.3 Exponential density

The continuous random variable X is said to have an *exponential* density with parameter $\lambda > 0$ if its probability density function is of the form

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The parameter λ is a scale parameter. An exponentially distributed random variable X takes on only positive values. Figure 10.3 displays the exponential density function with $\lambda = 1$. The exponential distribution is often used as a probability model for the time until a *rare* event occurs. Examples are the time elapsed until the next earthquake in a certain region and the decay time of a radioactive particle. Also, the exponential distribution is frequently used to model times between independent events such as arrivals at a service facility. The exponential distribution is intimately related to the Poisson arrival process that was discussed in Section 4.2.4. The expected value and the variance of the random variable X are given by

$$E(X) = \frac{1}{\lambda} \quad \text{and} \quad \text{var}(X) = \frac{1}{\lambda^2}.$$

The cumulative probability distribution function $F(x) = \int_{-\infty}^{x} f(y)\,dy$ equals

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0. \end{cases}$$

The exponential distribution is very important in probability. It not only models many real-world phenomena, but it allows for tractable mathematical analysis as well. The reason for its mathematical tractability is the *memoryless property* of the exponential distribution. The memoryless property states that, regardless of the value of t_0,

$$P(X > t_0 + x \mid X > t_0) = P(X > x) \qquad \text{for all } x > 0,$$

regardless of the value of t_0. In words, imagining that the exponentially distributed random variable X represents the lifetime of an item, the residual life of an item has the *same* exponential distribution as the original lifetime, regardless of how long the item has been already in use (see Section 10.4 for a proof).

In many situations the probability of exceeding some *extreme* level is approximately equal to an exponential tail probability, where an exponential tail probability is a probability of the form $\alpha e^{-\beta t}$ for constants α, $\beta > 0$. For example, in queueing systems, the probability of a customer waiting more than a time t is often approximately equal to an exponential tail probability when t is large. Another interesting example concerns the probability that a high tide of h meters or more above sea level will occur in any given year somewhere along the Dutch coastline. This probability is approximately equal to $e^{-2.97h}$ for values of h larger than 1.70 m. This empirical result was used in the design of the Delta works that were built following the 1953 disaster when the sea flooded a number of polders in the Netherlands.

10.2.4 Gamma density

A continuous random variable X is said to have a *gamma* density with parameters $\alpha > 0$ and $\lambda > 0$ if its probability density function is given by

$$f(x) = \begin{cases} c\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

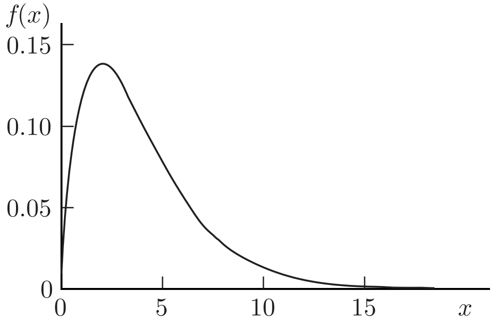


Fig. 10.4. Gamma density ($\alpha = 2.5$, $\lambda = 0.5$).

The constant c is determined by $\int_0^\infty f(x)dx = 1$. To specify c, we note that in advanced calculus the so-called *gamma function* is defined by

$$\Gamma(a) = \int_0^\infty e^{-y} y^{a-1} dy \qquad \text{for } a > 0.$$

This famous function has the property that

$$\Gamma(a+1) = a\Gamma(a) \qquad \text{for } a > 0.$$

This result is easily verified by partial integration. In particular

$$\Gamma(a) = (a-1)! \qquad \text{if } a \text{ is a positive integer.}$$

An easy consequence of the definition of $\Gamma(a)$ is that the constant c in the gamma density is given by

$$c = 1/\Gamma(\alpha).$$

The parameter α is a shape parameter, and the parameter λ is a scale parameter. A gamma-distributed random variable takes on only positive values. The gamma density with $\alpha = 1$ reduces to the exponential density. Figure 10.4 displays the gamma density with $\alpha = 2.5$ and $\lambda = 0.5$. The graph in Figure 10.4 is representative of the shape of the gamma density if the shape parameter α is larger than 1; otherwise, the shape of the gamma density is similar to that of the exponential density in Figure 10.3. The gamma distribution is a useful model in inventory and queueing applications to model demand sizes and service times. The expected value and the variance of the random variable X are given by

$$E(X) = \frac{\alpha}{\lambda} \quad \text{and} \quad \text{var}(X) = \frac{\alpha}{\lambda^2}.$$

Problem 10.21 Use properties of the gamma function to derive $E(X)$ and $E(X^2)$ for a gamma-distributed random variable X.

10.2.5 Weibull density

A continuous random variable X is said to have a *Weibull* density with parameters $\alpha > 0$ and $\lambda > 0$ if it has a probability density function of the form

$$f(x) = \begin{cases} \alpha\lambda(\lambda x)^{\alpha-1}e^{-(\lambda x)^{\alpha}} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The parameter α is a shape parameter, and the parameter λ is a scale parameter. The Weibull density has a similar shape as the gamma density. The expected value and the variance of the random variable X are given by

$$E(X) = \frac{1}{\lambda}\Gamma\left(1 + \frac{1}{\alpha}\right), \quad \text{var}(X) = \frac{1}{\lambda^2}\left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \left(\Gamma\left(1 + \frac{1}{\alpha}\right)\right)^2\right].$$

The Weibull distribution is a useful probability model for fatigue strengths of materials and is used in reliability models for lifetimes of devices.

10.2.6 Beta density

A continuous random variable X is said to have a *beta* density with parameters $\alpha > 0$ and $\beta > 0$ if its probability density function is of the form

$$f(x) = \begin{cases} cx^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

for an appropriate constant c. The constant c is determined by $\int_0^1 f(x)\,dx = 1$. Using advanced calculus, it can be shown that

$$c = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

Both parameters α and β are shape parameters. The beta distribution is a flexible distribution, and the graph of the beta density function can assume widely different shapes depending on the values of α and β. An extreme case is the uniform distribution on (0,1) corresponding to $\alpha = \beta = 1$. The graphs of several beta densities are given in Figure 10.5. The expected value and the variance of the random variable X are given by

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \text{var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

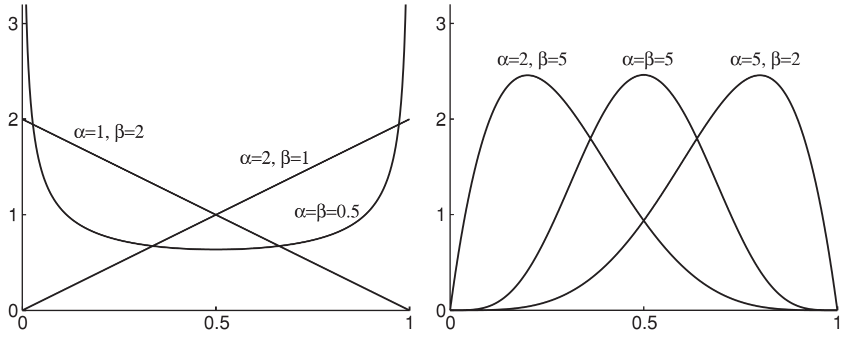


Fig. 10.5. Several beta densities.

The beta density is often used to model the distribution of a random proportion. It is common practice in Bayesian statistics to use a beta distribution for the prior distribution of the unknown value of the success probability in a Bernoulli experiment.

Problem 10.22 You perform an experiment that consists of ten independent Bernoulli trials. Before the experiment is done, your prior density of the success probability of the Bernoulli trials is a beta density with parameters α and β. Argue that the beta density with parameters $\alpha^* = \alpha + 7$ and $\beta^* = \beta^* + 3$ gives the posterior density of the success probability after you have done the experiment and observed seven successes. *Hint*: for Δp small, evaluate the conditional probability of having a success probability between p and $p + \Delta p$ given that seven successes occurred during the ten Bernoulli trials.

10.2.7 Normal density

This density was discussed extensively in Section 5.1. For completeness, we repeat the definition. A continuous random variable X is said to have a *normal* density with parameters μ and $\sigma > 0$ if its probability density function is given by

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

The parameter σ is a shape parameter, and the parameter μ is a scale parameter. The normal distribution also is referred to frequently as the *Gaussian* distribution. The expected value and the variance of the random variable X are given by

$$E(X) = \mu \quad \text{and} \quad \text{var}(X) = \sigma^2.$$

The notation X is $N(\mu, \sigma^2)$ is often used as a shorthand for X is a normally distributed random variable with parameters μ and σ. If $\mu = 0$ and $\sigma = 1$, the random variable X is said to have the *standard normal* distribution. The standard normal density function is $(1/\sqrt{2\pi})e^{-\frac{1}{2}x^2}$. Let

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2}\, dy$$

denote the standard normal distribution function. Then the cumulative probability distribution function of an $N(\mu, \sigma^2)$-distributed random variable X can be calculated as

$$P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

using the fact that $(X - \mu)/\sigma$ has the standard normal distribution.

Figure 5.2 in Chapter 5 displays the famous bell-shaped graph of the normal density function. Advanced calculus is required to prove that the area under the graph of the normal density function is indeed 1 (see Problem 10.24). The importance and applications of the normal density were discussed in Chapter 5. Although a normal random variable theoretically takes on values in the interval $(-\infty, \infty)$, it still may provide a useful model for a variable that takes on only positive values provided that the normal probability mass on the negative axis is negligible.

A nice property of an $N(\mu, \sigma^2)$-distributed random variable X is that $aX + b$ is $N(a\mu + b, a^2\sigma^2)$ distributed for any constants a, b with $a \neq 0$. This result can be directly verified by writing down the probability distribution function of $Y = aX + b$ and taking the derivative. Another useful property of the normal distribution is that $X + Y$ is $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ distributed if the random variables X and Y are independent and are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributed. This property will be proved in Section 14.2 of Chapter 14. Also, in Chapter 14 a proof of the central limit theorem will be outlined. This theorem states that the sum $X_1 + \cdots + X_n$ of n independent random variables $X_1, \ldots, X_n$ each having the same probability distribution with mean μ_1 and standard deviation σ_1 has approximately a normal distribution with mean $n\mu_1$ and standard deviation $\sigma_1\sqrt{n}$ for n large enough. Several applications of this very important theorem were discussed in Chapter 5.

Problem 10.23 Suppose that $X_1, \ldots, X_n$ are independent random variables that are uniformly distributed on (0,1). What is the probability that the rounded sum $X_1 + \cdots + X_n$ equals the sum of the rounded X_i when all rounding is to the nearest integer? Use the central limit theorem to verify that this probability is approximately equal to $1 - 2\Phi(\sqrt{3/n})$ for n sufficiently large.

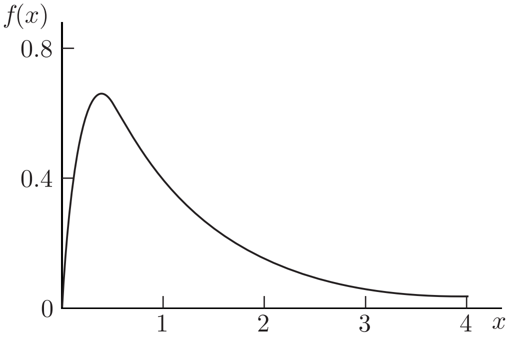


Fig. 10.6. Lognormal density ($\mu = 0, \sigma = 1$).

Problem 10.24 In order to prove that the normal probability density function integrates to 1 over the interval $(-\infty, \infty)$, evaluate the integral $I = \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx$ for the standard normal density. By changing to polar coordinates in the double integral $I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)}\,dx\,dy$, verify that $I = \sqrt{2\pi}$ (the polar coordinates r and θ satisfy $x = r\cos(\theta)$ and $y = r\sin(\theta)$ with $dx\,dy = r\,dr\,d\theta$). Also, verify that the change of variable $t = \frac{1}{2}x^2$ in $I = \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\,dx$ leads to $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

10.2.8 Lognormal density

A continuous random variable X is said to have a *lognormal* density with parameters μ and $\sigma > 0$ if its probability density function is given by

$$f(x) = \begin{cases} \frac{1}{\sigma x\sqrt{2\pi}}\, e^{-\frac{1}{2}[\ln(x)-\mu]^2/\sigma^2} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

A lognormally distributed random variable takes on only positive values. The graph of the lognormal density function with $\mu = 0$ and $\sigma = 1$ is displayed in Figure 10.6. It is not difficult to prove that the random variable X is lognormally distributed with parameters μ and σ if the random variable $\ln(X)$ is $N(\mu, \sigma^2)$ distributed (see also Example 10.6 in Section 10.3). Hence, using the relation $P(X \leq x) = P(\ln(X) \leq \ln(x))$ for $x > 0$

$$P(X \leq x) = \Phi\left(\frac{\ln(x) - \mu}{\sigma}\right) \qquad \text{for } x > 0.$$

The expected value and the variance of the random variable X are given by

$$E(X) = e^{\mu + \frac{1}{2}\sigma^2} \text{ and } \operatorname{var}(X) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right).$$

The lognormal distribution provides a useful probability model for income distributions. The explanation is that its probability density function $f(x)$ is skewed to the left and tends very slowly to zero as x approaches infinity (assuming that $\sigma > 1$). In other words, most outcomes of this lognormal distribution will be relatively small, but very large outcomes occur occasionally. Also, handling times of service requests at a call center and sizes of insurance claims often closely follow a lognormal distribution. The lognormal distribution is also often used to model future stock prices after a longer period of time. In general, the lognormal distribution arises when the underlying random variable is the result of a large number of independent multiplicative effects.

Problem 10.25 A population of bacteria has the initial size s_0. In each generation, independently of each other, it is equally likely that the population increases by 25% or decreases by 20%. What is the approximate probability density of the size of the population after n generations with n large?

10.2.9 Chi-square density

A continuous random variable X is said to have a *chi-square* distribution with *d degrees of freedom* if it can be represented as

$$X = Z_1^2 + Z_2^2 + \cdots + Z_d^2,$$

where $Z_1, Z_2, \ldots, Z_d$ are independent random variables, each having a standard normal distribution. The probability density function of X is

$$f(x) = \frac{1}{2^{\frac{1}{2}d}\Gamma\left(\frac{1}{2}d\right)} x^{\frac{1}{2}d-1} e^{-\frac{1}{2}x} \qquad \text{for } x > 0$$

(see Rule 14.5 in Section 14.2 for a proof). This density is a special case of the gamma density with shape parameter $\alpha = \frac{1}{2}d$ and scale parameter $\lambda = \frac{1}{2}$. Thus, the graph of the gamma density with $\alpha = 2.5$ and $\lambda = \frac{1}{2}$ in Figure 10.4 is also the graph of the chi-square density with $n = 5$. The expected value and the variance of the random variable X are given by

$$E(X) = d \qquad \text{and} \qquad \text{var}(X) = 2d.$$

The chi-square distribution plays an important role in statistics and is best known for its use in the so-called chi-square tests. Also, the chi-square distribution arises in the analysis of random walks: if $V_1, \ldots, V_d$ are independent random variables that are $N(0, \sigma^2)$ distributed, then the random variable

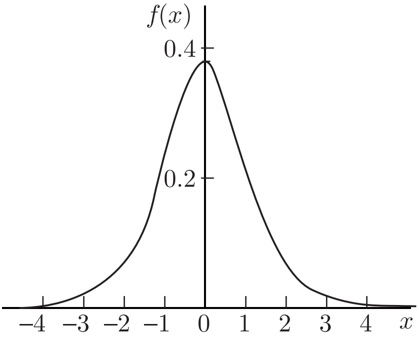


Fig. 10.7. Student-*t* density for $n = 5$.

$W = \sqrt{V_1^2 + \cdots + V_d^2}$ has the density function

$$f_W(w) = \frac{\sigma^{-d}}{2^{\frac{1}{2}d-1}\Gamma\left(\frac{1}{2}d\right)} w^{d-1} e^{-\frac{1}{2}w^2/\sigma^2} \qquad \text{for } w > 0.$$

The verification of this result is left as an exercise to the reader.

10.2.10 Student-*t* density

A continuous random variable X is said to have a *Student-t* distribution with n degrees of freedom if it can be represented as

$$X = \frac{Z}{\sqrt{U/n}},$$

where Z has a standard normal distribution, U has a chi-square distribution with n degrees of freedom and the random variables Z and U are independent. It can be shown that the density function of X is given by

$$f(x) = c\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} \qquad \text{for} - \infty < x < \infty.$$

The constant c is determined by $\int_{-\infty}^{\infty} f(x)dx = 1$. Using advanced calculus, it can be verified that $c = \frac{1}{\sqrt{\pi n}}\Gamma(\frac{1}{2}(n+1))/\,\Gamma(\frac{1}{2}n)$. In Figure 10.7, the Student-*t* density function is displayed for $n = 5$. The density function is very similar to that of the standard normal density but it has a longer tail than the $N(0, 1)$ density. The Student-*t* distribution is named after William Gosset, who invented this distribution in 1908 and used the pen name "A. Student" in his publication. Gosset worked for the Guiness brewery in Dublin which, at that time, did not allow its employees to publish research papers. The expected value and the variance of the random variable X are given by

$$E(X) = 0 \qquad \text{and} \qquad \operatorname{var}(X) = \frac{n}{n-2} \quad \text{for } n > 2.$$

The Student-t distribution is used in statistics, primarily when dealing with small samples from a *normal* population. In particular, this distribution is used for constructing an *exact* confidence interval in case the observations are generated from a normal distribution (confidence intervals were discussed in Section 5.7). This goes as follows. Suppose that $Y_1, \ldots, Y_n$ are independent samples from an $N(\mu, \sigma^2)$ distribution with (unknown) expected value μ. The construction of the confidence interval uses the sample mean $\overline{Y}(n)$ and the sample variance $\overline{S}^2(n)$ which are defined by

$$\overline{Y}(n) = \frac{1}{n}\sum_{k=1}^{n} Y_k \quad \text{and} \quad \overline{S}^2(n) = \frac{1}{n-1}\sum_{k=1}^{n}\left[Y_k - \overline{Y}(n)\right]^2.$$

It is stated without proof that the random variables $\overline{Y}(n)$ and $\overline{S}^2(n)$ are independent. Moreover, it can be shown that $(\overline{Y}(n) - \mu)/(\sigma/\sqrt{n})$ has a standard normal distribution and $(n-1)\overline{S}^2(n)/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom. Thus, the ratio

$$\frac{\overline{Y}(n) - \mu}{\sqrt{\overline{S}^2(n)/n}}$$

has a Student-t distribution with $n-1$ degrees of freedom. This important result holds for any value of n and enables us to give the following *exact* $100(1-\alpha)\%$ confidence interval for the unknown expected value μ

$$\overline{Y}(n) \pm t_{n-1,1-\frac{1}{2}\alpha}\sqrt{\frac{\overline{S}^2(n)}{n}},$$

where $t_{n-1,1-\frac{1}{2}\alpha}$ is the $(1-\frac{1}{2}\alpha)$th percentile of the Student-t density function with $n-1$ degrees of freedom. That is, the area under the graph of this symmetric density function between the points $-t_{n-1,1-\frac{1}{2}\alpha}$ and $t_{n-1,1-\frac{1}{2}\alpha}$ equals $1-\alpha$. This confidence interval for a sample from a normal population does not require a large n but can be used for any value of n. The statistic $(\overline{Y}(n) - \mu)/\sqrt{\overline{S}^2(n)/n}$ has the pleasant feature of being *robust*. This means that the statistic is not sensitive for small deviations from the normality assumption.

10.3 Transformation of random variables

In Chapter 2, we saw several methods for simulating random variates from a discrete distribution. Each of these methods used the tool of generating random numbers between 0 and 1. This tool is also indispensable for simulating random variates from a continuous distribution. This will be shown by an example. Let R be a continuous random variable with probability density function $h(r) = r\exp(-\frac{1}{2}r^2)$ for $r > 0$ and $h(r) = 0$ otherwise. This is the Rayleigh density with parameter 1, a much used density in physics. How to generate a random observation of R? To do so, we need the probability distribution function of the positive random variable R. Letting $H(r) = P(R \leq r)$, we have

$$H(r) = \int_0^r xe^{-\frac{1}{2}x^2}\,dx = -\int_0^r de^{-\frac{1}{2}x^2} = -e^{-\frac{1}{2}x^2}\Big|_0^r = 1 - e^{-\frac{1}{2}r^2}.$$

If u is a random number between 0 and 1, then the solution r to the equation

$$H(r) = u$$

is a random observation of R provided that this equation has a unique solution. Since $H(r)$ is strictly increasing on $(0, \infty)$, the equation has a unique solution. Also, the equation can be solved explicitly. The reader may easily verify that the equation $H(r) = u$ has the solution

$$r = \sqrt{-2\ln(1-u)}.$$

It will be clear that the above approach is generally applicable to simulate a random observation of a continuous random variable provided that the probability distribution function of the random variable is strictly increasing and allows for an easily computable inverse function. This approach is known as the *inverse-transformation* method.

As a by-product of the discussion above, we find that the transformation $\sqrt{-2\ln(1-U)}$ applied to the uniform random variable U on $(0, 1)$ yields a random variable with probability density function $r\exp(-\frac{1}{2}r^2)$ on $(0, \infty)$. This result can be put in a more general framework. Suppose that X is a continuous random variable with probability density function $f(x)$. What is the probability density function of the random variable $Y = v(X)$ for a given function $v(x)$? A simple formula can be given for the density function of $v(X)$ when the function $v(x)$ is either strictly increasing or strictly decreasing on the range of X. The function $v(x)$ then has a unique inverse function $a(y)$ (say). That is, for each attainable value y of $Y = v(X)$, the equation $v(x) = y$ has a unique solution $x = a(y)$. Note that $a(y)$ is strictly increasing (decreasing) if $v(x)$ is strictly increasing (decreasing). It is assumed that $a(y)$ is continuously differentiable.

Rule 10.1 *If the function $v(x)$ is strictly increasing or strictly decreasing, then the probability density of the random variable $Y = v(X)$ is given by*

$$f(a(y))|a'(y)|,$$

where $a(y)$ is the inverse function of $v(x)$.

The proof is simple and instructive. We first give the proof for the case that $v(x)$ is strictly increasing. Then, $v(x) \leq v$ if and only if $x \leq a(v)$. Thus

$$P(Y \leq y) = P(v(X) \leq y) = P(X \leq a(y)) = F(a(y)),$$

where $F(x)$ denotes the cumulative probability distribution function of X. Differentiating $P(Y \leq y)$ leads to

$$\frac{d}{dy} P(Y \leq y) = \frac{d}{da(y)} F\,(a(y))\,\frac{da(y)}{dy} = f\,(a(y))\,a'(y),$$

which gives the desired result, since $a'(y) > 0$ for a strictly increasing function $a(y)$. In the case of a strictly decreasing function $v(x)$, we have $v(x) \leq v$ if and only if $x \geq a(v)$ and so

$$P(Y \leq y) = P(v(X) \leq y) = P(X \geq a(y)) = 1 - F(a(y)).$$

By $a'(y) < 0$, differentiation of $P(Y \leq y)$ yields the desired result.

Example 10.6 Let the random variable Y be defined by $Y = e^X$, where X is an $N(\mu, \sigma^2)$-distributed random variable. What is the probability density of Y?

Solution. The inverse of the function $v(x) = e^x$ is given by $a(y) = \ln(y)$. The derivative of $a(y)$ is $1/y$. Applying Rule 10.1 gives that the probability density of Y is given by

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\ln(y)-\mu)^2/\sigma^2} \frac{1}{y} \qquad \text{for } y > 0.$$

In other words, the random variable Y has a lognormal density with parameters μ and σ.

In general, one best uses first principles to determine the probability density function of any given function of a continuous random variable X. This is illustrated by the following example.

Example 10.7 Suppose that the random variable X is $N(0, \sigma^2)$ distributed. What is the probability density function of the random variable $V = |X|$? What is the expected value of V?

Solution. Using the fact that X/σ is $N(0, 1)$ distributed, we have

$$\begin{aligned} P(V \leq v) &= P(-v \leq X \leq v) = P\left(\frac{-v}{\sigma} \leq \frac{X}{\sigma} \leq \frac{v}{\sigma}\right) \\ &= \Phi\left(\frac{v}{\sigma}\right) - \Phi\left(-\frac{v}{\sigma}\right) \qquad \text{for } v > 0. \end{aligned}$$

Differentiation gives that V has the probability density function

$$\frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}v^2/\sigma^2} \qquad \text{for } v > 0.$$

The expected value of V is calculated as

$$\begin{aligned} E(V) &= \int_0^\infty v \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}v^2/\sigma^2}\, dv = \frac{-2\sigma^2}{\sigma\sqrt{2\pi}} \int_0^\infty de^{-\frac{1}{2}v^2/\sigma^2} \\ &= \frac{-2\sigma}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2/\sigma^2}\Big|_0^\infty = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}. \end{aligned}$$

Problem 10.26 Verify that a random observation from the Weibull distribution with shape parameter α and scale parameter λ can be simulated by taking $X = \frac{1}{\lambda}[-\ln(1-U)]^{1/\alpha}$, where U is a random number from the interval (0,1). In particular, $X = -\frac{1}{\lambda}\ln(1-U)$ is a random observation from the exponential distribution with parameter λ.

Problem 10.27 Let X be a continuous random variable with probability density function $f(x)$. Suppose that the probability distribution function $F(x) = P(X \leq x)$ is strictly increasing on the range of X. Define the function $I(u)$ as the inverse function of $F(x)$. Verify that

(a) $P(I(U) \leq x) = P(X \leq x)$ for all x, where the continuous random variable U is uniformly distributed on (0, 1).

(b) For any function $g(x)$, $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx = \int_0^1 h(u)\,du$, where the function $h(u)$ is defined by $h(u) = g(I(u))$ for $0 < u < 1$.

10.4 Failure rate function

The concept of failure rate function applies to a positive random variable and can be best explained by considering a random variable X that represents the lifetime or the time to failure of an item. It is assumed that X has a probability distribution function $F(x) = P(X \leq x)$ with probability density $f(x)$. What is the probability that an item of age a will fail in the next Δa time units with Δa

small? This probability is given by

$$P(X \leq a + \Delta a \mid X > a) = \frac{P(a < X \leq a + \Delta a)}{P(X > a)}$$
$$\approx \frac{f(a)\Delta a}{1 - F(a)} \qquad \text{for } \Delta a \text{ small.}$$

Therefore, the *failure rate function* of the random variable X is defined as

$$r(x) = \frac{f(x)}{1 - F(x)} \qquad \text{for } x \geq 0.$$

The term hazard rate function is often used instead of failure rate function. The function $r(x)$ is not a probability density, but $r(x)$ represents the conditional probability intensity that an item of age x will fail in the next moment.

Noting that $r(x)$ is the derivative of $-\ln(1 - F(x))$, it follows that the failure rate function is related to the probability distribution function by

$$F(x) = 1 - e^{-\int_0^x r(t)dt} \qquad \text{for } x \geq 0.$$

As an example, consider an exponentially distributed lifetime X with expected value $1/\mu$. Then, $F(x) = 1 - e^{-\mu x}$ and $f(x) = \mu e^{-\mu x}$ and so $r(x) = \mu$ for all $x \geq 0$. Thus, the exponential distribution has a constant failure rate. In other words, new is as good as used when an item has an exponentially distributed lifetime. This characteristic is fairly accurate for many kinds of electronic devices. More generally, if X has a Weibull distribution with parameters α and λ, the failure rate function $r(x)$ follows as

$$r(x) = \alpha\lambda(\lambda x)^{\alpha-1},$$

using the formulas for $F(x)$ and $f(x)$ in Section 10.2.5. The Weibull distribution has an increasing failure rate if $\alpha > 1$ and a decreasing failure rate if $0 < \alpha < 1$ (the Weibull distribution with $\alpha = 1$ reduces to the exponential distribution). Most complex systems usually exhibit a failure rate that initially decreases to become nearly constant for a while, and then finally increases. This form of failure rate is known as the U-shaped failure rate or bathtub failure rate. An item with a bathtub failure rate has a fairly high failure rate when it is first put into operation. If the item survives the first period, then a nearly constant failure rate applies for some period. Finally, the failure rate begins to increase as wearout becomes a factor. More complicated probability distributions are required to model the bathtub-shaped failure rate function. The existence of such probability distribution functions is guaranteed by the following rule.

Rule 10.2 *Any function $r(x)$ with $r(x) \geq 0$ for all $x \geq 0$ and $\int_0^\infty r(t)dt = \infty$ is the failure rate function of a unique probability distribution function.*

The proof is simple. Define $F(x)$ by $F(x) = 1 - e^{-\int_0^x r(t)dt}$ for $x \geq 0$ and $F(x) = 0$ for $x < 0$. To prove that $F(x)$ is a probability distribution function of a positive random variable, we must verify that $F(x)$ is increasing in x with $F(0) = 0$ and $\lim_{x\to\infty} F(x) = 1$. By $r(x) \geq 0$ for all $x \geq 0$, the function $\int_0^x r(t)dt$ is increasing in x, implying that $F(x)$ is increasing in x. It is obvious that $F(0) = 1 - 1 = 0$, while $\lim_{x\to\infty} F(x) = 1$ is a consequence of the fact that $\int_0^\infty r(t)dt = \infty$.

11
Jointly distributed random variables

In experiments, one is often interested not only in individual random variables, but also in relationships between two or more random variables. For example, if the experiment is the testing of a new medicine, the researcher might be interested in cholesterol level, blood pressure, and the glucose level of a test person. Similarly, a political scientist investigating the behavior of voters might be interested in the income and level of education of a voter. There are many more examples in the physical sciences, medical sciences, and social sciences. In applications, one often wishes to make inferences about one random variable on the basis of observations of other random variables. The purpose of this chapter is to familiarize the student with the notations and the techniques relating to experiments whose outcomes are described by two or more real numbers. The discussion is restricted to the case of pairs of random variables. Extending the notations and techniques to collections of more than two random variables is straightforward.

11.1 Joint probability densities

It is helpful to discuss the joint probability mass function of two discrete random variables before discussing the concept of the joint density of two continuous random variables. In fact, Section 9.3 has dealt with the joint distribution of discrete random variables. If X and Y are two discrete random variables defined on a same sample space with probability measure P, the mass function $p(x, y)$ defined by

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

is called the *joint probability mass function* of X and Y. As noted before, $P(X = x, Y = y)$ is the probability assigned by P to the intersection of the two sets

Table 11.1. *The joint probability mass function* $p(x, y)$.

$x\backslash y$	2	3	4	5	6	7	8	9	10	11	12	$p_X(x)$
1	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	0	0	0	$\frac{11}{36}$
2	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	0	0	$\frac{9}{36}$
3	0	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	0	$\frac{7}{36}$
4	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	0	$\frac{5}{36}$
5	0	0	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{2}{36}$	0	$\frac{3}{36}$
6	0	0	0	0	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$
$p_Y(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	sum $= 1$

$A = \{\omega : X(\omega) = x\}$ and $B = \{\omega : Y(\omega) = y\}$, with ω representing an element of the sample space. The joint probability mass function uniquely determines the probability distributions $p_X(x) = P(X = x)$ and $p_Y(y) = P(Y = y)$ by

$$p_X(x) = \sum_y P(X = x, Y = y), \quad p_Y(y) = \sum_x P(X = x, Y = y).$$

These distributions are called the *marginal distributions* of X and Y.

Example 11.1 Two fair dice are rolled. Let the random variable X represent the smallest of the outcomes of the two rolls, and let Y represent the sum of the outcomes of the two rolls. What is the joint probability mass function of X and Y?

Solution. The random variables X and Y are defined on the same sample space. The sample space is the set of all 36 pairs (i, j) for $i, j = 1, \ldots, 6$, where i and j are the outcomes of the first and second dice. A probability of $\frac{1}{36}$ is assigned to each element of the sample space. In Table 11.1, we give the joint probability mass function $p(x, y) = P(X = x, Y = y)$. For example, $P(X = 2, Y = 5)$ is the probability of the intersection of the sets $A = \{(2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 2), (4, 2), (5, 2), (6, 2)\}$ and $B = \{(1, 4), (4, 1), (2, 3), (3, 2)\}$. The set $\{(2, 3), (3, 2)\}$ is the intersection of these two sets and has probability $\frac{2}{36}$.

Problem 11.1 You roll a pair of dice. What is the joint probability mass function of the low and high points rolled?

Problem 11.2 Let X denote the number of hearts and Y the number of diamonds in a bridge hand. What is the joint probability mass function of X and Y?

The following example provides a good starting point for a discussion of joint probability densities.

Example 11.2 A point is picked at random inside a circular disc with radius r. Let the random variable X denote the length of the line segment between the center of the disc and the randomly picked point, and let the random variable Y denote the angle between this line segment and the horizontal axis (Y is measured in radians and so $0 \leq Y < 2\pi$). What is the joint distribution of X and Y?

Solution. The two continuous random variables X and Y are defined on a common sample space. The sample space consists of all points (v, w) in the two-dimensional plane with $v^2 + w^2 \leq r^2$, where the point $(0, 0)$ represents the center of the disc. The probability $P(A)$ assigned to each well-defined subset A of the sample space is taken as the area of region A divided by πr^2. The probability of the event of X taking on a value less than or equal to a and Y taking on a value less than or equal to b is denoted by $P(X \leq a,\ Y \leq b)$. This event occurs only if the randomly picked point falls inside the disc segment with radius a and angle b. The area of this disc segment is $\frac{b}{2\pi}\pi a^2$. Dividing this by πr^2 gives

$$P\left(X \leq a, Y \leq b\right) = \frac{b}{2\pi}\frac{a^2}{r^2} \qquad \text{for } 0 \leq a \leq r \text{ and } 0 \leq b \leq 2\pi.$$

We are now in a position to introduce the concept of joint density. Let X and Y be two random variables that are defined on the same sample space with probability measure P. The *joint cumulative probability distribution function* of X and Y is defined by $P(X \leq x, Y \leq y)$ for all x, y, where $P(X \leq x, Y \leq y)$ is a shorthand for $P(\{\omega : X(\omega) \leq x \text{ and } Y(\omega) \leq y\})$ and the symbol ω represents an element of the sample space.

Definition 11.1 *The continuous random variables X and Y are said to have a joint probability density function $f(x, y)$ if the joint cumulative probability distribution function $P(X \leq a, Y \leq b)$ allows for the representation*

$$P(X \leq a, Y \leq b) = \int_{x=-\infty}^{a}\int_{y=-\infty}^{b} f(x, y)\, dx\, dy, \qquad -\infty < a, b < \infty,$$

where the function $f(x, y)$ satisfies

$$f(x, y) \geq 0 \qquad \textit{for all } x, y \quad \textit{and} \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y) dx dy = 1.$$

Just as in the one-dimensional case, $f(a, b)$ allows for the interpretation

$$f(a,b)\,\Delta a\,\Delta b$$
$$\approx P\left(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a, b - \frac{1}{2}\Delta b \leq Y \leq b + \frac{1}{2}\Delta b\right)$$

for small positive values of Δa and Δb provided that $f(x, y)$ is continuous in the point (a, b). In other words, the probability that the random point (X, Y) falls into a small rectangle with sides of lengths Δa, Δb around the point (a, b) is approximately given by $f(a, b)\,\Delta a\,\Delta b$.

To obtain the joint probability density function $f(x, y)$ of the random variables X and Y in Example 11.2, we take the partial derivatives of $P(X \leq x, Y \leq y)$ with respect to x and y. It then follows from

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} P(X \leq x, Y \leq y)$$

that

$$f(x,y) = \begin{cases} \frac{1}{2\pi}\frac{2x}{r^2} & \text{for } 0 < x < r \text{ and } 0 < y < 2\pi, \\ 0 & \text{otherwise.} \end{cases}$$

In general, the joint probability density function is found by determining first the cumulative joint probability distribution function and taking next the partial derivatives. However, sometimes it is easier to find the joint probability density function by using its probabilistic interpretation. This is illustrated with the next example.

Example 11.3 The pointer of a spinner of radius r is spun three times. The three spins are performed independently of each other. With each spin, the pointer stops at an unpredictable point on the circle. The random variable L_i corresponds to the length of the arc from the top of the circle to the point where the pointer stops on the ith spin. The length of the arc is measured clockwise. Let $X = \min(L_1, L_2, L_3)$ and $Y = \max(L_1, L_2, L_3)$. What is the joint probability density function $f(x, y)$ of the two continuous random variables X and Y?

Solution. We can derive the joint probability density function $f(x, y)$ by using the interpretation that the probability $P(x < X \leq x + \Delta x, y < Y \leq y + \Delta y)$ is approximately equal to $f(x, y)\Delta x \Delta y$ for Δx and Δy small. The event $\{x < X \leq x + \Delta x, y < Y \leq y + \Delta y\}$ occurs only if one of the L_i takes on a value between x and $x + \Delta x$, one of the L_i a value between y and $y + \Delta y$, and the remaining L_i a value between x and y, where $0 < x < y$. There are $3 \times 2 \times 1 = 6$ ways in which L_1, L_2, L_3 can be arranged and the probability that for fixed i the random variable L_i takes on a value between a and b equals

$(b-a)/(2\pi r)$ for $0 < a < b < 2\pi r$ (explain!). Thus, by the independence of L_1, L_2, and L_3 (see the general Definition 9.2)

$$\begin{aligned} &P(x < X \leq x + \Delta x, y < Y \leq y + \Delta y) \\ &\quad = 6 \frac{(x + \Delta x - x)}{2\pi r} \frac{(y + \Delta y - y)}{2\pi r} \frac{(y - x)}{2\pi r}. \end{aligned}$$

Hence, the joint probability density function of X and Y is given by

$$f(x, y) = \begin{cases} \frac{6(y-x)}{(2\pi r)^3} & \text{for } 0 < x < y < 2\pi r \\ 0 & \text{otherwise.} \end{cases}$$

In general, if the random variables X and Y have a joint probability density function $f(x, y)$

$$P((X, Y) \in C) = \int\int_C f(x, y)\, dx\, dy$$

for any set C of pairs of real numbers. In calculating a double integral over a nonnegative integrand, it does not matter whether we integrate over x first or over y first. This is a basic fact from calculus. The double integral can be written as a repeated one-dimensional integral. The expression for $P((X, Y) \in C)$ is very useful to determine the probability distribution function of any function $g(X, Y)$ of X and Y. To illustrate this, we derive the useful result that the sum $Z = X + Y$ has the probability density

$$f_Z(z) = \int_{-\infty}^{\infty} f(u, z - u)\, du.$$

To prove this *convolution formula*, note that

$$\begin{aligned} P(Z \leq z) &= \iint\limits_{\substack{(x,y):\\ x+y \leq z}} f(x, y)\, dx\, dy = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{z-x} f(x, y)\, dx\, dy \\ &= \int_{v=-\infty}^{z} \int_{u=-\infty}^{\infty} f(u, v - u)\, du\, dv, \end{aligned}$$

using the change of variables $u = x$ and $v = x + y$. Next, differentiation of $P(Z \leq z)$ yields the convolution formula for $f_Z(z)$. If the random variables X and Y are nonnegative, the convolution formula reduces to

$$f_Z(z) = \int_0^z f(u, z - u)\, du \qquad \text{for } z > 0.$$

Uniform distribution over a region

Another useful result is the following. Suppose that a point (X, Y) is picked at random inside a bounded region R in the two-dimensional plane. Then, the joint probability density function $f(x, y)$ of X and Y is given by the uniform density

$$f(x, y) = \frac{1}{\text{area of region } R} \qquad \text{for } (x, y) \in R.$$

The proof is simple. For any subset $C \subseteq R$

$$P((X, Y) \in C) = \frac{\text{area of } C}{\text{area of } R},$$

being the mathematical definition of the random selection of a point inside the region R. Integral calculus tells us that area of $C = \iint_C dxdy$. Thus, for any subset $C \subseteq R$

$$P((X, Y) \in C) = \iint_C \frac{1}{\text{area of } R}\, dx\, dy,$$

showing that the random point (X, Y) has the above density $f(x, y)$.

In the following problems you are asked to apply the basic expression $P((X, Y) \in C) = \iint_C f(x, y)\, dx\, dy$ yourselves in order to find the probability density of a given function of X and Y.

Problem 11.3 A point (X, Y) is picked at random inside the triangle consisting of the points (x, y) in the plane with $x, y \geq 0$ and $x + y \leq 1$. What is the joint probability density of the point (X, Y)? Determine the probability density of each of the random variables $X + Y$ and $\max(X, Y)$.

Problem 11.4 Let X and Y be two random variables with a joint probability density

$$f(x, y) = \begin{cases} \frac{1}{(x+y)^3} & \text{for } x, y > c \\ 0 & \text{otherwise,} \end{cases}$$

for an appropriate constant c. Verify that $c = \frac{1}{4}$ and calculate the probability $P(X > a, Y > b)$ for $a, b > c$.

Problem 11.5 Independently of each other, two points are chosen at random in the interval (0, 1). What is the joint probability density of the smallest and the largest of these two random numbers? What is the probability density of the length of the middle interval of the three intervals that result from the two random points in (0,1)? What is the probability that the smallest of the three resulting intervals is larger than a?

Problem 11.6 Independently of each other, two numbers X and Y are chosen at random in the interval (0, 1). Let $Z = X/Y$ be the ratio of these two random numbers.

(a) Use the joint density of X and Y to verify that $P(Z \leq z)$ equals $\frac{1}{2}z$ for $0 < z < 1$ and equals $1 - 1/(2z)$ for $z \geq 1$.

(b) What is the probability that the first significant (nonzero) digit of Z equals 1? What about the digits $2, \ldots, 9$?

(c) What is the answer to Question (b) for the random variable $V = XY$?

(d) What is the density function of the random variable $(X/Y)U$ when U is a random number from (0, 1) that is independent of X and Y?

11.2 Marginal probability densities

If the two random variables X and Y have a joint probability density function $f(x, y)$, then each of the random variables X and Y has a probability density itself. Using the fact that $\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n)$ for any nondecreasing sequence of events A_n, it follows that

$$P(X \leq a) = \lim_{b\to\infty} P\left(X \leq a, Y \leq b\right) = \int_{-\infty}^{a} \left[\int_{-\infty}^{\infty} f(x, y)\, dy\right] dx.$$

This representation shows that X has probability density function

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy, \qquad -\infty < x < \infty.$$

In the same way, the random variable Y has probability density function

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx, \qquad -\infty < y < \infty.$$

The probability density functions $f_X(x)$ and $f_Y(y)$ are called the *marginal probability density functions* of X and Y. The following interpretation can be given to the marginal density $f_X(x)$ at the point $x = a$ when a is a continuity point of $f_X(x)$. For Δa small, $f_X(a)\Delta a$ gives approximately the probability that (X, Y) falls in a vertical strip in the two-dimensional plane with width Δa and around the vertical line $x = a$. A similar interpretation applies to $f_Y(b)$ for any continuity point b of $f_Y(y)$.

Example 11.4 A point (X, Y) is chosen at random inside the unit circle. What is the marginal density of X?

Solution. Denote by $C = \{(x, y) \mid x^2 + y^2 \leq 1\}$ the unit circle. The joint probability density function $f(x, y)$ of X and Y is given by $f(x, y) = 1/(\text{area of } C)$

for $(x, y) \in C$. Hence

$$f(x, y) = \begin{cases} \frac{1}{\pi} & \text{for } (x, y) \in C \\ 0 & \text{otherwise.} \end{cases}$$

Using the fact that $f(x, y)$ is equal to zero for those y satisfying $y^2 > 1 - x^2$, if follows that

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\, dy,$$

and so

$$f_X(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & \text{for } -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Can you explain why the marginal density of X is not the uniform density on $(-1, 1)$? *Hint*: interpret $P(x < X \leq x + \Delta x)$ as the area of a vertical strip in the unit circle.

Problem 11.7 A point (X, Y) is chosen at random in the equilateral triangle having $(0, 0)$, $(1, 0)$, and $(\frac{1}{2}, \frac{1}{2}\sqrt{3})$ as corner points. Determine the marginal densities of X and Y. Before determining the function $f_X(x)$, can you explain why $f_X(x)$ must be largest at $x = \frac{1}{2}$?

A general condition for the independence of the jointly distributed random variables X and Y is stated in Definition 9.2. In terms of the marginal densities, the continuous analog of Rule 9.6 for the discrete case is:

Rule 11.1 *The jointly distributed random variables X and Y are independent if and only if*

$$f(x, y) = f_X(x) f_Y(y) \qquad \textit{for all } x, y.$$

Let us illustrate this with the random variables X and Y from Example 11.2. Then, we obtain from $f_X(x) = \int_0^{2\pi} \frac{x}{\pi r^2}\, dy$ that

$$f_X(x) = \begin{cases} \frac{2x}{r^2} & \text{for } 0 < x < r, \\ 0 & \text{otherwise.} \end{cases}$$

In the same way, we obtain from $f_Y(y) = \int_0^r \frac{x}{\pi r^2}\, dx$ that

$$f_Y(y) = \begin{cases} \frac{1}{2\pi} & \text{for } 0 < y < 2\pi, \\ 0 & \text{otherwise.} \end{cases}$$

The calculations lead to the intuitively obvious result that the angle Y has a uniform distribution on $(0, 2\pi)$. A somewhat more surprising result is that the distance X and the angle Y are independent random variables, though there is dependence between the components of the randomly picked point. The independence of X and Y follows from the observation that $f(x, y) = f_X(x) f_Y(y)$ for all x, y.

To conclude this subsection, we give a very important result for the exponential distribution.

Example 11.5 Suppose that X and Y are independent random variables, where X is exponentially distributed with expected value $1/\alpha$ and Y is exponentially distributed with expected value $1/\beta$. What is the probability distribution of $\min(X, Y)$? What is the probability that X is less than Y?

Solution. The answer to the first question is that $\min(X, Y)$ is exponentially distributed with expected value $1/(\alpha + \beta)$. It holds that

$$P(\min(X, Y) \leq z) = 1 - e^{-(\alpha+\beta)z} \quad \text{for } z \geq 0 \quad \text{and} \quad P(X < Y) = \frac{\alpha}{\alpha + \beta}.$$

The proof is simple. Noting that $P(\min(X, Y) \leq z) = 1 - P(X > z, Y > z)$, we have

$$P(\min(X, Y) \leq z) = 1 - \int_{x=z}^{\infty} \int_{y=z}^{\infty} f_X(x) f_Y(y)\, dx\, dy.$$

Also,

$$P(X < Y) = \int_{x=0}^{\infty} \int_{y=x}^{\infty} f_X(x) f_Y(y)\, dx\, dy.$$

Using the fact that $f_X(x) = \alpha e^{-\alpha x}$ and $f_Y(y) = \beta e^{-\beta y}$, it is next a matter of simple algebra to derive the results. The details are left to the reader.

Problem 11.8 The continuous random variables X and Y are nonnegative and independent. Verify that the density function of $Z = X + Y$ is given by the convolution formula

$$f_Z(z) = \int_0^z f_X(z - y) f_Y(y) dy \qquad \text{for } z \geq 0.$$

Problem 11.9 The nonnegative random variables X and Y are independent and uniformly distributed on (c, d). What is the probability density of $Z = X + Y$? What is the probability density function of $V = X^2 + Y^2$? Use the latter density to calculate the expected value of the distance of a point chosen at random inside the unit square to the center of the unit square.

11.2.1 Substitution rule

The expected value of a given function of jointly distributed random variables X and Y can be calculated by the two-dimensional substitution rule. In the continuous case, we have:

Rule 11.2 *If the random variables X and Y have a joint probability density function $f(x, y)$, then*

$$E\left[g(X, Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy$$

for any function $g(x, y)$ provided that the integral is well defined.

An easy consequence of Rule 11.2 is that

$$E(aX + bY) = aE(X) + bE(Y)$$

for any constants a, b provided that $E(X)$ and $E(Y)$ exist. To see this, note that

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax + by) f(x, y)\, dx\, dy \\ &\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} axf(x, y)\, dx\, dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} byf(x, y)\, dx\, dy \\ &\quad = \int_{x=-\infty}^{\infty} ax\, dx \int_{y=-\infty}^{\infty} f(x, y)\, dy + \int_{y=-\infty}^{\infty} by\, dy \int_{x=-\infty}^{\infty} f(x, y)\, dx \\ &\quad = a\int_{-\infty}^{\infty} xf_X(x)\, dx + b\int_{-\infty}^{\infty} yf_Y(y)\, dy,\end{aligned}$$

which proves the desired result. It is left to the reader to verify from Rules 11.1 and 11.2 that

$$E(XY) = E(X)E(Y) \qquad \text{for independent } X \text{ and } Y.$$

An illustration of the substitution rule is provided by Problem 2.21: what is the expected value of the distance between two points that are chosen at random in the interval (0, 1)? To answer this question, let X and Y be two independent random variables that are uniformly distributed on (0, 1). The joint density function of X and Y is given by $f(x, y) = 1$ for all $0 < x, y < 1$. The substitution rule gives

$$\begin{aligned}
E(|X-Y|) &= \int_0^1 \int_0^1 |x-y|\,dxdy \\
&= \int_0^1 dx \left[\int_0^x (x-y)\,dy + \int_x^1 (y-x)\,dy \right] \\
&= \int_0^1 \left[\frac{1}{2}x^2 + \frac{1}{2} - \frac{1}{2}x^2 - x(1-x) \right] dx = \frac{1}{3}.
\end{aligned}$$

Hence, the answer to the question is $\frac{1}{3}$.

As another illustration of Rule 11.2, consider Example 11.2 again. In this example, a point is picked at random inside a circular disk with radius r and the point (0, 0) as center. What is the expected value of the rectangular distance from the randomly picked point to the center of the disk? This rectangular distance is given by $|X\cos(Y)| + |X\sin(Y)|$ (the rectangular distance from point (a, b) to (0, 0) is defined by $|a| + |b|$). For the function $g(x, y) = |x\cos(y)| + |x\sin(y)|$, we find

$$\begin{aligned}
E\left[g(X,Y)\right] &= \int_0^r \int_0^{2\pi} \{x|\cos(y)| + x|\sin(y)|\} \frac{x}{\pi r^2}\,dx\,dy \\
&= \frac{1}{\pi r^2} \int_0^{2\pi} |\cos(y)|\,dy \int_0^r x^2\,dx + \frac{1}{\pi r^2} \int_0^{2\pi} |\sin(y)|\,dy \int_0^r x^2\,dx \\
&= \frac{r^3}{3\pi r^2} \left[\int_0^{2\pi} |\cos(y)|\,dy + \int_0^{2\pi} |\sin(y)|\,dy \right] = \frac{8r}{3\pi}.
\end{aligned}$$

The same ideas hold in the discrete case with the probability mass function assuming the role of the density function

$$E[g(X,Y)] = \sum_x \sum_y g(x,y)p(x,y)$$

when the random variables X and Y have the joint probability mass function $p(x, y) = P(X = x, Y = y)$.

11.3 Transformation of random variables

In statistical applications, one sometimes needs the joint density of two random variables V and W that are defined as functions of two other random variables X and Y having a joint density $f(x, y)$. Suppose that the random variables V and W are defined by $V = g(X, Y)$ and $W = h(X, Y)$ for given functions g and h. What is the joint probability density function of V and W? An answer to this question will be given under the assumption that the transformation is one-to-one. That is, it is assumed that the equations $v = g(x, y)$ and $w = h(x, y)$ can be solved uniquely to yield functions $x = a(v, w)$ and $y = b(v, w)$. Also assume that the partial derivatives of the functions $a(v, w)$ and $b(v, w)$ with respect to v and w are continuous in (v, w). Then the following transformation rule holds:

Rule 11.3 *The joint probability density function of V and W is given by*

$$f(a(v, w), b(v, w))|J(v, w)|,$$

where the Jacobian $J(v, w)$ is given by the determinant

$$\begin{vmatrix} \frac{\partial a(v,w)}{\partial v} & \frac{\partial a(v,w)}{\partial w} \\ \frac{\partial b(v,w)}{\partial v} & \frac{\partial b(v,w)}{\partial w} \end{vmatrix} = \frac{\partial a(v, w)}{\partial v}\frac{\partial b(v, w)}{\partial w} - \frac{\partial a(v, w)}{\partial w}\frac{\partial b(v, w)}{\partial v}.$$

The proof of this rule is omitted. This transformation rule looks intimidating, but is easy to use in many applications. In the next section it will be shown how Rule 11.3 can be used to devise a method for simulating from the normal distribution. However, we first give a simple illustration of Rule 11.3. Suppose that X and Y are independent $N(0, 1)$ random variables. Then, the random variables $V = X + Y$ and $W = X - Y$ are normally distributed and independent. To verify this, note that the inverse functions $a(v, w)$ and $b(v, w)$ are given by $x = \frac{v+w}{2}$ and $y = \frac{v-w}{2}$. Thus, the Jacobian $J(v, w)$ is equal to

$$\begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Since X and Y are independent $N(0, 1)$ random variables, it follows from Rule 11.1 that their joint density function is given by

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

Applying Rule 11.3, we obtain that the joint density function of V and W is given by

$$\begin{aligned} f_{V,W}(v, w) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{v+w}{2})^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{v-w}{2})^2} \times \frac{1}{2} \\ &= \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{1}{2}v^2/2} \times \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{1}{2}w^2/2}, \qquad -\infty < v, w < \infty. \end{aligned}$$

This implies that $f_{V,W}(v, w) = f_V(v) f_W(w)$ for all v, w with the marginal density functions $f_V(v) = \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{1}{2}v^2/2}$ and $f_W(w) = \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-\frac{1}{2}w^2/2}$. Using

Rule 11.1 again, it now follows that $V = X + Y$ and $W = X - Y$ are $N(0, 2)$ distributed and independent.

11.3.1 Simulating from a normal distribution

A natural transformation of two independent standard normal random variables leads to a practically useful method for simulating random observations from the standard normal distribution. Suppose that X and Y are independent random variables each having the standard normal distribution. Using Rule 11.1, the joint probability density function of X and Y is given by

$$f(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}.$$

The random vector (X, Y) can be considered as a point in the two-dimensional plane. Let the random variable V be the distance from the point (0, 0) to the point (X, Y) and let W be the angle that the line through the points (0, 0) and (X, Y) makes with the horizontal axis. The random variables V and W are functions of X and Y (the function $g(x, y) = \sqrt{x^2 + y^2}$ and $h(x, y) = \arctan(y/x)$). The inverse functions $a(v, w)$ and $b(v, w)$ are very simple. By basic geometry, $x = v\cos(w)$ and $y = v\sin(w)$. We thus obtain the Jacobian

$$\begin{vmatrix} \cos(w) & -v\sin(w) \\ \sin(w) & v\cos(w) \end{vmatrix} = v\cos^2(w) + v\sin^2(w) = v,$$

using the celebrated identity $\cos^2(w) + \sin^2(w) = 1$. Hence, the joint probability density function of V and W is given by

$$f_{V,W}(v, w) = \frac{v}{2\pi} e^{-\frac{1}{2}\left(v^2\cos^2(w)+v^2\sin^2(w)\right)} = \frac{v}{2\pi} e^{-\frac{1}{2}v^2}$$

for $0 < v < \infty$ and $0 < w < 2\pi$. The marginal densities of V and W are given by

$$f_V(v) = \frac{1}{2\pi}\int_0^{2\pi} v e^{-\frac{1}{2}v^2}\, dw = v e^{-\frac{1}{2}v^2}, \qquad 0 < v < \infty$$

and

$$f_W(w) = \frac{1}{2\pi}\int_0^{\infty} v e^{-\frac{1}{2}v^2}\, dv = \frac{1}{2\pi}, \qquad 0 < w < 2\pi.$$

Since $f_{V,W}(v, w) = f_V(v)f_W(w)$, we have the remarkable finding that V and W are independent random variables. The random variable V has the probability density function $ve^{-\frac{1}{2}v^2}$ for $v > 0$ and W is uniformly distributed on $(0, 2\pi)$. This result is extremely useful for simulation purposes. Using the inverse-transformation method from Section 10.3, it is a simple matter to simulate random observations from the probability distributions of V and W. If we let U_1 and U_2 denote two independent random numbers from the interval (0,1), it follows from results in Section 10.3 that random observations of V and W are given by

$$V = \sqrt{-2\ln(1-U_1)} \quad \text{and} \quad W = 2\pi U_2.$$

Next, one obtains two random observations X and Y from the standard normal distribution by taking

$$X = V\cos(W) \quad \text{and} \quad Y = V\sin(W).$$

Theoretically, X and Y are independent of each other. However, if a pseudo-random generator is used to generate U_1 and U_2, one uses only one of two variates X and Y. It surprisingly appears that the points (X, Y) lie on a spiral in the plane when a multiplicative generator is used for the pseudo-random numbers. The explanation of this subtle dependency lies in the fact that pseudo-random numbers are not truly random. The method described above for generating normal variates is known as the Box-Muller method.

Problem 11.10 A point (V, W) is chosen inside the unit circle as follows. First, a number R is chosen at random between 0 and 1. Next, a point is chosen at random on the circumference of the circle with radius R. Use the transformation formula to find the joint density function of this point (V, W). What is the marginal density function of each of the components of the point (V, W)? Can you intuitively explain why the point (V, W) is not uniformly distributed over the unit circle?

Problem 11.11 Let (X, Y) be a point chosen at random inside the unit circle. Define V and W by $V = X\sqrt{-2\ln(Q)/Q}$ and $W = Y\sqrt{-2\ln(Q)/Q}$, where $Q = X^2 + Y^2$. Verify that the random variables V and W are independent and $N(0, 1)$ distributed. This method for generating normal variates is known as Marsaglia's polar method.

Problem 11.12 The independent random variables Z and Y have a standard normal distribution and a chi-square distribution with ν degrees of freedom. Use the transformation $V = Y$ and $W = Z/\sqrt{Y/\nu}$ to prove that the random variable $W = Z/\sqrt{Y/\nu}$ has a Student-t density with ν degrees of freedom. *Hint*: in evaluating $f_W(w)$ from $\int_0^\infty f_{V,W}(v, w)\, dv$, use the fact that the gamma density $\lambda^\alpha x^{\alpha-1}e^{-\lambda x}/\,\Gamma(\alpha)$ integrates to 1 over $(0, \infty)$.

11.4 Covariance and correlation coefficient

Let the random variables X and Y be defined on the same sample space with probability measure P. A basic rule in probability is that the expected value of the sum $X + Y$ equals the sum of the expected values of X and Y. Does a similar rule hold for the variance of the sum $X + Y$? To answer this question, we apply the definition of variance. The variance of $X + Y$ equals

$$\begin{aligned} &E[\{X + Y - E(X + Y)\}^2] \\ &\quad = E[(X - E(X))^2 + 2(X - E(X))(Y - E(Y)) + (Y - E(Y))^2] \\ &\quad = \text{var}(X) + 2E[(X - E(X))(Y - E(Y))] + \text{var}(Y). \end{aligned}$$

This leads to the following general definition.

Definition 11.2 *The covariance cov(X, Y) of two random variables X and Y is defined by*

$$\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

whenever the expectations exist.

The formula for cov(X, Y) can be written in the equivalent form

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

by expanding $(X - E(X))(Y - E(Y))$ into $XY - XE(Y) - YE(X) + E(X)E(Y)$ and noting that the expectation is a linear operator. Using the fact that $E(XY) = E(X)E(Y)$ for independent random variables, the alternative formula for cov(X, Y) has as direct consequence:

Rule 11.4 *If X and Y are independent random variables, then*

$$\text{cov}(X, Y) = 0.$$

However, the converse of this result is not always true. A simple example of two dependent random variables X and Y having covariance zero is given in Section 9.4. Another counterexample is provided by the random variables $X = Z$ and $Y = Z^2$, where Z has the standard normal distribution. Nevertheless, cov(X, Y) is often used as a measure of the dependence of X and Y. The covariance appears over and over in practical applications (see the discussion in Section 5.2).

Using the definition of covariance and the above expression for var$(X + Y)$, we find the general rule:

Rule 11.5 *For any two random variables X and Y*

$$\text{var}(X + Y) = \text{var}(X) + 2\text{cov}(X, Y) + \text{var}(Y).$$

If the random variables X and Y are independent, then

$$\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y).$$

The units of $\mathrm{cov}(X, Y)$ are not the same as the units of $E(X)$ and $E(Y)$. Therefore, it is often more convenient to use the *correlation coefficient* of X and Y which is defined by

$$\rho(X, Y)=\frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}},$$

provided that $\mathrm{var}(X)>0$ and $\mathrm{var}(Y)>0$. The correlation coefficient is a dimensionless quantity with the property that

$$-1 \leq \rho(X, Y) \leq 1.$$

The reader is asked to prove this property in Problem 11.14. The random variables X and Y are said to be *uncorrelated* if $\rho(X, Y)=0$. Independent random variables are always uncorrelated, but the converse is not always true. If $\rho(X, Y)=\pm 1$, then Y is fully determined by X. In this case it can be shown that $Y=aX+b$ for constants a and b with $a \neq 0$.

The problem section of Chapter 5 contains several exercises on the covariance and correlation coefficient. Here are some more exercises.

Problem 11.13 The continuous random variables X and Y have the joint density $f(x, y)=4y^2$ for $0<x<y<1$ and $f(x, y)=0$ otherwise. What is the correlation coefficient of X and Y? Can you intuitively explain why this correlation coefficient is positive?

Problem 11.14 Verify that

$$\mathrm{var}(aX+b)=a^2\mathrm{var}(X) \quad \text{and} \quad \mathrm{cov}(aX, bY)=ab\,\mathrm{cov}(X, Y)$$

for any constants a, b. Next, evaluate the variance of the random variable $Z = Y/\sqrt{\mathrm{var}(Y)}-\rho(X, Y)X/\sqrt{\mathrm{var}(X)}$ to prove that $-1 \leq \rho(X, Y) \leq 1$. Also, for any constants a, b, c, and d, verify that $\mathrm{cov}(aX+bY, cV+dW)$ can be worked out as $ac\mathrm{cov}(X, V)+ad\mathrm{cov}(X, W)+bc\mathrm{cov}(Y, V)+bd\mathrm{cov}(Y, W)$.

Problem 11.15 The amounts of rainfall in Amsterdam during each of the months January, February, . . ., December are independent random variables with expected values of 62.1, 43.4, 58.9, 41.0, 48.3, 67.5, 65.8, 61.4, 82.1, 85.1, 89.0, and 74.9 mm and with standard deviations of 33.9, 27.8, 31.1, 24.1, 29.3, 33.8, 36.8, 32.1, 46.6, 42.4, 40.0, and 36.2 mm. What are the expected value and the standard deviation of the annual rainfall in Amsterdam? Calculate an approximate value for the probability that the total rainfall in Amsterdam next year will be larger than 1,000 mm.

Problem 11.16 Let the random variables $X_1, \ldots, X_n$ be defined on a common probability space. Prove that

$$\mathrm{var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{var}(X_i) + 2\sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathrm{cov}(X_i, X_j).$$

Next, evaluate $\mathrm{var}(\sum_{i=1}^{n} t_i X_i)$ in order to verify that $\sum_{i=1}^{n} \sum_{j=1}^{n} t_i t_j \sigma_{ij} \geq 0$ for all real numbers $t_1, \ldots, t_n$, where $\sigma_{ij} = \mathrm{cov}(X_i, X_j)$. In other words, the covariance matrix $C = (\sigma_{ij})$ is positive semi-definite.

Problem 11.17 The hypergeometric distribution describes the probability mass function of the number of red balls drawn when n balls are randomly chosen from an urn containing R red and W white balls. Show that the variance of the number of red balls drawn is given by $n\frac{R}{R+W}(1 - \frac{R}{R+W})\frac{R+W-n}{R+W-1}$. *Hint:* the number of red balls drawn can be written as $X_1 + \ldots + X_R$, where X_i equals 1 if the ith red ball is selected and 0 otherwise.

Problem 11.18 What is the variance of the number of distinct birthdays within a randomly formed group of 100 persons? *Hint*: define the random variable X_i as 1 if the ith day is among the 100 birthdays, and as 0 otherwise.

Problem 11.19 You roll a pair of dice. What is the correlation coefficient of the high and low points rolled?

Problem 11.20 What is the correlation coefficient of the Cartesian coordinates of a point picked at random in the unit circle?

11.4.1 Linear predictor

Suppose that X and Y are two dependent random variables. In statistical applications, it is often the case that we can observe the random variable X but we want to know the dependent random variable Y. A basic question in statistics is: what is the best *linear* predictor of Y with respect to X? That is, for which linear function $y = \alpha + \beta x$ is

$$E[(Y - \alpha - \beta X)^2]$$

minimal? The answer to this question is

$$y = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X),$$

where $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X = \sqrt{\mathrm{var}(X)}$, $\sigma_Y = \sqrt{\mathrm{var}(Y)}$, and $\rho_{XY} = \rho(X, Y)$. The derivation is simple. Rewriting $y = \alpha + \beta x$ as $y = \mu_Y + \beta(x - \mu_X) - (\mu_Y - \alpha - \beta\mu_X)$, it follows after some algebra that $E[(Y - \alpha - \beta X)^2]$

can be evaluated as

$$\begin{aligned}
&E[\{Y-\mu_Y-\beta(X-\mu_X)+\mu_Y-\alpha-\beta\mu_X\}^2]\\
&\quad=E[\{Y-\mu_Y-\beta(X-\mu_X)\}^2]+(\mu_Y-\alpha-\beta\mu_X)^2\\
&\qquad+2(\mu_Y-\alpha-\beta\mu_X)E[Y-\mu_Y-\beta(X-\mu_X)]\\
&\quad=\sigma_Y^2+\beta^2\sigma_X^2-2\beta\rho_{XY}\sigma_X\sigma_Y+(\mu_Y-\alpha-\beta\mu_X)^2.
\end{aligned}$$

In order to minimize this quadratic function in α and β, we put the partial derivatives of the function with respect to α and β equal to zero. This leads after some simple algebra to

$$\beta=\frac{\rho_{XY}\sigma_Y}{\sigma_X}\quad\text{and}\quad\alpha=\mu_Y-\frac{\rho_{XY}\sigma_Y}{\sigma_X}\mu_X.$$

For these values of α and β, we have the minimal value

$$E\left[(Y-\alpha-\beta X)^2\right]=\sigma_Y^2\left(1-\rho_{XY}^2\right).$$

This minimum is sometimes called the residual variance of Y.

The phenomenon of *regression to the mean* can be explained with the help of the best linear predictor. Think of X as the height of a 25-year-old father and think of Y as the height his newborn son will have at the age of 25 years. It is reasonable to assume that $\mu_X=\mu_Y=\mu$, $\sigma_X=\sigma_Y=\sigma$, and $\rho=\rho(X,Y)$ is positive. The best linear predictor $\hat{Y}$ of Y then satisfies $\hat{Y}-\mu=\rho(X-\mu)$ with $0<\rho<1$. If the height of the father scores above the mean, the best linear prediction is that the height of the son will score closer to the mean. Very tall fathers tend to have somewhat shorter sons and very short fathers somewhat taller ones! Regression to the mean shows up in a wide variety of places: it helps explain why great movies have often disappointing sequels, and disastrous presidents have often better successors.

12
Multivariate normal distribution

Do the one-dimensional normal distribution and the one-dimensional central limit theorem allow for a generalization to dimension two or higher? The answer is yes. Just as the one-dimensional normal density is completely determined by its expected value and variance, the bivariate normal density is completely specified by the expected values and the variances of its marginal densities and by its correlation coefficient. The bivariate normal distribution appears in many applied probability problems. This probability distribution can be extended to the multivariate normal distribution in higher dimensions. The multivariate normal distribution arises when you take the sum of a large number of independent random vectors. To get this distribution, all you have to do is to compute a vector of expected values and a matrix of covariances. The multidimensional central limit theorem explains why so many natural phenomena have the multivariate normal distribution. A nice feature of the multivariate normal distribution is its mathematical tractability. The fact that any linear combination of multivariate normal random variables has a univariate normal distribution makes the multivariate normal distribution very convenient for financial portfolio analysis, among others.

12.1 Bivariate normal distribution

A random vector (X, Y) is said to have a *standard bivariate normal distribution* with parameter ρ if it has a joint probability density function of the form

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\, e^{-\frac{1}{2}(x^2-2\rho xy+y^2)/(1-\rho^2)}, \qquad -\infty < x, y < \infty,$$

where ρ is a constant with $-1 < \rho < 1$. Before showing that ρ can be interpreted as the correlation coefficient of X and Y, we derive the marginal densities of X and Y. Therefore, we first decompose the joint density function $f(x, y)$ as

$$f(x, y) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\; \frac{1}{\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\rho x)^2/(1-\rho^2)}.$$

Next observe that, for *fixed* x,

$$g(y) = \frac{1}{\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\rho x)^2/(1-\rho^2)}$$

is an $N(\rho x, 1-\rho^2)$ density. This implies that $\int_{-\infty}^{\infty} g(y)\,dy = 1$ and so

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}, \qquad -\infty < x < \infty.$$

In other words, the marginal density $f_X(x)$ of X is the standard normal density. Also, for reasons of symmetry, the marginal density $f_Y(y)$ of Y is the standard normal density. Next, we prove that ρ is the correlation coefficient $\rho(X, Y)$ of X and Y. Since $\text{var}(X) = \text{var}(Y) = 1$, it suffices to verify that $\text{cov}(X, Y) = \rho$. To do so, we use again the above decomposition of $f(x, y)$. By $E(X) = E(Y) = 0$, we have $\text{cov}(X, Y) = E(XY)$. Thus, letting $\tau^2 = 1 - \rho^2$,

$$\begin{aligned}
\text{cov}(X, Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x, y)\,dx\,dy \\
&= \int_{x=-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\,dx \int_{y=-\infty}^{\infty} y\frac{1}{\tau\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\rho x)^2/\tau^2}\,dy \\
&= \int_{-\infty}^{\infty} \rho x^2 \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\,dx = \rho,
\end{aligned}$$

where the third equality uses the fact that the expected value of an $N(\rho x, \tau^2)$ random variable is ρx and the last equality uses the fact that $E(Z^2) = \sigma^2(Z) = 1$ for a standard normal random variable Z.

A random vector (X, Y) is said to be *bivariate normal* distributed with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ if the standardized random vector

$$\left(\frac{X-\mu_1}{\sigma_1}, \frac{Y-\mu_2}{\sigma_2}\right)$$

has the standard bivariate normal distribution with parameter ρ. In this case the joint density $f(x, y)$ of the random variables X and Y is given by

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}[(\frac{x-\mu_1}{\sigma_1})^2 - 2\rho(\frac{x-\mu_1}{\sigma_1})(\frac{y-\mu_2}{\sigma_2}) + (\frac{y-\mu_2}{\sigma_2})^2]/(1-\rho^2)}.$$

Rule 12.1 *Suppose that the random vector (X, Y) has a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Then,*

(a) *The marginal densities $f_X(x)$ and $f_Y(y)$ of X and Y are the $N(\mu_1, \sigma_1^2)$ density and the $N(\mu_2, \sigma_2^2)$ density.*
(b) *The correlation coefficient of X and Y is given by $\rho(X, Y) = \rho$.*

The result (a) follows directly from the fact that $(X - \mu_1)/\sigma_1$ and $(Y - \mu_2)/\sigma_2$ are $N(0, 1)$ distributed, as was verified above. Also, it was shown above that the covariance of $(X - \mu_1)/\sigma_1$ and $(Y - \mu_2)/\sigma_2$ equals ρ. Using the basic formula $\text{cov}(aX + b, cY + d) = ac\text{cov}(X, Y)$ for any constants a, b, c, and d, we next find the desired result

$$\rho = \text{cov}\left(\frac{X - \mu_1}{\sigma_1}, \frac{Y - \mu_2}{\sigma_2}\right) = \frac{1}{\sigma_1\sigma_2}\text{cov}(X, Y) = \rho(X, Y).$$

In general, uncorrelatedness is a necessary but not sufficient condition for independence of two random variables. However, for a bivariate normal distribution, uncorrelatedness is a necessary and sufficient condition for independence:

Rule 12.2 *Bivariate normal random variables X and Y are independent if and only if they are uncorrelated.*

This important result is a direct consequence of Rule 11.1, since the above representation of the bivariate normal density $f(x, y)$ reveals that $f(x, y) = f_X(x)f_Y(y)$ if and only if $\rho = 0$.

As already pointed out, the bivariate normal distribution has the important property that its marginal distributions are one-dimensional normal distributions. More generally, it can be shown that the random variables X and Y have a bivariate normal distribution if and only if $aX + bY$ is normally distributed for any constants a and b.[†] The "only if" part of this result can be proved by elementary means. The reader is asked to do this in Problem 12.1. The proof of the "if" part is more advanced and requires the technique of moment-generating functions (see Problem 14.15 in Chapter 14). To conclude that (X, Y) has a bivariate normal distribution it is not sufficient that X and Y are normally distributed, but normality of $aX + bY$ should be required for all constants a and b not both equal to zero. A counterexample is as follows. Let the random variable Y be equal to X with probability 0.5 and equal to $-X$ with probability 0.5, where X has a standard normal distribution. Then, the random variable Y also

[†] To be precise, this result requires the following convention: if X is normally distributed and $Y = aX + b$ for constants a and b, then (X, Y) is said to have a bivariate normal distribution. This is a singular bivariate distribution: the probability mass of the two-dimensional vector (X, Y) is concentrated on the one-dimensional line $y = ax + b$. Also, a random variable X with $P(X = \mu) = 1$ for a constant μ is said to have a degenerate $N(\mu, 0)$ distribution with its mass concentrated at a single point.

has a standard normal distribution. It is readily verified that $\text{cov}(X, Y) = 0$. This would imply that X and Y are independent if (X, Y) has a bivariate normal distribution. However, X and Y are obviously dependent, showing that (X, Y) does not have a bivariate normal distribution.

Problem 12.1 Prove that $aX + bY$ is normally distributed for any constants a and b if (X, Y) has a bivariate normal distribution. How do you calculate $P(X > Y)$?

Problem 12.2 The rates of return on two stocks A and B have a bivariate normal distribution with parameters $\mu_1 = 0.08$, $\mu_2 = 0.12$, $\sigma_1 = 0.05$, $\sigma_2 = 0.15$, and $\rho = -0.50$. What is the probability that the average of the returns on stocks A and B will be larger than 0.11?

Problem 12.3 Suppose that the probability density function $f(x, y)$ of the random variables X and Y is given by the bivariate standard normal density with parameter ρ. Verify that the probability density function $f_Z(z)$ of the ratio $Z = X/Y$ is given by the so-called Cauchy density

$$\int_{-\infty}^{\infty} |y| f(zy, y)\, dy = \frac{(1/\pi)\sqrt{1-\rho^2}}{(z-\rho)^2 + 1 - \rho^2}, \qquad -\infty < z < \infty.$$

Problem 12.4 Use the decomposition of the standard bivariate normal density to verify that $P(X \leq a, Y \leq b)$ can be calculated as

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(a-\mu_1)/\sigma_1} \Phi\left(\frac{(b-\mu_2)/\sigma_2 - \rho x}{\sqrt{1-\rho^2}}\right) e^{-\frac{1}{2}x^2}\, dx$$

if the random vector (X, Y) is bivariate normal distributed with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Here, $\Phi(x)$ is the standard normal distribution function.

As stated before, a fundamental result is that (X, Y) has a bivariate normal distribution if and only if $aX + bY$ has a univariate normal distribution for all constants a and b. Use this result to solve the following problems.

Problem 12.5 Let (X, Y) have a bivariate normal distribution. Define the random variables V and W by $V = a_1X + b_1Y + c_1$ and $W = a_2X + b_2Y + c_2$, where a_i, b_i and c_i are constants for $i = 1, 2$. Argue that (V, W) has a bivariate normal distribution.

Problem 12.6 The random variables Z_1 and Z_2 are independent and $N(0, 1)$ distributed. Define the random variables X_1 and X_2 by $X_1 = \mu_1 + \sigma_1 Z_1$ and $X_2 = \mu_2 + \sigma_2\rho Z_1 + \sigma_2\sqrt{1-\rho^2}Z_2$, where μ_1, μ_2, σ_1, σ_2 and ρ are constants with $\sigma_1 > 0$, $\sigma_2 > 0$ and $-1 < \rho < 1$. Prove that (X_1, X_2) has a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

Problem 12.7 Let (X, Y) have a bivariate normal distribution with $\sigma^2(X) = \sigma^2(Y)$. Prove that the random variables $X + Y$ and $X - Y$ are independent and normally distributed.

12.1.1 The drunkard's walk

The drunkard's walk is one of the most useful probability models in the physical sciences. Let us formulate this model in terms of a particle moving on the two-dimensional plane. The particle starts at the origin $(0, 0)$. In each step, the particle travels a unit distance in a randomly chosen direction between 0 and 2π. The direction of each successive step is determined independently of the others. What is the joint probability density function of the (x, y) coordinates of the position of the particle after n steps?

Let the random variable Θ denote the direction taken by the particle in any step. In each step the x-coordinate of the position of the particle changes with an amount that is distributed as $\cos(\Theta)$ and the y-coordinate with an amount that is distributed as $\sin(\Theta)$. The continuous random variable Θ has a uniform distribution on $(0, 2\pi)$. Let X_k and Y_k be the changes of the x-coordinate and the y-coordinate of the position of the particle in the kth step. Then the position of the particle after n steps can be represented by the random vector (S_{n1}, S_{n2}), where

$$S_{n1} = X_1 + \cdots + X_n \qquad \text{and} \qquad S_{n2} = Y_1 + \cdots + Y_n.$$

For each n the random vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and have the same distribution. The reader who is familiar with the central limit theorem from Chapter 5 for the sum of one-dimensional random variables will not be surprised to learn that the random vector $(S_{n1}, S_{n2}) = (X_1 + \cdots + X_n, Y_1 + \cdots + Y_n)$ satisfies the conditions of the two-dimensional version of the central limit theorem. In general form, the two-dimensional version of the central limit theorem reads as

$$\lim_{n\to\infty} P\left(\frac{S_{n1} - n\mu_1}{\sigma_1\sqrt{n}} \le x, \frac{S_{n2} - n\mu_2}{\sigma_2\sqrt{n}} \le y\right)$$
$$= \frac{1}{2\pi\sqrt{(1-\rho^2)}} \int_{-\infty}^{x}\int_{-\infty}^{y} e^{-\frac{1}{2}(v^2 - 2\rho v w + w^2)/(1-\rho^2)}\, dv\, dw,$$

where $\mu_1 = E(X_i), \mu_2 = E(Y_i), \sigma_1^2 = \sigma^2(X_i), \sigma_2^2 = \sigma^2(Y_i)$, and $\rho = \rho(X, Y)$. In the particular case of the drunkard's walk, we have

$$\mu_1 = \mu_2 = 0, \ \sigma_1 = \sigma_2 = \frac{1}{\sqrt{2}} \ \text{and} \ \rho = 0.$$

The derivation of this result is simple and instructive. The random variable Θ has the uniform density function $f(\theta) = \frac{1}{2\pi}$ for $0 < \theta < 2\pi$. Applying the substitution rule gives

$$\mu_1 = E[\cos(\Theta)] = \int_0^{2\pi} \cos(\theta) f(\theta)\, d\theta = \frac{1}{2\pi} \int_0^{2\pi} \cos(\theta)\, d\theta = 0.$$

In the same way, $\mu_2 = 0$. Using the formula $\sigma^2(X) = E(X^2) - [E(X)]^2$ with $X = \cos(\Theta)$, we find

$$\sigma_1^2 = E[\cos^2(\Theta)] = \int_0^{2\pi} \cos^2(\theta) f(\theta)\, d\theta = \frac{1}{2\pi} \int_0^{2\pi} \cos^2(\theta)\, d\theta.$$

In the same way, $\sigma_2^2 = \frac{1}{2\pi} \int_0^{2\pi} \sin^2(\theta)\, d\theta$. Invoking the celebrated formula $\cos^2(\theta) + \sin^2(\theta) = 1$ from goniometry, we obtain $\sigma_1^2 + \sigma_2^2 = 1$. Hence, for reasons of symmetry, $\sigma_1^2 = \sigma_2^2 = \frac{1}{2}$. Finally

$$\text{cov}(X_1, Y_1) = E\left[(\cos(\Theta) - 0)\,(\sin(\Theta) - 0)\right] = \frac{1}{2\pi} \int_0^{2\pi} \cos(\theta) \sin(\theta)\, d\theta.$$

This integral is equal to zero since $\cos(x + \frac{\pi}{2}) \sin(x + \frac{\pi}{2}) = -\cos(x) \sin(x)$ for each of the ranges $0 \leq x \leq \frac{\pi}{2}$ and $\pi \leq x \leq \frac{3\pi}{2}$. This verifies that $\rho = 0$.

Next we can formulate two interesting results using the two-dimensional central limit theorem. The first result states that

$$P\left(S_{n1} \leq x, S_{n2} \leq y\right) \approx \frac{1}{\pi n} \int_{-\infty}^{x} \int_{-\infty}^{y} e^{-(t^2+u^2)/n}\, dt\, du$$

for n large. In other words, the position of the particle after n steps has approximately the bivariate normal density function

$$\phi_n(x, y) = \frac{1}{\pi n} e^{-(x^2+y^2)/n}$$

when n is large. That is, the probability of finding the particle after n steps in a small rectangle with sides Δa and Δb around the point (a, b) is approximately equal to $\phi_n(a, b)\Delta a \Delta b$ for n large. In Figure 12.1, we display the bivariate normal density function $\phi_n(x, y)$ for $n = 25$. The correlation coefficient of the bivariate normal density $\phi_n(x, y)$ is zero. Hence, in accordance with our intuition, the coordinates of the position of the particle after many steps are practically independent of each other. Moreover, by the decomposition

$$\phi_n(x, y) = \frac{1}{\sqrt{n/2}\sqrt{2\pi}} e^{-\frac{1}{2}x^2/\frac{1}{2}n} \times \frac{1}{\sqrt{n/2}\sqrt{2\pi}} e^{-\frac{1}{2}y^2/\frac{1}{2}n},$$

each of the coordinates of the position of the particle after n steps is approximately $N(0, \frac{1}{2}n)$ distributed for n large.

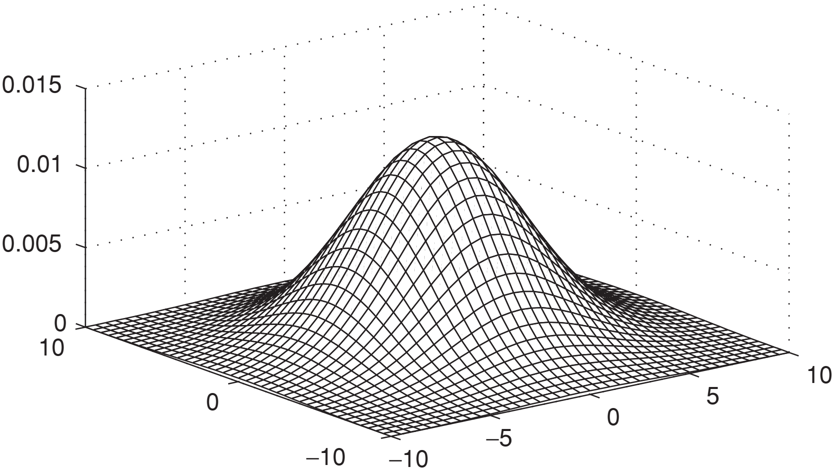


Fig. 12.1. The density of the particle's position after 25 steps.

The second result states that

$$E(D_n) \approx \frac{1}{2}\sqrt{\pi n}$$

for n large, where the random variable D_n is defined by

$$D_n = \text{the distance from the origin to the position of the particle after } n \text{ steps.}$$

The proof of these results goes as follows. Rewrite $P(S_{n1} \leq x, S_{n2} \leq y)$ as

$$P\left(\frac{S_{n1} - n\mu_1}{\sigma_1\sqrt{n}} \leq \frac{x - n\mu_1}{\sigma_1\sqrt{n}}, \frac{S_{n2} - n\mu_2}{\sigma_2\sqrt{n}} \leq \frac{y - n\mu_2}{\sigma_2\sqrt{n}}\right).$$

Substituting the values of μ_1, μ_2, σ_1, σ_2, and ρ, it next follows from the two-dimensional central limit theorem that

$$P(S_{n1} \leq x, S_{n2} \leq y) \approx \frac{1}{2\pi} \int_{-\infty}^{x/\sqrt{n/2}} \int_{-\infty}^{y/\sqrt{n/2}} e^{-\frac{1}{2}(v^2+w^2)}\, dv\, dw$$

for n large. By the change of variables $t = v\sqrt{n/2}$ and $u = w\sqrt{n/2}$, the first result is obtained. To find the approximation formula for $E(D_n)$, note that

$$D_n = \sqrt{S_{n1}^2 + S_{n2}^2}.$$

An application of Rule 11.2 yields that

$$E(D_n) \approx \frac{1}{\pi n} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sqrt{x^2 + y^2} e^{-(x^2+y^2)/n}\, dx\, dy$$

for n large. To evaluate this integral, we use several results from advanced calculus. By a change to polar coordinates $x = r\cos(\theta)$ and $y = r\sin(\theta)$ with $dxdy = rdrd\theta$ and using the identity $\cos^2(\theta) + \sin^2(\theta) = 1$, we find

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sqrt{x^2+y^2}e^{-(x^2+y^2)/n}\,dx\,dy \\
&= \int_0^{\infty}\int_0^{2\pi} \sqrt{r^2\cos^2(\theta)+r^2\sin^2(\theta)}e^{-(r^2\cos^2(\theta)+r^2\sin^2(\theta))/n} r\,dr\,d\theta \\
&= \int_0^{\infty}\int_0^{2\pi} r^2e^{-r^2/n}\,dr\,d\theta = 2\pi\int_0^{\infty} r^2e^{-r^2/n}\,dr.
\end{aligned}$$

Obviously

$$\begin{aligned}
\int_0^{\infty} r^2e^{-r^2/n}\,dr &= -\frac{n}{2}\int_0^{\infty} rde^{-r^2/n} = -\frac{n}{2}re^{-r^2/n}\Big|_0^{\infty} + \frac{n}{2}\int_0^{\infty} e^{-r^2/n}\,dr \\
&= \frac{1}{2}\frac{n}{2}\sqrt{n/2}\sqrt{2\pi}\int_{-\infty}^{\infty}\frac{1}{\sqrt{n/2}\sqrt{2\pi}}e^{-\frac{1}{2}r^2/(\frac{1}{2}n)}\,dr = \frac{1}{4}n\sqrt{n\pi},
\end{aligned}$$

using the fact that the $N(0, \frac{1}{2}n)$ density integrates to 1 over $(\infty, -\infty)$. Putting the pieces together, we get $E(D_n) \approx \frac{1}{2}\sqrt{\pi n}$. This is an excellent approximation for $n = 10$ onwards. Using the relation

$$P(D_n \le u) \approx \iint_{C(u)} \frac{1}{\pi n}e^{-(x^2+y^2)/n}\,dx\,dy \qquad \text{for } n \text{ large}$$

with $C(u) = \{(x, y) : \sqrt{x^2+y^2} \le u\}$, a slight modification of the above analysis shows that

$$P(D_n \le u) \approx \frac{2}{n}\int_0^u re^{-r^2/n}\,dr \qquad \text{for } u > 0.$$

Hence, $P(D_n \le u) \approx 1 - e^{-u^2/n}$ for n large and so the approximate probability density of D_n is $\frac{2u}{n}e^{-u^2/n}$ for $u > 0$. This is the Rayleigh density.

Problem 12.8 The rectangular distance from the origin to the position of the particle after n steps is defined by $R_n = |S_{n1}| + |S_{n2}|$. Verify that $E(R_n) \approx \sqrt{\frac{4n}{\pi}}$ and R_n has the approximate density $\frac{4}{\sqrt{2\pi n}}e^{-\frac{1}{4}r^2/n}[\Phi(\frac{r}{\sqrt{n}}) - \Phi(\frac{-r}{\sqrt{n}})]$ for n large, where $\Phi(x)$ is the standard normal distribution function.

Problem 12.9 Two particles carry out a drunkard's walk on the two-dimensional plane, independently of each other. Both particles start at the origin (0, 0). One particle performs n steps and the other m steps. Can you give an intuitive explanation why the expected distance between the final positions of the two particles is equal to $\frac{1}{2}\sqrt{\pi}\sqrt{n+m}$?

12.1.2 Drunkard's walk in dimension three or higher

When the drunkard's walk occurs in three-dimensional space, it can be shown that the joint probability density function of the (x, y, z) coordinates of the position of the particle after n steps is approximately given by the trivariate normal probability density function

$$\frac{1}{(2\pi n/3)^{3/2}} e^{-\frac{3}{2}(x^2+y^2+z^2)/n}$$

for n large. Thus, for n large, the coordinates of the particle after n steps are practically independent of each other and are each approximately $N(0, \frac{1}{3}n)$ distributed. The same result holds for the drunkard's walk in dimension d. Each of the coordinates of the particle after n steps then has an approximate $N(0, \frac{1}{d}n)$ distribution. Also, for the drunkard's walk in dimension d, the following result can be given for the probability distribution of the distance D_n between the origin and the position of the particle after n steps with n large

$$P(D_n \leq u) \approx \int_0^{u/\sqrt{n}} \frac{d^{\frac{1}{2}d}}{2^{\frac{1}{2}d-1}\Gamma\left(\frac{1}{2}d\right)} e^{-\frac{1}{2}dr^2} r^{d-1}\, dr \qquad \text{for } u > 0,$$

where $\Gamma(a)$ is the gamma function. The probability distribution of D_n is related to the chi-square distribution with d degrees of freedom (see Section 10.2.9). It is matter of some algebra to derive from the approximate density of D_n that

$$E(D_n) \approx \frac{\alpha_d}{d^{\frac{1}{2}}2^{\frac{1}{2}d-1}\Gamma\left(\frac{1}{2}d\right)}\sqrt{n},$$

where $\alpha_m = \int_0^\infty x^m e^{-\frac{1}{2}x^2}\, dx$. Using partial integration, it is not difficult to verify that the α_m can be recursively computed from

$$\alpha_m = (m-1)\alpha_{m-2} \qquad \text{for } m = 2, 3, \ldots$$

with $\alpha_0 = \sqrt{\frac{1}{2}\pi}$ and $\alpha_1 = 1$. In particular, using the fact that $\Gamma(\frac{3}{2}) = \frac{1}{2}\sqrt{\pi}$, we find for the three-dimensional space that $E(D_n) \approx \sqrt{\frac{8n}{3\pi}}$ for n large. An application of this formula in physics can be found in Section 2.4.

12.2 Multivariate normal distribution

The multivariate normal distribution is a very useful probability model to describe dependencies between two or more random variables. In finance, the multivariate normal distribution is frequently used to model the joint distribution of the returns in a portfolio of assets.

First we give a general definition of the multivariate normal distribution.

Definition 12.1 *A d-dimensional random vector $(S_1, S_2, \ldots, S_d)$ is said to be multivariate normal distributed if for any d-tuple of real numbers $\alpha_1, \ldots, \alpha_d$ the one-dimensional random variable*

$$\alpha_1 S_1 + \alpha_2 S_2 + \cdots + \alpha_d S_d$$

has a (univariate) normal distribution.

Recall the convention that a degenerate random variable X with $P(X = \mu) = 1$ for a constant μ is considered as an $N(\mu, 0)$-distributed random variable. Definition 12.1 implies that each of the individual random variables $S_1, \ldots, S_d$ is normally distributed. Let us define the vector $\mu = (\mu_i),\ i = 1, \ldots, d$ of expected values and the matrix $\mathbf{C} = (\sigma_{ij}),\ i,\ j = 1, \ldots, d$ of covariances by

$$\mu_i = E(X_i) \quad \text{and} \quad \sigma_{ij} = \text{cov}(X_i, X_j).$$

Note that $\sigma_{ii} = \text{var}(X_i)$. The multivariate normal distribution is called *nonsingular* if the determinant of the covariance matrix $\mathbf{C}$ is nonzero; otherwise, the distribution is called singular. By a basic result from linear algebra, a singular covariance matrix $\mathbf{C}$ means that the probability mass of the multivariate normal distribution is concentrated on a subspace with a dimension lower than d. In applications, the covariance matrix of the multivariate normal distribution is often singular. In the example of the drunkard's walk on the two-dimensional plane, however, the approximate multivariate normal distribution of the position of the particle after n steps has the nonsingular covariance matrix

$$\begin{pmatrix} \frac{1}{2}n & 0 \\ 0 & \frac{1}{2}n \end{pmatrix}.$$

A very useful result for practical applications is the fact that the multivariate normal distribution is *uniquely determined* by the vector of expected values and the covariance matrix. Note that the covariance matric C is symmetric and positive semi-definite (see also Problem 11.16).

Further study of the multivariate normal distribution requires matrix analysis and advanced methods in probability theory such as the theory of the so-called characteristic functions. Linear algebra is indispensable for multivariate analysis in probability and statistics. The following important result for the multivariate normal distribution is stated without proof: the random variables $S_1, \ldots, S_d$ can be expressed as linear combinations of independent standard normal random variables. That is

$$S_i = \mu_i + \sum_{j=1}^{d} a_{ij} Z_j \qquad \text{for } i = 1, \ldots, d,$$

where $Z_1, \ldots, Z_n$ are independent random variables each having the standard normal distribution. The matrix $\mathbf{A} = (a_{ij})$ satisfies $\mathbf{C} = \mathbf{A}\mathbf{A}^T$ with $\mathbf{A}^T$ denoting the transpose of the matrix $\mathbf{A}$ (you may directly verify this result by writing out $\text{cov}(S_i, S_j)$ from the decomposition formula for the S_i). Moreover, using the fact that the covariance matrix C is symmetric and positive semi-definite, it follows from a basic diagonalization result in linear algebra that the matrix A can be computed from

$$\mathbf{A} = \mathbf{U}\mathbf{D}^{1/2},$$

where the matrix $\mathbf{D}^{1/2}$ is a diagonal matrix with the square roots of the eigenvalues of the covariance matrix $\mathbf{C}$ on its diagonal (these eigenvalues are real and nonnegative). The orthogonal matrix $\mathbf{U}$ has the normalized eigenvectors of the matrix $\mathbf{C}$ as column vectors (Cholesky decomposition is a convenient method to compute the matrix $\mathbf{A}$ when $\mathbf{C}$ is nonsingular). The decomposition result for the vector $(S_1, \ldots, S_d)$ is particularly useful when simulating random observations from the multivariate normal distribution. In Section 11.3, we explained how to simulate from the one-dimensional standard normal distribution.

Remark 12.1 The result that the S_i are distributed as $\mu_i + \sum_{j=1}^{d} a_{ij} Z_j$ has a useful corollary. By taking the inproduct of the vector $(S_1 - \mu_1, \ldots, S_d - \mu_d)$ with itself, it is a matter of basic linear algebra to prove that $\sum_{j=1}^{d}(S_j - \mu_j)^2$ is distributed as $\sum_{j=1}^{d} \lambda_j {Z_j}^2$. This is a useful result for establishing the chi-square test in Section 12.4. If the eigenvalues λ_k of the covariance matrix $\mathbf{C}$ are 0 or 1, then the random variable $\sum_{j=1}^{d} \lambda_j {Z_j}^2$ has a chi-square distribution. These matters are quite technical but are intended to give you better insight into the chi-square test that will be discussed in Section 12.4.

Remark 12.2 If the covariance matrix $\mathbf{C}$ of the multivariate normal distribution is nonsingular, it is possible to give an explicit expression for the corresponding multivariate probability density function. To do so, let us define the matrix $\mathbf{Q} = (q_{ij})$ by

$$q_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \qquad \text{for } i, j = 1, \ldots, d,$$

where σ_ℓ is a shorthand for $\sqrt{\sigma_{\ell\ell}}$. Denote by γ_{ij} the (i, j)th element of the inverse matrix $\mathbf{Q}^{-1}$ and let the polynomial $Q(x_1, \ldots, x_d)$ denote

$$Q(x_1, \ldots, x_d) = \sum_{i=1}^{d} \sum_{j=1}^{d} \gamma_{ij} x_i x_j.$$

Then, the standardized vector $\left(\frac{S_1-\mu_1}{\sigma_1}, \ldots, \frac{S_d-\mu_d}{\sigma_d}\right)$ can be shown to have the standard multivariate normal probability density function

$$\frac{1}{(2\pi)^{d/2}\sqrt{\det(\mathbf{Q})}} e^{-\frac{1}{2}Q(x_1,\ldots,x_d)}.$$

This multidimensional density function reduces to the standard bivariate normal probability density function from Section 12.1 when $d = 2$.

12.3 Multidimensional central limit theorem

The central limit theorem is the queen of all theorems in probability theory. The one-dimensional version is extensively discussed in Chapter 5. The analysis of the drunkard's walk on the two-dimensional plane used the two-dimensional version. The multidimensional version of the central limit theorem is as follows. Suppose that

$$\mathbf{X}_1 = (X_{11}, \ldots, X_{1d}),\ \mathbf{X}_2 = (X_{21}, \ldots, X_{2d}), \ldots,\ \mathbf{X}_n = (X_{n1}, \ldots, X_{nd})$$

are independent random vectors of dimension d. The random vector $\mathbf{X}_k$ has the one-dimensional random variable X_{kj} as its jth component. The random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are said to be independent if

$$P(\mathbf{X}_1 \in A_1, \ldots, \mathbf{X}_n \in A_n) = P(\mathbf{X}_1 \in A_1) \cdots P(\mathbf{X}_n \in A_n)$$

for any subsets $A_1, \ldots, A_n$ of the d-dimensional Euclidean space. Note that, for fixed k, the random variables $X_{k1}, \ldots, X_{kd}$ need not be independent. Also assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ have the same individual distributions, that is, $P(\mathbf{X}_1 \in A) = \ldots = P(\mathbf{X}_n \in A)$ for any subset A of the d-dimensional space. Under this assumption, let

$$\mu_j^{(0)} = E(X_{1j}) \quad \text{and} \quad \sigma_{ij}^{(0)} = \text{cov}(X_{1i}, X_{1j})$$

for $i, j = 1, \ldots, d$, assuming that the expectations exist. For $j = 1, \ldots, d$, we now define the random variable S_{nj} by

$$S_{nj} = X_{1j} + X_{2j} + \cdots + X_{nj}.$$

Multidimensional central limit theorem *For n large, the random vector* $\mathbf{S}_n = (S_{n1}, S_{n2}, \ldots, S_{nd})$ *has approximately a multivariate normal distribution. The vector* μ *of expected values and the covariance matrix* $\mathbf{C}$ *are given by*

$$\mu = \left(n\mu_1^{(0)}, \ldots, n\mu_d^{(0)}\right) \quad \textit{and} \quad \mathbf{C} = \left(n\sigma_{ij}^{(0)}\right)$$

when the random vectors $\mathbf{X}_k$ *are identically distributed.*

In the next section we discuss two applications of the multidimensional central limit theorem. In the first application, we will use the fact that the assumption of identically distributed random vectors $\mathbf{X}_k$ may be weakened in the multidimensional central limit theorem.

Problem 12.10 The annual rates of return on the three stocks A, B, and C have a trivariate normal distribution. The rate of return on stock A has expected value 7.5% and standard deviation 7%, the rate of return on stock B has expected value 10% and standard deviation 12%, and the rate of return on stock C has expected value 20% and standard deviation 18%. The correlation coefficient of the rates of return on stocks A and B is 0.7, the correlation coefficient is -0.5 for the stocks A and C, and the correlation coefficient is -0.3 for the stocks B and C. An investor has \$100,000 in cash. Any cash that is not invested in the three stocks will be put in a riskless asset that offers an annual interest rate of 5%.

(a) Suppose the investor puts \$20,000 in stock A, \$20,000 in stock B, \$40,000 in stock C, and \$20,000 in the riskless asset. What are the expected value and the standard deviation of the portfolio's value next year?

(b) Can you find a portfolio whose risk is smaller than the risk of the portfolio from Question (a) but whose expected return is not less than that of the portfolio from Question (a)?

(c) For the investment plan from Question (a), find the probability that the portfolio's value next year will be less than \$112,500 and the probability that the portfolio's value next year will be more than \$125,000.

Problem 12.11 The random vector (X_1, X_2, X_3) has a trivariate normal distribution. What is the joint distribution of X_1 and X_2?

12.3.1 Predicting election results

The multivariate normal distribution is also applicable to the problem of predicting election results. In Section 3.6, we discuss a polling method whereby a respondent is not asked to choose a favorite party, but instead is asked to indicate how likely the respondent is to vote for each party. Consider the situation in which there are three parties A, B, and C and n representative voters are interviewed. A probability distribution (p_{iA}, p_{iB}, p_{iC}) with $p_{iA} + p_{iB} + p_{iC} = 1$ describes the voting behavior of respondent i for $i = 1, \ldots, n$. That is, p_{iP} is the probability that respondent i will vote for party P on election day. Let the random variable S_{nA} be the number of respondents of the n interviewed voters who actually vote for party A on election day. The random variables S_{nB} and S_{nC} are defined in a similar manner. The vector $\mathbf{S}_n = (S_{nA}, S_{nB}, S_{nC})$ can be written as the sum of n random vectors $\mathbf{X}_1 = (X_{1A}, X_{1B}, X_{1C}), \ldots, \mathbf{X}_n = (X_{nA}, X_{nB}, X_{nC})$, where the random variable X_{iP} is defined by

$$X_{iP} = \begin{cases} 1 & \text{if respondent } i \text{ votes for party } P \\ 0 & \text{otherwise.} \end{cases}$$

The random vector $\mathbf{X}_i = (X_{iA}, X_{iB}, X_{iC})$ describes the voting behavior of respondent i. The simplifying assumption is made that the random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent. These random vectors do not have the same individual distributions. However, under the crucial assumption of independence, the multidimensional central limit theorem can be shown to remain valid and thus the random vector (S_{nA}, S_{nB}, S_{nC}) has approximately a multivariate normal distribution for n large. This multivariate normal distribution has

$$\mu = \left(\sum_{i=1}^{n} p_{iA}, \sum_{i=1}^{n} p_{iB}, \sum_{i=1}^{n} p_{iC} \right)$$

as vector of expected values and

$$\mathbf{C} = \begin{pmatrix} \sum_{i=1}^{n} p_{iA}(1-p_{iA}) & -\sum_{i=1}^{n} p_{iA}p_{iB} & -\sum_{i=1}^{n} p_{iA}p_{iC} \\ -\sum_{i=1}^{n} p_{iA}p_{iB} & \sum_{i=1}^{n} p_{iB}(1-p_{iB}) & -\sum_{i=1}^{n} p_{iB}p_{iC} \\ -\sum_{i=1}^{n} p_{iA}p_{iC} & -\sum_{i=1}^{n} p_{iB}p_{iC} & \sum_{i=1}^{n} p_{iC}(1-p_{iC}) \end{pmatrix}$$

as covariance matrix (this matrix is singular, since for each row the sum of the elements is zero). The result for the vector μ of expected values is obvious, but a few words of explanation are in order for the covariance matrix $\mathbf{C}$. By the independence of $X_{1A}, \ldots, X_{nA}$

$$\sigma^2(S_{nA}) = \sigma^2\left(\sum_{i=1}^{n} X_{iA} \right) = \sum_{i=1}^{n} \sigma^2(X_{iA}) = \sum_{i=1}^{n} p_{iA}(1-p_{iA}).$$

Also, by the independence of X_{iA} and X_{jB} for $j \neq i$ and since X_{iA} and X_{iB} cannot both be positive, it follows that $\text{cov}(S_{nA}, S_{nB})$ is given by

$$\begin{aligned} & E\left[\left(\sum_{i=1}^{n} X_{iA} \right) \left(\sum_{j=1}^{n} X_{jB} \right) \right] - E\left(\sum_{i=1}^{n} X_{iA} \right) E\left(\sum_{j=1}^{n} X_{jB} \right) \\ & = \sum_{i=1}^{n} \sum_{j \neq i} E(X_{iA}X_{jB}) - \left(\sum_{i=1}^{n} p_{iA} \right) \left(\sum_{j=1}^{n} p_{jB} \right) \\ & = \sum_{i=1}^{n} \sum_{j \neq i} p_{iA}p_{jB} - \sum_{i=1}^{n} \sum_{j=1}^{n} p_{iA}p_{jB} = -\sum_{i=1}^{n} p_{iA}p_{iB}. \end{aligned}$$

Similarly, the other terms in matrix $\mathbf{C}$ are explained.

Table 12.1. *Voting probabilities.*

No. of voters	$(p_{iA},\ p_{iB},\ p_{iC})$
230	(0.20, 0.80, 0)
140	(0.65, 0.35, 0)
60	(0.70, 0.30, 0)
120	(0.45, 0.55, 0)
70	(0.90, 0.10, 0)
40	(0.95, 0, 0.05)
130	(0.60, 0.35, 0.05)
210	(0.20, 0.55, 0.25)

It is standard fare in statistics to simulate random observations from the multivariate normal distribution. This means that computer simulation provides a fast and convenient tool to estimate probabilities of interest such as the probability that party A will receive the most votes or the probability that the two parties A and C will receive more than half of the votes.

Numerical illustration

Suppose that a representative group of $n = 1{,}000$ voters is polled. The probabilities assigned by each of the 1,000 voters to parties A, B, and C are summarized in Table 12.1: the vote of each of 230 persons will go to parties A, B, and C with probabilities 0.80, 0.20, and 0, the vote of each of 140 persons will go to parties A, B, and C with probabilities 0.65, 0.35, and 0, and so on. Each person votes independently. Let the random variable S_A be defined as

$S_A =$ the number of votes on party A when the 1,000 voters
actually vote on election day.

Similarly, the random variables S_B and S_C are defined. How do we calculate probabilities such as the probability that party A will become the largest party and the probability that parties A and C together will get the majority of the votes? These probabilities are given by $P(S_A > S_B,\ S_A > S_C)$ and $P(S_A + S_C > S_B)$. Simulating from the trivariate normal approximation for the random vector (S_A, S_B, S_C) provides a simple and fast method to get approximate values for these probabilities. The random vector (S_A, S_B, S_C) has approximately a trivariate normal distribution. Using the data from Table 12.1, the vector of expected values and the covariance matrix of this trivariate normal distribution are estimated by

$$\mu = (454, 485, 61) \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} 183.95 & -167.65 & -16.30 \\ -167.65 & 198.80 & -31.15 \\ -16.30 & -31.15 & 47.45 \end{pmatrix}.$$

In order to simulate random observations from this trivariate normal distribution, the eigenvalues $\lambda_1,\ \lambda_2,\ \lambda_3$ and the corresponding normalized eigenvectors $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$ of matrix $\mathbf{C}$ must be first calculated. Using standard software, we find

$$\lambda_1 = 70.6016,\ \mathbf{u}_1 = \begin{pmatrix} 0.4393 \\ 0.3763 \\ -0.8157 \end{pmatrix},\ \lambda_2 = 359.5984,\ \mathbf{u}_2 = \begin{pmatrix} -0.6882 \\ 0.7246 \\ -0.0364 \end{pmatrix},$$

$$\lambda_3 = 0,\ \mathbf{u}_3 = \begin{pmatrix} 0.5774 \\ 0.5774 \\ 0.5774 \end{pmatrix}.$$

The diagonal matrix $\mathbf{D}^{1/2}$ has $\sqrt{\lambda_1}$, $\sqrt{\lambda_2}$, and $\sqrt{\lambda_3}$ on its diagonal and the orthogonal matrix $\mathbf{U}$ has $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$ as column vectors. The matrix product $\mathbf{UD}^{1/2}$ gives the desired decomposition matrix

$$\mathbf{A} = \begin{pmatrix} 3.6916 & -13.0508 & 0 \\ 3.1622 & 13.7405 & 0 \\ -6.8538 & -0.6897 & 0 \end{pmatrix}.$$

Thus, the random vector (S_A, S_B, S_C) can approximately be represented as

$$\begin{aligned} S_A &= 454 + 3.6916Z_1 - 13.0508Z_2 \\ S_B &= 485 + 3.1622Z_1 + 13.7405Z_2 \\ S_C &= 61 - 6.8538Z_1 - 0.6897Z_2, \end{aligned}$$

where Z_1 and Z_2 are independent random variables each having the standard normal distribution. Note that the condition $S_A + S_B + S_C = 1,000$ is preserved in this decomposition. Using the decomposition, it is standard fare to simulate random observations from the trivariate normal approximation to (S_A, S_B, S_C). A simulation study with 100,000 random observations leads to the following estimates

$P(\text{party } A \text{ becomes the largest party}) = 0.123\ (\pm 0.002)$
$P(\text{party } B \text{ becomes the largest party}) = 0.877\ (\pm 0.002)$
$P(\text{parties } A \text{ and } C \text{ get the majority of the votes}) = 0.855\ (\pm 0.002)$,

where the numbers between the parentheses indicate the 95% confidence intervals. How accurate is the model underlying these predictions? They are based on an approximately multivariate normal distribution. To find out how well this approximative model works, we use the bootstrap method to simulate directly from the data in Table 12.1 (see Section 3.6 for more details on this powerful method). Performing 100,000 simulation runs, we obtain the values 0.120 (± 0.002), 0.872 (± 0.002), and 0.851 (± 0.002) for the three probabilities above. The approximative values of the multivariate normal model are very close to the exact values of the bootstrap method. This justifies the use of this model which is computationally less demanding than the bootstrap method.

12.3.2 Lotto r/s

In the Lotto 6/45, six balls are drawn out of a drum with 45 balls numbered from 1 to 45. More generally, in the Lotto r/s, r balls are drawn from a drum with s balls. For the Lotto r/s, define the random variable S_{nj} by

$$S_{nj} = \text{the number of times ball number } j \text{ is drawn in } n \text{ drawings}$$

for $j = 1, \ldots, s$. Letting

$$X_{kj} = \begin{cases} 1 & \text{if ball number } j \text{ is drawn at the } k\text{th drawing} \\ 0 & \text{otherwise,} \end{cases}$$

we can represent S_{nj} in the form

$$S_{nj} = X_{1j} + X_{2j} + \cdots + X_{nj}.$$

Thus, by the multidimensional central limit theorem, the random vector $\mathbf{S}_n = (S_{n1}, \ldots, S_{ns})$ approximately has a multivariate normal distribution. The quantities $\mu_j^{(0)} = E(X_{1j})$ and $\sigma_{ij}^{(0)} = \text{cov}(X_{1i}, X_{1j})$ are given by

$$\mu_j^{(0)} = \frac{r}{s},\ \sigma_{jj}^{(0)} = \frac{r}{s}\left(1 - \frac{r}{s}\right) \quad \text{and} \quad \sigma_{ij}^{(0)} = -\frac{r(s-r)}{s^2(s-1)} \qquad \text{for } i \neq j.$$

It is left to the reader to verify this with the help of

$$P(X_{1j} = 1) = \frac{r}{s},\ P(X_{1j} = 0) = 1 - \frac{r}{s} \qquad \text{for all } j$$

and

$$P(X_{1i} = 1, X_{1j} = 1) = \frac{r}{s} \times \frac{r-1}{s-1} \qquad \text{for all } i, j \text{ with } i \neq j.$$

The covariance matrix $\mathbf{C} = (n\sigma_{ij}^{(0)})$ is singular. The reason is that the sum of the elements of each row is zero. The matrix $\mathbf{C}$ has rank $s - 1$.

For the lotto, an interesting random walk is the stochastic process that describes how the random variable $\max_{1 \le j \le s} S_{nj} - \min_{1 \le j \le s} S_{nj}$ behaves as a function of n. This random variable gives the difference between the number of occurrences of the most-drawn-ball number and that of the least-drawn-ball number in the first n drawings. Simulation experiments reveal that the sample paths of the random walk exhibit the tendency to increase proportionally to $\sqrt{n}$ as n gets larger. The central limit theorem is at work here. In particular, it can be proved that a constant c exists such that

$$E\left[\max_{1\leq j\leq s} S_{nj} - \min_{1\leq j\leq s} S_{nj}\right] \approx c\sqrt{n}$$

for n large. Using computer simulation, we find the value $c = 1.52$ for the Lotto 6/45 and the value $c = 1.48$ for the Lotto 6/49.

Problem 12.12 For the Lotto 6/45, simulate from the multivariate normal distribution in order to find approximately the probability

$$P\left(\max_{1\leq j\leq s} S_{nj} - \min_{1\leq j\leq s} S_{nj} > 1.5\sqrt{n}\right) \qquad \text{for } n = 100, 300, \text{ and } 500.$$

Use computer simulation to find the exact value of this probability.

12.4 The chi-square test

The chi-square (χ^2) test is tailored to measure how well an assumed distribution fits given data when the data are the result of independent repetitions of an experiment with a finite number of possible outcomes. Let's consider an experiment with d possible outcomes $j = 1, \ldots, d$, where the outcome j occurs with probability p_j for $j = 1, \ldots, d$. It is assumed that the probabilities p_j are not estimated from the data but are known. Suppose that n independent repetitions of the experiment are done. Define the random variable X_{kj} by

$$X_{kj} = \begin{cases} 1 & \text{if the outcome of the } k\text{th experiment is } j \\ 0 & \text{otherwise.} \end{cases}$$

Then, the random vectors $\mathbf{X}_1 = (X_{11}, \ldots, X_{1d}), \ldots, \mathbf{X}_n = (X_{n1}, \ldots, X_{nd})$ are independent and identically distributed. Let the random variable N_j represent the number of times that the outcome j appears in the n repetitions of the experiment. That is

$$N_j = X_{1j} + \cdots + X_{nj} \qquad \text{for } j = 1, \ldots, d.$$

A convenient measure of the distance between the random variables N_j and their expected values np_j is the weighted sum of squares

$$\sum_{j=1}^{d} w_j(N_j - np_j)^2$$

for appropriately chosen weights $w_1, \ldots, w_d$. How do we choose the constants w_j? Naturally, we want to make the distribution of the weighted sum of squares as simple as possible. This is achieved by choosing $w_j = (np_j)^{-1}$. For large n, the test statistic

$$D = \sum_{j=1}^{d} \frac{(N_j - np_j)^2}{np_j}$$

has approximately a chi-square distribution with $d - 1$ degrees of freedom (one degree of freedom is lost because of the linear relationship $\sum_{j=1}^{d} N_j = n$). We briefly outline the proof of this very useful result that goes back to Karl Pearson (1857–1936), one of the founders of modern statistics. Using the multidimensional central limit theorem, it can be shown that, for large n, the random vector

$$\left(\frac{N_1 - np_1}{\sqrt{np_1}}, \ldots, \frac{N_d - np_d}{\sqrt{np_d}} \right)$$

has approximately a multivariate normal distribution with the zero vector as its vector of expected values and the matrix $\mathbf{C} = \mathbf{I} - \sqrt{\mathbf{p}}\sqrt{\mathbf{p}}^T$ as its covariance matrix, where $\mathbf{I}$ is the identity matrix, and the column vector $\sqrt{\mathbf{p}}$ has $\sqrt{p_j}$ as its jth component. Using the fact that $\sum_{j=1}^{d} p_j = 1$, the reader familiar with linear algebra may easily verify that one of the eigenvalues of the matrix $\mathbf{C}$ is equal to zero, and all other $d - 1$ eigenvalues are equal to 1. Thus, by appealing to a result stated in Remark 12.1 in Section 12.2, the random variable $\sum_{j=1}^{d} \left(\frac{N_j - np_j}{\sqrt{np_j}} \right)^2$ is approximately distributed as the sum of the squares of $d - 1$ independent $N(0, 1)$ random variables and thus has an approximate chi-square distribution with $d - 1$ degrees of freedom.

To get an idea as to how well the chi-square approximation performs, consider Question 9 from Chapter 1 again. The problem deals with an experiment having the six possible outcomes $1, \ldots, 6$, where the corresponding probabilities are hypothesized to be $\frac{1}{6}, \ldots, \frac{1}{6}$. In 1,200 rolls of a fair die the outcomes 1, 2, 3, 4, 5, and 6 occurred 196, 202, 199, 198, 202, and 203 times. In this case the test statistic D takes on the value

$$\frac{4^2 + 2^2 + 1^2 + 2^2 + 2^2 + 3^2}{200} = 0.19.$$

We immediately notice that the value 0.19 lies far below the expected value 5 of the χ_5^2-distribution. Since the frequencies of the outcomes are very close to the expected values, the suspicion is that the data are fabricated. Therefore, we determine the probability that the test statistic $D = \sum_{j=1}^{6}(N_j - 200)^2/200$ is smaller than or equal to 0.19, where N_j is the number of occurrences of outcome j in 1,200 rolls of a fair die. The chi-square approximation of this probability is equal to $P(\chi_5^2 \leq 0.19) = 0.00078$. This approximation is very close to the simulated value $P(D \leq 0.19) = 0.00083$ obtained from four million simulation runs of 1,200 rolls of a fair die. The very small value of this probability indicates that the data are most likely fabricated. The finding that $P(\chi_5^2 \leq 0.19)$ is an excellent approximation to the exact value of $P(D \leq 0.19)$ confirms the widely used rule of thumb that the chi-square approximation can be applied when $np_j \geq 5$ for all j.

Problem 12.13 In a classical experiment, Gregor Mendel observed the shape and color of peas that resulted from certain crossbreedings. A sample of 556 peas was studied with the result that 315 produced round yellow, 108 produced round green, 101 produced wrinkled yellow, and 32 produced wrinkled green. According to Mendelian theory, the frequencies should be in the ratio 9 : 3 : 3 : 1. What do you conclude from a chi-square test?

Problem 12.14 Use a chi-square test to investigate the quality of the random-number generator on your computer. Generate 10,000 random numbers and count the frequencies of the numbers falling in each of the intervals $\left(\frac{i-1}{10}, \frac{i}{10}\right)$ for $i = 1, \ldots, 10$.

Problem 12.15 In the Dutch Lotto six so-called main numbers and one so-called bonus number are drawn from the numbers $1, \ldots, 45$ and in addition one color is drawn from six differing colors. For each ticket you are asked to mark six distinct numbers and one color. You win the jackpot (first prize) by matching the six main numbers and the color; the second prize by matching the six main numbers, but not the color; the third prize by matching five main numbers, the color, and the bonus number; the fourth prize by matching five main numbers and the bonus number but not the color; the fifth prize by matching five main numbers and the color, but not the bonus number; and the sixth prize by matching only five main numbers. A total of 98,364,597 tickets filled in during a half-year period resulted in 2 winners of the jackpot, 6 winners of the second prize, 9 winners of the third prize, 35 winners of the fourth prize, 411 winners of the fifth prize, and 2,374 winners of the sixth prize. Use a chi-square test to find out whether or not the tickets were randomly filled in.

Problem 12.16 A study of D. Kadell and D. Ylvisaker entitled "Lotto play: the good, the fair and the truly awful," *Chance Magazine* **4** (1991): 22–25 analyzes the behavior of players in the lotto. They took 111,221,666 tickets that were manually filled in for a specific draw of the California Lotto 6/53 and counted how many combinations were filled in exactly k times for $k = 0, 1, \ldots, 20$.

k	N_k	k	N_k
0	288,590	11	217,903
1	1,213,688	12	126,952
2	2,579,112	13	77,409
3	3,702,310	14	50,098
4	4,052,043	15	33,699
5	3,622,666	16	23,779
6	2,768,134	17	17,483
7	1,876,056	18	13,146
8	1,161,423	19	10,158
9	677,368	20	7,969
10	384,186	> 20	53,308

In the table we give the observed values of the N_k, where N_k denotes the number of combinations filled in k times. Use a chi-square test to find out whether the picks chosen by the players are random or not.

13

Conditional distributions

In Chapter 8, conditional probabilities are introduced by conditioning upon the occurrence of an event B of nonzero probability. In applications, this event B is often of the form $Y = b$ for a discrete random variable Y. However, when the random variable Y is continuous, the condition $Y = b$ has probability zero for any number b. The purpose of this chapter is to develop techniques for handling a condition provided by the observed value of a continuous random variable. We will see that the conditional probability density function of X given $Y = b$ for continuous random variables is analogous to the conditional probability mass function of X given $Y = b$ for discrete random variables. The conditional distribution of X given $Y = b$ enables us to define the natural concept of conditional expectation of X given $Y = b$. This concept allows for an intuitive understanding and is of utmost importance. In statistical applications, it is often more convenient to work with conditional expectations instead of the correlation coefficient when measuring the strength of the relationship between two dependent random variables. In applied probability problems, the computation of the expected value of a random variable X is often greatly simplified by conditioning on an appropriately chosen random variable Y. Learning the value of Y provides additional information about the random variable X and for that reason the computation of the conditional expectation of X given $Y = b$ is often simple.

13.1 Conditional probability densities

Suppose that the random variables X and Y are defined on the same sample space Ω with probability measure P. A basic question for dependent random variables X and Y is: if the observed value of Y is y, what distribution now describes the distribution of X? To answer this question, we first consider the case of discrete random variables X and Y with joint probability mass function $p(x, y) = P(X = x, Y = y)$. The *conditional probability mass function* of X given that $Y = b$ is denoted and defined by

$$P\,(X = x \mid Y = b) = \frac{P(X = x, Y = b)}{P(Y = b)}$$

for any fixed b with $P(Y = b) > 0$. This definition is just $P(A \mid B) = \frac{P(AB)}{P(B)}$ written in terms of random variables, where $A = \{\omega : X(\omega) = x\}$ and $B = \{\omega : Y(\omega) = b\}$ with ω denoting an element of the sample space. The notation $p_X(x \mid b)$ is often used for the conditional mass function $P(X = x \mid Y = b)$. Writing

$$P(X = a, Y = b) = P(X = a \mid Y = b)P(Y = b)$$

and using the fact that $\sum_b P(X = a, Y = b) = P(X = a)$, we have the useful relation

$$P(X = a) = \sum_b P(X = a \mid Y = b)P(Y = b),$$

in agreement with the law of conditional probabilities from Section 8.1.3.

Example 13.1 Two fair dice are rolled. Let the random variable X represent the smallest of the outcomes of the two rolls, and let Y represent the sum of the outcomes of the two rolls. What are the conditional probability mass functions of X and Y?

Solution. The joint probability mass function $p(x, y) = P(X = x, Y = y)$ of X and Y is given in Table 11.1. The conditional mass functions follow directly from this table. For example, the conditional mass function $p_X(x \mid 7) = P(X = x \mid Y = 7)$ is given by

$$p_X(1 \mid 7) = \frac{2/36}{6/36} = \frac{1}{3},\ p_X(2 \mid 7) = \frac{2/36}{6/36} = \frac{1}{3},\ p_X(3 \mid 7) = \frac{2/36}{6/36} = \frac{1}{3},$$
$$p_X(x \mid 7) = 0 \qquad \text{for } x = 4, 5, 6.$$

This conditional distribution is a discrete uniform distribution on $\{1, 2, 3\}$. We also give the conditional mass function $p_Y(y \mid 3) = P(Y = y \mid X = 3)$

$$p_Y(6 \mid 3) = \frac{1/36}{7/36} = \frac{1}{7},\ p_Y(7 \mid 3) = p_Y(8 \mid 3) = p_Y(9 \mid 3) = \frac{2/36}{7/36} = \frac{2}{7}$$
$$p_Y(y \mid 3) = 0 \qquad \text{for } y = 2, 3, 4, 5, 10, 11, 12.$$

What is the continuous analog of the conditional probability mass function when X and Y are continuous random variables with a joint probability density function $f(x, y)$? In this situation, we have the complication that $P(Y = y) = 0$

for each real number y. Nevertheless, this situation also allows for a natural definition of the concept of conditional distribution. Toward this end, we need the probabilistic interpretations of the joint density function $f(x, y)$ and the marginal densities $f_X(x)$ and $f_Y(y)$ of the random variables X and Y. For small values of $\Delta a > 0$ and $\Delta b > 0$

$$\begin{aligned} &P\left(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a \mid b - \frac{1}{2}\Delta b \leq Y \leq b + \frac{1}{2}\Delta b\right) \\ &\quad = \frac{P\left(a - \frac{1}{2}\Delta a \leq X \leq a + \frac{1}{2}\Delta a, b - \frac{1}{2}\Delta b \leq Y \leq b + \frac{1}{2}\Delta b\right)}{P\left(b - \frac{1}{2}\Delta b \leq Y \leq b + \frac{1}{2}\Delta b\right)} \\ &\quad \approx \frac{f(a, b)\Delta a \Delta b}{f_Y(b)\Delta b} = \frac{f(a, b)}{f_Y(b)}\Delta a \end{aligned}$$

provided that (a, b) is a continuity point of $f(x, y)$ and $f_Y(b) > 0$. This leads to the following definition.

Definition 13.1 *If X and Y are continuous random variables with joint probability density function $f(x, y)$ and $f_Y(y)$ is the marginal density function of Y, then the conditional probability density function of X given that $Y = b$ is defined by*

$$f_X(x \mid b) = \frac{f(x, b)}{f_Y(b)}, \qquad -\infty < x < \infty$$

for any fixed b with $f_Y(b) > 0$.

A probabilistic interpretation can be given to $f_X(a \mid b)$: given that the observed value of Y is b, the probability of the other random variable X taking on a value in a small interval of length Δa around point a is approximately equal to $f_X(a \mid b)\Delta a$ if a is a continuity point of $f_X(x \mid b)$. The conditional probability that the random variable X takes on a value smaller than or equal to x given that $Y = b$ is denoted by $P(X \leq x \mid Y = b)$ and is defined by

$$P(X \leq x \mid Y = b) = \int_{-\infty}^{x} f_X(u \mid b)\, du.$$

Before discussing implications of this definition, we illustrate the concept of conditional probability density function with two examples.

Example 13.2 A point (X, Y) is chosen at random inside the unit circle. What is the conditional density of X?

Solution. In Example 11.4, we determined the joint density function $f(x, y)$ of X and Y together with the marginal density function $f_Y(y)$ of Y. This gives for any fixed b with $-1 < b < 1$,

$$f_X(x \mid b) = \begin{cases} \frac{1}{2\sqrt{1-b^2}} & \text{for } -\sqrt{1-b^2} < x < \sqrt{1-b^2} \\ 0 & \text{otherwise.} \end{cases}$$

In other words, the conditional distribution of X given that $Y = b$ is the uniform distribution on the interval $(-\sqrt{1-b^2}, \sqrt{1-b^2})$. The same distribution as that of the x-coordinate of a randomly chosen point of the horizontal chord through the point $(0, b)$. This chord has length $2\sqrt{1-b^2}$, by Pythagoras.

Example 13.3 Suppose that the random variables X and Y have a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. What are the conditional probability densities of X and Y?

Solution. The joint density function $f(x, y)$ is specified in Section 12.1. Also, in this section we find that the marginal probability densities $f_X(x)$ and $f_Y(y)$ of X and Y are given by the $N(\mu_1, \sigma_1^2)$ density and the $N(\mu_2, \sigma_2^2)$ density. Substituting the expressions for these densities in the formulas for the conditional densities, we find after simple algebra that the conditional probability density $f_X(x \mid b)$ of X given that $Y = b$ is the

$$N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(b - \mu_2), \sigma_1^2(1 - \rho^2)\right)$$

density and the conditional probability density $f_Y(y \mid a)$ of Y given that $X = a$ is the

$$N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(a - \mu_1), \sigma_2^2(1 - \rho^2)\right)$$

density. Thus the expected values of the conditional densities are linear functions of the conditional variable, and the conditional variances are constants.

The relation $f_X(x \mid y) = \frac{f(x,y)}{f_Y(y)}$ can be written in the more insightful form

$$f(x, y) = f_X(x \mid y) f_Y(y),$$

in analogy with $P(AB) = P(A \mid B)P(B)$. This representation of $f(x, y)$ may be helpful in simulating a random observation from the joint probability distribution of X and Y. First, a random observation for Y is generated from the density function $f_Y(y)$. If the value of this observation is y, a random observation for X is generated from the conditional density function $f_X(x \mid y)$. For example, the results of Examples 11.4 and 13.2 show that a random point (X, Y) in the unit circle can be simulated by generating first a random observation Y from the density function $\frac{2}{\pi}\sqrt{1-y^2}$ on $(-1, 1)$ and next a random observation X from the uniform density on $(-\sqrt{1-y^2}, \sqrt{1-y^2})$. How to obtain a random observation from the density of Y? A generally applicable method to generate a random observation from any given univariate density function is the acceptance-rejection method (see Problem 13.7). It is a generalization of the hit-or-miss method from Section 2.9.

Example 13.4 A very tasty looking toadstool growing in the forest is nevertheless so poisonous that it is fatal to squirrels that consume more than half of it. Squirrel 1, however, does partake of it, and later on squirrel 2 does the same. What is the probability that both squirrels survive? Assume that the first squirrel consumes a uniformly distributed amount of the toadstool, and the second squirrel a uniformly distributed amount of the remaining part of the toadstool.

Solution. To answer the question, let the random variable X represent the proportion of the toadstool consumed by squirrel 1 and let Y be the proportion of the toadstool consumed by squirrel 2. Using the uniformity assumption, it follows that $f_X(x) = 1$ for all $0 < x < 1$ and $f_Y(y \mid x) = \frac{1}{1-x}$ for $0 < y < 1 - x$. Applying the representation $f(x, y) = f_X(x) f_Y(y \mid x)$ leads to

$$f(x, y) = \frac{1}{1-x} \qquad \text{for } 0 < x < 1 \quad \text{and} \quad 0 < y < 1 - x.$$

The probability of both squirrels surviving is equal to

$$\begin{aligned} P(X \leq \frac{1}{2}, Y \leq \frac{1}{2}) &= \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}} f(x, y)\, dx\, dy = \int_0^{\frac{1}{2}} \frac{dx}{1-x} \int_0^{\frac{1}{2}} dy \\ &= \frac{1}{2} \int_{\frac{1}{2}}^1 \frac{du}{u} = \frac{1}{2} \ln(2) = 0.3466. \end{aligned}$$

13.2 Law of conditional probabilities

For discrete random variables X and Y, the unconditional probability $P(X = a)$ can be calculated from

$$P(X = a) = \sum_b P(X = a \mid Y = b) P(Y = b).$$

This *law of conditional probabilities* is a special case of Rule 8.1 in Chapter 8. In the situation of continuous random variables X and Y, the continuous analog of the law of conditional probabilities is:

Rule 13.1 *If the random variables X and Y are continuously distributed with a joint density function $f(x, y)$ and $f_Y(y)$ is the marginal density function of Y,* *then*

$$P(X \leq a) = \int_{-\infty}^{\infty} P(X \leq a \mid Y = y) f_Y(y)\, dy.$$

This statement is a direct consequence of the definition of the conditional probability $P(X \leq a \mid Y = y) = \int_{-\infty}^{a} f_X(x \mid y)\, dx$. Thus

$$\begin{aligned}\int_{-\infty}^{\infty} P(X \leq a \mid Y = y) f_Y(y)\, dy &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{a} \frac{f(x,y)}{f_Y(y)}\, dx\right] f_Y(y)\, dy\\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{a} f(x,y)\, dx\right] dy = \int_{-\infty}^{a} \left[\int_{-\infty}^{\infty} f(x,y)\, dy\right] dx\\
&= \int_{-\infty}^{a} f_X(x)\, dx = P(X \leq a).\end{aligned}$$

The importance of the continuous analog of the law of conditional probabilities can be hardly overestimated. In applications, the conditional probability $P(X \leq a \mid Y = y)$ is often calculated without explicitly using the joint distribution of X and Y, but through a direct physical interpretation of the conditional probability in the context of the concrete application. To illustrate this, let's return to Example 13.4 and calculate the probability that squirrel 2 will survive. This probability can be obtained as

$$\begin{aligned}P\left(Y \leq \frac{1}{2}\right) &= \int_{0}^{1} P\left(Y \leq \frac{1}{2} \mid X = x\right) f_X(x)\, dx =\\
&= \int_{0}^{\frac{1}{2}} \frac{0.5}{1-x}\, dx + \int_{\frac{1}{2}}^{1} 1\, dx = \frac{1}{2}\ln(2) + 0.5 = 0.8466.\end{aligned}$$

In the following example the law of conditional probabilities is used for the situation of a discrete random variable X and a continuous random variable Y. A precise definition of $P(X \leq a \mid Y = y)$ for this situation requires some technical machinery and will not be given. However, in the context of the concrete problem, it is immediately obvious what is meant by the conditional probability.

Example 13.5 Every morning at exactly the same time, Mr. Johnson rides the metro to work. He waits for the metro at the same place in the metro station. Every time the metro arrives at the station, the exact spot where it comes to a stop is a surprise. From experience, Mr. Johnson knows that the distance between him and the nearest metro door once the metro has stopped is uniformly distributed between 0 and 2 meters. Mr. Johnson is able to find a place to sit with probability $1 - \sqrt{\frac{1}{2}y}$ if the nearest door is y meters from where he is standing.

On any given morning, what is the probability that Mr. Johnson will succeed in finding a place to sit down?

Solution. The probability is not $1 - \sqrt{\frac{1}{2} \times 1} = 0.293$ as some people believe (they substitute the expected value of the distance to the nearest door for y into the formula $1 - \sqrt{\frac{1}{2}y}$). The correct value can be obtained as follows. Define the random variable X as equal to 1 when Mr. Johnson finds a seat and 0 otherwise. Now, define the random variable Y as the distance from Mr. Johnson's waiting place to the nearest metro door. Obviously

$$P(X = 1 \mid Y = y) = 1 - \sqrt{\frac{1}{2}y}.$$

The random variable Y has the probability density function $f_Y(y) = \frac{1}{2}$ for $0 < y < 2$. Hence, the unconditional probability that Mr. Johnson will succeed in finding a place to sit down on any given morning is equal to

$$P(X = 1) = \int_0^2 \left(1 - \sqrt{\frac{1}{2}y}\right) \frac{1}{2}\, dy = \frac{1}{3}.$$

The next example deals with a continuous version of the game of chance discussed in the Problems 2.29 and 3.22.

Example 13.6 Two players A and B in turn draw one or two random numbers between 0 and 1. For each player, the decision whether to go for a second draw depends on the result of the first draw. The object of the game is to have the highest total score, from one or two draws, without going over 1. Player A takes the first draw of one or two random numbers and then waits for the opponent's score. The opponent has the advantage of knowing the score of player A. What strategy maximizes the probability of player A winning? What is the value of this probability?

Solution. In analyzing this problem, it is natural to condition on the outcome of the first draw of player A. Denote by the random variable U the number player A gets at the first draw. Define the conditional probability

$$P_1(a) = \text{the winning probability of player } A \text{ if player } A \\ \text{stops after the first draw given that } U = a.$$

Also, define $P_2(a)$ as the winning probability of player A if player A continues for a second draw given that $U = a$. It will be seen below that there is a unique number a^* such that $P_2(a) \geq P_1(a)$ for all $0 \leq a \leq a^*$ and $P_1(a) \geq P_2(a)$ for all $a^* \leq a \leq 1$. Then, the optimal strategy for player A is to stop after the first draw if this draw gives a number larger than a^* and to continue otherwise. By the law of conditional probabilities, the overall winning probability of player A under this strategy is given by

$$P_A = \int_0^{a^*} P_2(a) f_U(a)\, da + \int_{a^*}^1 P_1(a) f_U(a)\, da,$$

where $f_U(a) = 1$ for all $0 \le a \le 1$. It remains to find $P_1(a)$, $P_2(a)$ and a^*.

Under the condition that player A has stopped with a score of a, player B can win in two possible ways: (1) the first draw of player B results in a number $y > a$, and (2) the first draw of player B results in a number $y \le a$ and his second draw gives a number between $a - y$ and $1 - y$. Denoting by the random variable Y the number player B gets at the first draw and using again the law of conditional probabilities, we then find

$$\begin{aligned} 1 - P_1(a) &= \int_a^1 1 \times f_Y(y)\, dy + \int_0^a [1 - y - (a - y)] f_Y(y)\, dy \\ &= 1 - a + (1 - a)a = 1 - a^2, \end{aligned}$$

showing that $P_1(a) = a^2$. To obtain $P_2(a)$, denote by the random variable V the number player A gets at the second draw. If player A has a total score of $a + v$ after the second draw with $a + v \le 1$, then player A will win with probability $(a + v)^2$, in view of the result $P_1(x) = x^2$ (only the final score of player A matters for player B). Thus

$$\begin{aligned} P_2(a) &= \int_0^{1-a} (a + v)^2 f_V(v)\, dv + \int_{1-a}^1 0 \times f_V(v)\, dv \\ &= \int_a^1 w^2\, dw = \frac{1}{3}(1 - a^3). \end{aligned}$$

The function $P_1(a) - P_2(a)$ is negative for $a = 0$ and positive for $a = 1$. Also, the function is increasing on (0, 1) (the derivative is positive). This proves the existence of a number a^* as claimed above. A numerical search procedure gives the solution $a^* = 0.53209$ to the equation $P_1(a) - P_2(a) = 0$. The winning probability of player A can next be calculated as $P_A = 0.4538$.

Problem 13.1 The length of time required to unload a ship has an $N(\mu, \sigma^2)$ distribution. The crane to unload the ship has just been overhauled and the time it will operate until the next breakdown has an exponential distribution with an expected value of $1/\lambda$. What is the probability of no breakdown during the unloading of the ship?

Problem 13.2 Consider the three-players variant of Example 13.6. Calculate the optimal strategy for the first player A, assuming that the other two players B and C play optimally. What is the probability distribution function of the final score of player A under his optimal strategy? What are the win probabilities of the players A, B, and C?

Problem 13.3 An opaque bowl contains B balls, where B is given. Each ball is red or white. The number of red balls in the bowl is unknown, but has a binomial distribution with parameters B and p. You randomly select r balls out of the urn without replacing any. Use the law of conditional probabilities to obtain the probability distribution of the number of red balls among the selected balls. Does surprise the result you? Can you give a direct probabilistic argument for the result obtained?

Problem 13.4 The random variables X_1 and X_2 are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributed. Let the random variable V be distributed as X_1 with given probability p and as X_2 with probability $1 - p$. What is the probability density of V? Is this probability density the same as the probability density of the random variable $W = pX_1 + (1 - p)X_2$ when X_1 and X_2 are independent?

Problem 13.5 Use twice the law of conditional probabilities to calculate the probability that the quadratic equation $Ax^2 + Bx + C = 0$ will have two real roots when A, B, and C are independent samples from the uniform distribution on (0, 1).

Problem 13.6 You leave work at random times between 5.45 p.m. and 6.00 p.m. to take the bus home. Bus numbers 1 and 3 bring you home. You take the first bus that arrives. Bus number 1 arrives exactly every 15 minutes starting from the hour, whereas bus number 3 arrives according to a Poisson process with the same average frequency as bus number 1 (that is, the interarrival times of buses number 3 are independent and exponentially distributed with an expected value of 15 minutes). What is the probability that you take bus number 1 home on any given day? Use the law of conditional probabilities and the memoryless property of the exponential distribution to verify that this probability equals $1 - e^{-1}$ with $e = 2.71828\ldots$. Can you give an intuitive explanation why the probability is larger than $\frac{1}{2}$?

Problem 13.7 Suppose that U and Y are independent random variables, where U is uniformly distributed on (0, 1) and Y is a continuous random variable with probability density $g(y)$. Let $f(y)$ be another probability density function such that, for some constant c (≥ 1), $f(y) \leq cg(y)$ for all y. Verify that

$$P\left(Y \leq x \mid U \leq \frac{f(Y)}{cg(Y)}\right) = \int_{-\infty}^{x} f(y)dy \qquad \text{for all } x.$$

Remark: This result underlies the so-called *acceptance-rejection method* for simulating from a “difficult” density $f(x)$ via an “easy” density $g(y)$. This method proceeds as follows: Step 1. Generate Y having density $g(y)$ and generate a random number U. Step 2. If $U \leq \frac{f(Y)}{cg(Y)}$, then accept Y as a sample from $f(x)$; otherwise, return to step 1. Intuitively, the acceptance-rejection method generates a random point $(Y, U \times cg(Y))$ under the graph of $cg(y)$ and accepts the point only when it also falls under the graph of $f(y)$ as is the case when $U \times cg(Y) \leq f(Y)$. The hit-or-miss method from Section 2.9 can be seen as a special case of the acceptance-rejection method.

13.3 Law of conditional expectations

In the case that the random variables X and Y have a discrete distribution, the *conditional expectation* of X given that $Y = b$ is defined by

$$E(X \mid Y = b) = \sum_x x\, P(X = x \mid Y = b)$$

for each b with $P(Y = b) > 0$ (assuming that the sum is well defined). In the case that X and Y are continuously distributed with joint probability density function $f(x, y)$, the *conditional expectation* of X given that $Y = b$ is defined by

$$E(X \mid Y = b) = \int_{-\infty}^{\infty} x\, f_X(x \mid b)\, dx$$

for each b with $f_Y(b) > 0$ (assuming that the integral is well defined).

Just as the law of conditional probabilities directly follows from the definition of the conditional distribution of X given that $Y = y$, the law of conditional expectations is a direct consequence of the definition of $E(X \mid Y = y)$. In the discrete case the *law of conditional expectations* reads as

$$E(X) = \sum_y E(X \mid Y = y)\, P(Y = y),$$

while for the continuous case the law reads as

$$E(X) = \int_{-\infty}^{\infty} E(X \mid Y = y)\, f_Y(y)\, dy.$$

In words, the law of conditional expectations says that the unconditional expected value of X may be obtained by first conditioning on an appropriate random variable Y to get the conditional expected value of X given that $Y = y$ and then taking the expectation of this quantity with respect to Y.

Example 13.7 The relationship between household expenditure and net income of households in Fantasia is given by the joint density function

$$f(x, y) = \begin{cases} c(x-10)(y-10) & \text{for } 10 < x < y < 30 \\ 0 & \text{otherwise,} \end{cases}$$

where the normalizing constant $c = \frac{1}{20{,}000}$. What is the expected value of the household expenditure of a randomly chosen household given that the income of the household is y? What is the probability that the household expenditure is more than 20 given that the income is 25?

Solution. To answer the questions, let the random variables X and Y represent the household expenditure and the net income of a randomly selected household. How do we find $E(X \mid Y = y)$? We first determine the marginal density of the conditioning variable Y

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x, y)\, dx = c \int_{10}^{y} (x-10)(y-10)\, dx \\ &= \frac{1}{2} c (y-10)^3 \qquad \text{for } 10 < y < 30. \end{aligned}$$

Using the relation $f_X(x \mid y) = \frac{f(x,y)}{f_Y(y)}$, we next obtain, for fixed y, the conditional density function

$$f_X(x \mid y) = \begin{cases} \frac{2(x-10)}{(y-10)^2} & \text{for } 10 < x < y \\ 0 & \text{otherwise.} \end{cases}$$

This gives the desired result

$$E(X \mid Y = y) = \int_{10}^{y} x \frac{2(x-10)}{(y-10)^2}\, dx = 10 + \frac{2}{3}(y-10) \quad \text{for } 10 < y < 30.$$

Further, using the general formula $P(X \in A \mid Y = y) = \int_A f_X(x \mid y)\, dx$

$$P(X > 20 \mid Y = 25) = \int_{20}^{25} \frac{2(x-10)}{(25-10)^2}\, dx = \frac{5}{9}.$$

Remark 13.1 For two dependent random variables X and Y, let $m(x) = E(Y \mid X = x)$. The curve of the function $y = m(x)$ is called the *regression curve* of Y on X. It is a better measure for the dependence between X and Y than the correlation coefficient (recall that dependence does not necessarily imply a nonzero correlation coefficient).† In statistical applications it is often

† It is not generally true that $E(Y \mid X = x) = E(Y)$ for all x and $E(X \mid Y = y) = E(X)$ for all y are sufficient conditions for the independence of X and Y. This is shown by the example in which (X, Y) has the probability mass function $p(x, y) = \frac{1}{8}$ for $(x, y) = (1, 1), (1,–1), (–1,1), (–1,–1)$ and $p(x, y) = \frac{1}{2}$ for $(x, y) = (0, 0)$.

the case that we can observe the random variable X but we want to know the dependent random variable Y. The function value $y = m(x)$ can be used as a prediction of the value of the random variable Y given the observation x of the random variable X. The function $m(x) = E(Y \mid X = x)$ is an optimal prediction function in the sense that this function minimizes

$$E[(Y - g(X))^2]$$

over all functions $g(x)$. We only sketch the proof of this result. For any random variable U, the minimum of $E[(U - c)^2]$ over all constants c is achieved for the constant $c = E(U)$. This follows by differentiating $E[(U - c)^2] = E(U^2) - 2cE(U) + c^2$ with respect to c. Using the law of conditional expectations, $E[(Y - g(X))^2]$ can be expressed as

$$E[(Y - g(X))^2] = \int_{-\infty}^{\infty} E[(Y - g(X))^2|X = x]f_X(x)dx.$$

For every x the inner side of the integral is minimized by $g(x) = E(Y|X = x)$, yielding that $m(X)$ is the minimum mean squared error predictor of Y from X. By the law of conditional expectation, the statistic $m(X)$ has the same expected value as Y. But the predictor $m(X)$ has the nice feature that its variance is usually smaller than var(Y) itself. An intuitive explanation of this fact is that the conditional distribution of Y given the value of X involves more information than the distribution of Y alone. For the case that X and Y have a bivariate normal distribution, it follows from the results in Example 13.3 that the optimal prediction function $m(x)$ coincides with the best linear prediction function discussed in Section 11.4. The best linear prediction function uses only the expected values, the variances, and the correlation coefficient of the random variables X and Y.

Conditional expectations as a tool

In applied probability problems, the law of conditional expectations is a very useful result to calculate unconditional expectations. Beginning students have often difficulties in choosing the conditioning variable when they do a mathematical analysis. However, in a simulation program this "difficult" step would offer no difficulties at all and would be naturally done. So, our advice to students is as follows: if the first step in the analysis looks difficult to you, think of what you would do in a simulation program of the problem. As illustration, we give the following example.

Example 13.8 Someone purchases a liability insurance policy. The probability that a claim will be made on the policy is 0.1. In case of a claim, the size of the claim has an exponential distribution with an expected value of \$1,000,000. The maximum insurance policy payout is \$2,000,000. What is the expected value of the insurance payout?

Solution. The insurance payout is a mixed random variable: it takes on one of the discrete values 0 and 2×10^6 or a value in the continuous interval $(0,\ 2 \times 10^6)$. Its expected value is calculated through a two-stage process. In a simulation program you would first simulate whether a claim occurs or not. Hence, we first condition on the outcome of the random variable I, where $I = 0$ if no claim is made and $I = 1$ otherwise. The insurance payout is 0 if I takes on the value 0, and otherwise the insurance payout is distributed as $\min(2 \times 10^6, D)$, where the random variable D has an exponential distribution with parameter $\lambda = 1/10^6$. Thus, by conditioning

$$E(\text{insurance payout}) = 0.9 \times 0 + 0.1 \times E[\min(2 \times 10^6, D)].$$

Using the substitution rule, it follows that

$$E[\min(2 \times 10^6, D)] = \int_0^\infty \min(2 \times 10^6, x)\lambda e^{-\lambda x}\, dx$$
$$= \int_0^{2\times 10^6} x\lambda e^{-\lambda x}\, dx + \int_{2\times 10^6}^\infty (2 \times 10^6)\lambda e^{-\lambda x}\, dx.$$

It is left to the reader to verify the calculations leading to

$$E[\min(2 \times 10^6, D)] = 10^6(1 - e^{-2}) = 864{,}665 \quad \text{dollars}.$$

Hence, we can conclude that $E(\text{insurance payout}) = \$86{,}466.50$.

Thinking recursively can be very rewarding for the calculation of expected values. This is shown in the next example. The concepts of state and state transition are hidden in the solution of this example. These concepts stand central in Chapter 15 on Markov chains.

Example 13.9 In any drawing of the Lotto 6/45 six different numbers are chosen at random from the numbers $1, 2, \ldots, 45$. What is the expected value of the number of draws until each of the numbers $1, \ldots, 45$ has been drawn?

Solution. Define μ_i as the expected value of the remaining number of draws that are needed to obtain each of the numbers $1, 2, \ldots, 45$ when i of those numbers are still missing for $i = 1, 2, \ldots, 45$.[†] To find the desired value μ_{45},

[†] This is a natural definition: in a simulation program you would automatically use a state variable that keeps track of how many numbers are still missing.

we use a recurrence relation for the μ_i. By conditioning on the result of the next draw, we find

$$\mu_i = 1 + \sum_{k=0}^{6} \mu_{i-k} \frac{\binom{i}{k}\binom{45-i}{6-k}}{\binom{45}{6}} \qquad \text{for } i = 1, 2, \ldots, 45,$$

with the convention that $\mu_j = 0$ for $j \leq 0$. Applying this recurrence relation, we obtain $\mu_{45} = 31.497$.

Problem 13.8 The percentage of zinc content and iron content in ore from a certain location has the joint density $f(x, y) = \frac{1}{350}(5x + y - 30)$ for $2 < x < 3$, $20 < y < 30$ and $f(x, y) = 0$ otherwise. What is the expected value of the zinc content in a sample of ore given that the iron content is y? What is the probability that the zinc content is more than 2.5% given that the iron content is 25%?

Problem 13.9 A farming operation is located in a remote area that is more or less unreachable in the winter. As early as September, the farmer must order fuel oil for the coming winter. The amount of fuel oil he needs each winter is random, and depends on the severity of the winter weather to come. The winter will be normal with probability 2/3 and very cold with probability 1/3. The number of gallons of oil the farmer needs to get through the winter is $N(\mu_1, \sigma_1^2)$ distributed in a normal winter and $N(\mu_2, \sigma_2^2)$ distributed in a very cold winter. The farmer decides in September to stock up Q gallons of oil for the coming winter. What is the probability that he will run out of oil in the coming winter? What is the expected value of the number of gallons the farmer will come up short for the coming winter? What is the expected value of the number of gallons he will have left over at the end of the winter?

Problem 13.10 Nobel airlines has a direct flight from Amsterdam to Palermo. This particular flight uses an aircraft with $N = 150$ seats. The number of people who seek to reserve seats for a given flight has a Poisson distribution with expected value $\lambda = 170$. The airline management has decided to book up to $Q = 165$ passengers in order to protect themselves against no-shows. The probability of a booked passenger not showing up is $q = 0.07$. The booked passengers act independently of each other. What is the expected value of the number of people who show up for a given flight? What is the expected value of the number of people who show up but cannot be seated due to overbooking?

Problem 13.11 Consider the casino game Red Dog from Problem 3.25 again. Suppose that the initial stake of the player is \$10. What are the expected values of the total amount staked and the payout in any given play? Use the law of conditional expectations to find these expected values.

Problem 13.12 Let's return to Problem 13.6. Use the law of conditional expectations to verify that the expected value of your waiting time until the next bus arrival is equal to $\frac{15}{e}$.

Problem 13.13 A fair coin is tossed no more than n times, where n is fixed in advance. You stop the coin-toss experiment as soon as the proportion of heads exceeds $\frac{1}{2}$ or as soon as n tosses are done, whichever occurs first. Use the law of conditional expectations to calculate, for $n = 5, 10, 25$, and 50, the expected value of the proportion of heads at the moment the coin-toss experiment is stopped. *Hint*: define the random variable $X_k(i)$ as the proportion of heads upon stopping given that k tosses are still possible and heads turned up i times so far. Set up a recursion equation for $E[X_k(i)]$.

Problem 13.14 In the game of Pig each player's turn consists of repeatedly rolling a die. If the player rolls a 1, the player's turn ends and nothing is added to the player's score. If the player rolls a number other than 1, the player's turn continues and the player has the choice between rolling the die again or holding. If the player holds, the accumulated points during the turn are added to the player's total score. Use a recursion to find the probability of getting 20 or more points during a single turn. Also, use a recursion to find the expected value of the number of turns needed to reach a total score of 100 or more points when the player's strategy is to hold the turn when the accumulated points during the turn are 20 or more.

Problem 13.15 You spin a game board spinner in a round box whose circumference is marked with a scale from 0 to 1. When the spinner comes to rest, it points to a random number between 0 and 1. After your first spin, you have to decide whether to spin the spinner for a second time. Your payoff is \$1,000 times the total score of your spins as long as this score does not exceed 1; otherwise, your payoff is zero. What strategy maximizes the expected value of your payoff? What is the expected value of your payoff under the optimal strategy?

Problem 13.16 Fix a number a with $0 < a < 1$. You draw repeatedly a random number from an interval until you obtain a random number below a. The first random number is chosen from the interval (0,1) and each subsequent random number is chosen from the interval between zero and the previously chosen random number. What is the expected value of the number of drawings until you have a random number below a?

14
Generating functions

Generating functions were introduced by the Swiss genius Leonhard Euler (1707–1783) in the eighteenth century to facilitate calculations in counting problems. However, this important concept is also extremely useful in applied probability, as was first demonstrated by the work of Abraham de Moivre (1667–1754) who discovered the technique of generating functions independently of Euler. In modern probability theory, generating functions are an indispensable tool in combination with methods from numerical analysis. The purpose of this chapter is to give the basic properties of generating functions and to show the utility of this concept. First, the generating function is defined for a discrete random variable on nonnegative integers. Next, we consider the more general moment-generating function, which is defined for any random variable. The (moment) generating function is a powerful tool for both theoretical and computational purposes. In particular, it can be used to prove the central limit theorem. A sketch of the proof will be given.

14.1 Generating functions

We first introduce the concept of generating functions for a discrete random variable X whose possible values belong to the set of nonnegative integers.

Definition 14.1 *If X is a nonnegative, integer-valued random variable, then the generating function of X is defined by*

$$G_X(z) = \sum_{k=0}^{\infty} z^k P(X = k), \qquad |z| \leq 1.$$

The power series $G_X(z)$ is absolutely convergent for any $|z| \leq 1$ (why?). For any z, we can interpret $G_X(z)$ as

$$G_X(z) = E(z^X),$$

as follows by applying Rule 9.2. The probability mass function of X is uniquely determined by the generating function of X. To see this, use the fact that the derivative of an infinite series is obtained by differentiating the series term by term. Thus

$$\frac{d^r}{dz^r}G_X(z) = \sum_{k=r}^{\infty} k(k-1)\cdots(k-r+1)z^{k-r}P(X=k), \qquad r = 1, 2, \ldots.$$

In particular, by taking $z = 0$

$$P(X = r) = \frac{1}{r!}\frac{d^r}{dz^r}G_X(z)|_{z=0}, \qquad r = 1, 2, \ldots.$$

This proves that the generating function uniquely determines the probability mass function. This basic result explains the importance of the generating function. In many applications, it is relatively easy to obtain the generating function of a random variable X even when the probability mass function is not explicitly given. An example will be given below. Once we know the generating function of a random variable X, it is a simple matter to obtain the factorial moments of the random variable X. The rth factorial moment of the random variable X is defined by $E[X(X-1)\cdots(X-r+1)]$ for $r = 1, 2, \ldots$. In particular, the first factorial moment of X is the expected value of X. The variance of X is determined by the first and the second factorial moment of X. Putting $z = 1$ in the above expression for the rth derivative of $G_X(z)$, we obtain

$$E\left[X(X-1)\cdots(X-r+1)\right] = \frac{d^r}{dz^r}G_X(z)|_{z=1}, \qquad r = 1, 2, \ldots.$$

In particular

$$E(X) = G'_X(1) \quad \text{and} \quad E(X^2) = G''_X(1) + G'_X(1).$$

Example 14.1 Suppose that the random variable X has a Poisson distribution with expected value λ. Then

$$\sum_{k=0}^{\infty} z^k e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda z)^k}{k!} = e^{-\lambda}e^{\lambda z},$$

using the series expansion $e^x = \sum_{n=0}^{\infty} x^n/n!$. Hence

$$G_X(z) = e^{-\lambda(1-z)}, \qquad |z| \leq 1.$$

Differentiating $G_X(z)$ gives $G'_X(1) = \lambda$ and $G''_X(1) = \lambda^2$. Hence, $E(X) = \lambda$ and $E(X^2) = \lambda^2 + \lambda$. This implies that both the expected value and the variance of a Poisson-distributed random variable with parameter λ are given by λ, in agreement with earlier results in Example 9.7.

14.1.1 Convolution rule

The importance of the concept of generating function comes up especially when calculating the probability mass function of a sum of independent random variables that are nonnegative and integer-valued.

Rule 14.1 *Let X and Y be two nonnegative, integer-valued random variables. If the random variables X and Y are independent, then*

$$G_{X+Y}(z) = G_X(z)G_Y(z), \qquad |z| \leq 1.$$

Rule 14.1 is known as the *convolution rule* for generating functions and can be directly extended to the case of a finite sum of independent random variables. The proof is simple. If X and Y are independent, then the random variables $U = z^X$ and $V = z^Y$ are independent for any fixed z (see Rule 9.5). Also, by Rule 9.7, $E(UV) = E(U)E(V)$ for independent U and V. Thus

$$E\big(z^{X+Y}\big) = E(z^X z^Y) = E(z^X)E(z^Y),$$

proving that $G_{X+Y}(z) = G_X(z)G_Y(z)$. The converse of the statement in Rule 14.1 is, in general, not true. The random variables X and Y are not necessarily independent if $G_{X+Y}(z) = G_X(z)G_Y(z)$. It is left to the reader to verify that a counterexample is provided by the random vector (X, Y) that takes on the values (1,1), (2,2) and (3,3) each with probability $\frac{1}{9}$ and the values (1,2), (2,3) and (3,1) each with probability $\frac{2}{9}$. This counterexample was communicated to me by Fred Steutel.

Example 14.2 Suppose that X and Y are independent random variables that are Poisson distributed with respective parameters λ and μ. What is the probability mass function of $X + Y$?

Solution. Using the result from Example 14.1, we find

$$G_{X+Y}(z) = e^{-\lambda(1-z)}e^{-\mu(1-z)} = e^{-(\lambda+\mu)(1-z)}, \qquad |z| \leq 1.$$

Since a Poisson-distributed random variable with parameter $\lambda + \mu$ has the generating function $e^{-(\lambda+\mu)(1-z)}$ and the generating function $G_{X+Y}(z)$ uniquely determines the probability mass function of $X + Y$, it follows that $X + Y$ has a Poisson distribution with parameter $\lambda + \mu$.

Problem 14.1 Suppose that the random variable X has a binomial distribution with parameters n and p. Use the fact that X can be represented as the sum of n independent Bernoulli variables to derive the generating function of X. In a similar way, derive the generating function of the random variable X having a negative binomial distribution.

Problem 14.2 Use the results of Problem 14.1 to obtain the expected value and the variance of both the binomial distribution and the negative binomial distribution.

Problem 14.3 Suppose that you draw a number at random from the unit interval r times. A draw is called a "record draw" when the resulting number is larger than the previously drawn numbers. Determine the generating function of the number of record draws. What are the expected value and variance of the number of record draws?

Problem 14.4 The nonnegative, integer-valued random variables X and Y are independent and identically distributed. Verify that X and Y are Poisson distributed when the sum $X + Y$ is Poisson distributed.

Problem 14.5 The number of claims an insurance company will receive is a random variable N having a Poisson distribution with expected value μ. The claim sizes are independent random variables with a common probability mass function a_k for $k = 1, 2 \ldots$. Let the total claim size S be defined by $S = \sum_{i=1}^{N} X_i$, where $X_1, X_2, \ldots$ represent the individual claim sizes. Prove that the generating function of the random sum S is given by $e^{-\mu[1-A(z)]}$, where $A(z)$ is the generating function of the individual claim sizes. Also, verify that $E(S) = E(N)E(X_1)$ and $\text{var}(S) = E(N)\text{var}(X_1) + \text{var}(N)[E(X_1)]^2$ with $E(N) = \text{var}(N) = \mu$. *Remark*: the probability distribution of the random sum S is called the compound Poisson distribution.

Problem 14.6 The number of customers asking for a new product has a Poisson distribution with expected value μ. The demands of the customers are independent random variables each having the probability mass function $a_k = -\alpha^k[k \ln(1 - \alpha)]^{-1}$, $k = 1, 2, \ldots$ for a given $0 < \alpha < 1$. Let the random variable S denote the total customer demand. Use the expansion $-\ln(1 - x) = \sum_{k=0}^{\infty} x^k/k$ for $0 < x < 1$ to find the generating function of S. *Remark*: this generating function can be analytically inverted to obtain that $P(S = k) = \frac{\Gamma(s+k)}{\Gamma(s)\Gamma(k+1)} p^s(1 - p)^k$ for $k = 0, 1, \ldots$, where $s = -\mu/\ln(1 - \alpha)$ and $p = 1 - \alpha$. This is an extension of the negative binomial distribution.

Inversion of the generating function

In many applications, it is possible to derive an explicit expression for the generating function of a random variable X whose probability mass function is not readily available and has a complicated form. Is this explicit expression for the generating function of practical use apart from calculating the moments of

X? The answer is yes! If an explicit expression for the generating function of the random variable X is available, then the numerical values of the (unknown) probability mass function of X can be calculated by appealing to the discrete Fast Fourier Transform method from numerical analysis (this algorithm functions in the seemingly mystical realm of complex numbers, which world nonetheless is of great real-world significance). An explanation of how this extremely powerful method works is beyond the scope of this book. However, it is useful to know that this method exists. In practice, it is often used to calculate convolutions of discrete probability distributions.

Example 14.3 In the coupon collector's problem from Section 3.2, we calculated the expected value of the random variable X representing the number of bags of chips that must be purchased in order to get a complete set of n distinct flippos. How do we calculate the probability distribution of the random variable X?

Solution. The calculations can be done with the help of the generating function of X. The random variable X can be written as

$$X = Y_1 + Y_2 + \cdots + Y_n,$$

where the random variable Y_i denotes the number of bags of chips needed in order to go from $i-1$ to i different flippos. The random variables $Y_1, \ldots, Y_n$ are independent and the random variable Y_i has a geometric distribution with parameter $p_i = 1 - \frac{(i-1)}{n}$ for $i = 1, \ldots, n$. A random variable Y with the geometric distribution $P(Y = k) = (1-p)^{k-1}p$ for $k \geq 1$ satisfies

$$\begin{aligned}\sum_{k=0}^{\infty} P(Y=k)z^k &= \sum_{k=1}^{\infty}(1-p)^{k-1}pz^k = pz\sum_{k=1}^{\infty}((1-p)z)^{k-1} \\ &= pz\sum_{j=0}^{\infty}((1-p)z)^j = \frac{pz}{1-(1-p)z},\end{aligned}$$

using the fact that the geometric series $\sum_{j=0}^{\infty} x^j$ equals $\frac{1}{1-x}$ for $|x| < 1$ (see the Appendix). Since $G_X(z) = G_{Y_1}(z)G_{Y_2}(z)\cdots G_{Y_n}(z)$ by the independence of $Y_1, \ldots, Y_n$, it follows that

$$G_X(z) = \frac{p_1 p_2 \cdots p_n z^n}{(1-z+p_1 z)(1-z+p_2 z)\cdots(1-z+p_n z)}.$$

The coupon collector's problem with $n = 365$ flippos enables us to calculate how many persons are needed to have a group of persons in which all 365 possible birthdays (excluding February 29) are represented with a probability of at least 50%. Using the discrete Fast Fourier Transform method we can calculate that the group should consist of 2,287 randomly picked persons.

Problem 14.7 You participate in a game that consists of a series of independent plays. Any play results in a win with probability p, in a loss with probability q and in a draw with probability r, where $p+q+r=1$. One point is added to your score each time a play is won; otherwise, your score remains unchanged. The game is ended as soon as you lose a play. Let the random variable X denote your total score when the game is ended. Use a generating function to find the probability mass function of X. What is the probability mass function of X? *Hint:* condition on the outcome of the first play to verify that $E(z^X) = pzE(z^X) + q + rE(z^X)$ (this approach is called the method of *first-step analysis*).

Problem 14.8 Independently of each other, you generate integers at random from $0, 1, \ldots, 9$ until a zero is obtained. Use the method of first-step analysis to obtain the generating function of the sum of the generated integers.

Problem 14.9 Let X be the number of tosses of a fair coin until the number of heads first exceeds the number of tails. Use the method of first-step analysis to obtain the generating function of X. What is the expected value of X?

14.1.2 Branching processes and generating functions

The family name is inherited by sons only. Take a father who has one or more sons. In turn, each of his sons will have a random number of sons, each son of the second generation will have a random number of sons, and so forth. What is the probability that the family name will ultimately die out? The process describing the survival of family names is an example of a so-called branching process. Branching processes arise naturally in many situations. In physics, the model of branching processes can be used to study neutron chain reaction. A chance collision of a nucleus with a neutron yields a random number of new neutrons. Each of these secondary neutrons may hit some other nuclei, producing more additional neutrons, and so forth. In genetics, the model can be used to estimate the probability of long-term survival of genes that are subject to mutation. All of these examples possess the following structure. There is a population of individuals able to produce offspring of the same kind. Each individual will, by the end of its lifetime, have produced j new offspring with probability p_j for $j = 0, 1, \ldots$. All offspring behave independently. The number of individuals initially present, denoted by X_0, is called the size of the 0th generation. All offspring of the 0th generation constitute the first generation, and their number is denoted by X_1. In general, let X_n denote the size of the nth generation. We are interested in the probability that the population will eventually die out. To avoid uninteresting cases, it is assumed that $0 < p_0 < 1$. In order to find the *extinction probability*, it is no restriction to assume that $X_0 = 1$ (why?). Define the probability u_n by

$$u_n = P(X_n = 0).$$

Obviously, $u_0 = 0$ and $u_1 = p_0$. Noting that $X_n = 0$ implies $X_{n+1} = 0$, it follows that $u_{n+1} \geq u_n$ for all n. Since u_n is a nondecreasing sequence of numbers, $\lim_{n\to\infty} u_n$ exists. Denote this limit by u_∞. The probability u_∞ is the desired extinction probability. This requires some explanation. The probability that extinction will ever occur is defined as $P(X_n = 0 \text{ for some } n \geq 1)$. However, $\lim_{n\to\infty} P(X_n = 0) = P(X_n = 0 \text{ for some } n \geq 1)$, using the fact that $\lim_{n\to\infty} P(A_n) = P(\bigcup_{n=1}^{\infty} A_n)$ for any nondecreasing sequence of events A_n. The probability u_∞ can be computed by using the generating function $P(z) = \sum_{j=0}^{\infty} p_j z^j$ of the offspring distribution p_j. To do so, we first argue that

$$u_n = \sum_{k=0}^{\infty} (u_{n-1})^k p_k \qquad \text{for } n = 2, 3, \ldots.$$

This relation can be explained using the law of conditional probabilities. Fix $n \geq 2$. Now, condition on $X_1 = k$ and use the fact that the k subpopulations generated by the distinct offspring of the original parent behave independently and follow the same distributional law. The probability that any particular one of them will die out in $n-1$ generations is u_{n-1} by definition. Thus, the probability that all k subpopulations die out in $n-1$ generations is equal to $P(X_n = 0 \mid X_1 = k) = (u_{n-1})^k$ for $k \geq 1$. This relation is also true for $k = 0$, since $X_1 = 0$ implies that $X_n = 0$ for all $n \geq 2$. The equation for u_n next follows using the fact that

$$P(X_n = 0) = \sum_{k=0}^{\infty} P(X_n = 0 \mid X_1 = k) p_k,$$

by the law of conditional probabilities.

Using the definition of the generating function $P(z) = \sum_{k=0}^{\infty} p_k z^k$, the recursion equation for u_n can be rewritten as

$$u_n = P(u_{n-1}) \qquad \text{for } n = 2, 3, \ldots.$$

Next, by letting $n \to \infty$ in both sides of this equation and using a continuity argument, it can be shown that the desired probability u_∞ satisfies the equation

$$u = P(u).$$

This equation may have more than one solution. However, it can be shown that u_∞ is the smallest positive root of the equation $u = P(u)$. It may happen that $u_\infty = 1$, that is, the population is sure to die out ultimately. The case of

$u_\infty = 1$ can only happen if the expected value of the offspring distribution p_j is smaller than or equal to 1. The proof of this fact is omitted. As illustration, consider the numerical example with $p_0 = 0.25$, $p_1 = 0.25$ and $p_2 = 0.5$. The equation $u = P(u)$ then becomes the quadratic equation $u = \frac{1}{4} + \frac{1}{4}u + \frac{1}{2}u^2$. This equation has roots $u = 1$ and $u = \frac{1}{2}$. The smallest root gives the extinction probability $u_\infty = \frac{1}{2}$.

Problem 14.10 Every adult male in a certain society is married. Twenty percent of the married couples have no children. The other 80% have two or three children with respective probabilities $\frac{1}{3}$ and $\frac{2}{3}$, each child being equally likely to be a boy or a girl. What is the probability that the male line of a father with one son will eventually die out?

Problem 14.11 A population of bacteria begins with a single individual. In each generation, each individual dies with probability $\frac{1}{3}$ or splits in two with probability $\frac{2}{3}$. What is the probability that the population will die out by generation 3 and what is the probability that the population will die out eventually? What are these probabilities if the initial population consists of two individuals?

14.2 Moment-generating functions

How do we generalize the concept of generating function when the random variable is not integer-valued and nonnegative? The idea is to work with $E(e^{tX})$ instead of $E(z^X)$. Since e^{tX} is a nonnegative random variable, $E(e^{tX})$ is defined for any value of t. However, it may happen that $E(e^{tX}) = \infty$ for some values of t. For any nonnegative random variable X, we have that $E(e^{tX}) < \infty$ for any $t \leq 0$ (why?), but $E(e^{tX})$ need not be finite when $t > 0$. To illustrate this, suppose that the nonnegative random variable X has the one-sided Cauchy density function $f(x) = (2/\pi)/(1 + x^2)$ for $x > 0$. Then, $E(e^{tX}) = \int_0^\infty e^{tx} f(x)\, dx = \infty$ for any $t > 0$, since $e^{tx} \geq 1 + tx$ and $\int_0^\infty \frac{x}{1+x^2} dx = \infty$. In the case that the random variable X can take on both positive and negative values, then it may happen that $E(e^{tX}) = \infty$ for all $t \neq 0$. An example is provided by the random variable X having the two-sided Cauchy density function $f(x) = (1/\pi)/(1 + x^2)$ for $-\infty < x < \infty$. Fortunately, most random variables X of practical interest have the property that $E(e^{tX}) < \infty$ for all t in a neighborhood of 0.

Definition 14.2 *A random variable X is said to have a moment-generating function if $E(e^{tX}) < \infty$ for all t in an interval of the form $-\delta < t < \delta$ for some $\delta > 0$. For those t with $E(e^{tX}) < \infty$ the moment-generating function of X is* *defined and denoted by*

$$M_X(t) = E(e^{tX}).$$

If the random variable X has a probability density function $f(x)$, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx.$$

As an illustration, consider the case of an exponentially distributed random variable X. The density function $f(x)$ of X is equal to $\lambda e^{-\lambda x}$ for $x > 0$ and 0 otherwise. Then, $M_X(t) = \lambda \int_0^{\infty} e^{(t-\lambda)x}\, dx$. This integral is finite only if $t - \lambda < 0$. Thus, $M_X(t)$ is defined only for $t < \lambda$ and is then given by $M_X(t) = \lambda/(\lambda - t)$.

The explanation of the name moment-generating function is as follows. If the moment-generating function $M_X(t)$ of the random variable X exists, then it can be shown that

$$M_X(t) = 1 + tE(X) + t^2 \frac{E(X^2)}{2!} + t^3 \frac{E(X^3)}{3!} + \cdots$$

for $-\delta < t < \delta$. Heuristically, this result can be seen by using the expansion $E(e^{tX}) = E(\sum_{n=0}^{\infty} t^n \frac{X^n}{n!})$ and interchanging the order of expectation and summation. Conversely, the moments $E(X^r)$ for $r = 1, 2, \ldots$ can be obtained from the moment-generating function $M_X(t)$ when $E(e^{tX})$ exists in a neighborhood of $t = 0$. Assuming that X has a probability density function $f(x)$, it follows from advanced calculus that

$$\frac{d^r}{dt^r} \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \int_{-\infty}^{\infty} x^r e^{tx} f(x)\, dx$$

for $-\delta < t < \delta$. Taking $t = 0$, we obtain

$$E(X^r) = \frac{d^r}{dt^r} M_X(t)|_{t=0}, \qquad r = 1, 2, \ldots.$$

In particular,

$$E(X) = M'_X(0) \quad \text{and} \quad E(X^2) = M''_X(0).$$

A moment-generating function determines not only the moments of a random variable X, but it also determines uniquely the probability distribution of X. The following uniqueness theorem holds for the moment-generating function.

Rule 14.2 *If the moment-generating functions $M_X(t)$ and $M_Y(t)$ of the random variables X and Y exist and $M_X(t) = M_Y(t)$ for all t satisfying $-\delta < t < \delta$ for some $\delta > 0$, then the random variables X and Y are identically distributed.*

The proof of this rule is beyond the scope of this book. Also, we have the following very useful rule.

Rule 14.3 *Let X and Y be two random variables with generating functions $M_X(t)$ and $M_Y(t)$. If the random variables X and Y are independent, then*

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

for all t in a neighborhood of $t = 0$.

The proof is easy. If X and Y are independent, then the random variables e^{tX} and e^{tY} are independent for any fixed t (see Rule 9.5). Since $E(UV) = E(U)E(V)$ for independent random variables U and V, it follows that

$$E\left[e^{t(X+Y)}\right] = E[e^{tX}e^{tY}] = E(e^{tX})E(e^{tY}).$$

Example 14.4 Suppose that the random variable X has an $N(\mu, \sigma^2)$ distribution. Then

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}, \qquad -\infty < t < \infty.$$

The derivation is as follows. Let $Z = (X - \mu)/\sigma$. Then, Z has the $N(0, 1)$ distribution and

$$\begin{aligned} M_Z(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2}x^2} dx = e^{\frac{1}{2}t^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2} dx \\ &= e^{\frac{1}{2}t^2}, \end{aligned}$$

where the last equality uses the fact that for fixed t the function $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2}$ is the probability density function of an $N(t, 1)$ distribution. This implies that the integral of this function over the interval $(-\infty, \infty)$ equals 1. The desired expression for $M_X(t)$ next follows from

$$E(e^{tX}) = E\left(e^{t(\mu+\sigma Z)}\right) = e^{t\mu} E(e^{t\sigma Z}) = e^{t\mu} e^{\frac{1}{2}\sigma^2 t^2}.$$

The first and the second derivatives of $M_X(t)$ at the point $t = 0$ are given by $M'_X(0) = \mu$ and $M''_X(0) = \mu^2 + \sigma^2$, showing that the expected value and variance of an $N(\mu, \sigma^2)$-distributed random variable are indeed equal to μ and σ^2.

Remark 14.1 The moment-generating function $M_X(t)$ of the normal distribution enables us also to derive the expected value and the variance of the lognormal distribution. If X is $N(\mu, \sigma^2)$ distributed, then $Y = e^X$ has a lognormal distribution with parameters μ and σ. Taking $t = 1$ in the moment-generating function $M_X(t) = E(e^{tX})$, we obtain $E(Y)$. Also, by $e^{2X} = Y^2$, we obtain $E(Y^2)$ by putting $t = 2$ in $M_X(t) = E(e^{tX})$.

Using the result of Example 14.4, we easily verify that a linear combination of independent normal variates has, again, a normal distribution.

Rule 14.4 *Suppose that the random variables $X_1, \ldots, X_n$ are independent and normally distributed, where X_i has an $N(\mu_i, \sigma_i^2)$ distribution. Then, for any constants $a_1, \ldots, a_n$, the random variable $U = a_1X_1 + \cdots + a_nX_n$ has an $N(\mu, \sigma^2)$ distribution with*

$$\mu = a_1\mu_1 + \cdots + a_n\mu_n \quad \textit{and} \quad \sigma^2 = a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2.$$

It suffices to prove this result for $n = 2$. Next the general result follows by induction. Using Rule 14.3 and the result from Example 14.4, we find

$$\begin{aligned} E\left[e^{t(a_1X_1+a_2X_2)}\right] &= E(e^{ta_1X_1})E(e^{ta_2X_2}) \\ &= e^{\mu_1 a_1 t + \frac{1}{2}\sigma_1^2(a_1t)^2} e^{\mu_2 a_2 t + \frac{1}{2}\sigma_2^2(a_2t)^2} \\ &= e^{(a_1\mu_1 + a_2\mu_2)t + \frac{1}{2}(a_1^2\sigma_1^2 + a_2^2\sigma_2^2)t^2}, \end{aligned}$$

proving the desired result with an appeal to the uniqueness Rule 14.2.

The above example shows that the class of normal distributions is closed. A similar result can be shown for the class of gamma distributions (see Problem 14.13).

Example 14.5 Suppose that the random variable X has a gamma distribution with shape parameter α and scale parameter λ. Then

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha}, \qquad t < \lambda.$$

To verify this result, fix t with $t < \lambda$ and note that

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\, dx = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x}\, dx \\ &= \left(\frac{\lambda}{\lambda - t}\right)^{\alpha} \int_0^\infty \frac{(\lambda - t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x}\, dx. \end{aligned}$$

Using the fact that $(\lambda - t)^\alpha x^{\alpha-1} e^{-(\lambda-t)x} / \Gamma(\alpha)$ is a gamma density for any fixed t with $t < \lambda$ and thus integrates to 1, the desired result follows.

Rule 14.5 *Let $Z_1, \ldots, Z_n$ be independent random variables each having a standard normal distribution. Define the so-called chi-squared-distributed random* *variable U* by $U = Z_1^2 + \cdots + Z_n^2$. *Then, the random variable U has a gamma density with shape parameter $\frac{1}{2}n$ and scale parameter* 1.

Using the moment-generating function approach, this result is easily verified. Letting Z be an $N(0, 1)$ random variable, it follows that

$$\begin{aligned} E\big(e^{tZ^2}\big) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx^2} e^{-\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2t)x^2}\,dx \\ &= \frac{1}{\sqrt{1-2t}} \frac{1}{(1/\sqrt{1-2t})\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2/(1/\sqrt{1-2t})^2}\,dx \\ &= \frac{1}{\sqrt{1-2t}}, \qquad t < \frac{1}{2}. \end{aligned}$$

Next, by applying Rule 14.3

$$M_U(t) = \frac{1}{\sqrt{1-2t}} \cdots \frac{1}{\sqrt{1-2t}} = \frac{1}{(1-2t)^{n/2}}, \qquad t < \frac{1}{2}.$$

Comparing this expression with the moment-generating function of the gamma density in Example 14.5 and using the uniqueness Rule 14.2, it follows that the the chi-squared-distributed random variable U has a gamma density with shape parameter $\frac{1}{2}n$ and scale parameter 1.

Problem 14.12 Determine the moment-generating function of the random variable X having as density function the so-called *Laplace* density function $f(x) = \frac{1}{2}ae^{-a|x|}$ for $-\infty < x < \infty$, where a is a positive constant. Use the moment-generating function of X to find $E(X)$ and $\text{var}(X)$.

Problem 14.13 Let $X_1, \ldots, X_n$ be independent random variables each having a gamma density with the same scale parameter β. Denote by α_i the shape parameter of the gamma density of the X_i. Verify that $X_1 + \cdots + X_n$ has a gamma density with shape parameter $\alpha_1 + \cdots + \alpha_n$ and scale parameter β.

Problem 14.14 The random variables X and Y are independent and identically distributed. Use the moment-generating function to prove that X and Y are normally distributed if $X + Y$ has a normal distribution. *Remark*: the assumption that X and Y are identically distributed can be dropped, but this requires deep analysis.

Problem 14.15 The moment-generating function of two jointly distributed random variables X and Y is defined by $M_{X,Y}(v, w) = E(e^{vX+wY})$, provided that this integral is finite for all (v, w) in a neighborhood of (0, 0). A basic result is that $M_{X,Y}(v, w)$ uniquely determines the joint distribution of X and Y.[†]

[†] Using this uniqueness result, it is not difficult to verify that X and Y are independent if and only if $M_{X,Y}(v, w) = M_X(v)M_Y(w)$ for all v, w.

(a) What is $M_{X,Y}(v, w)$ if (X, Y) has a bivariate normal distribution?

(b) Suppose that the jointly distributed random variables X and Y have the property that $vX + wY$ is normally distributed for any constants v and w. Prove that (X, Y) has a bivariate normal density.

Problem 14.16 Consider Problem 14.5 again, but assume now that the individual claim sizes have a gamma density with shape parameter α and scale parameter λ. What is the moment-generating function of the total claim size?

14.2.1 The Chernoff bound

Let X be a random variable for which the moment-generating function $M_X(t)$ exists. The so-called *Chernoff bound* states that

$$P(X \geq c) \leq \min_{t>0}[e^{-ct}M_X(t)] \qquad \text{for any constant } c,$$

where the minimum is taken over all $t > 0$ for which $M_X(t)$ is finite. This is a very useful bound for tail probabilities.

The proof of the Chernoff bound is very simple. The bound follows directly from *Markov's inequality*, which states that

$$P(U \geq a) \leq \frac{1}{a}E(U) \qquad \text{for any constant } a > 0$$

when U is a nonnegative random variable. Apply Markov's inequality with $U = e^{tX}$ and $a = e^{ct} > 0$ and use the fact that

$$P(X \geq c) = P(tX \geq tc) = P(e^{tX} \geq e^{tc}) \qquad \text{for any } t > 0.$$

This gives $P(X \geq c) \leq e^{-ct}M_X(t)$ for any $t > 0$, implying the desired result. For its part, Markov's inequality is simply proved. For fixed $a > 0$, define the indicator variable I as equal to 1 if $U \geq a$ and 0 otherwise. Then, by $U \geq aI$ and $E(I) = P(U \geq a)$, it follows that $E(U) \geq aP(U \geq a)$. The Chernoff bound is more powerful than Chebyshev's inequality from Section 5.2. This inequality states that

$$P(|X - E(X)| \geq c) \leq \frac{\sigma^2(X)}{c^2} \qquad \text{for any constant } c > 0.$$

This bound can also be obtained directly from Markov's inequality by taking $U = (X - \mu)^2$ and $a = c^2$

$$P(|X - \mu| \geq c) = P((X - \mu)^2 \geq c^2) \leq \frac{E(X - \mu)^2}{c^2} = \frac{\sigma^2(X)}{c^2}.$$

Example 14.6 If the random variable X has the standard normal distribution, then the Chernoff bound implies that

$$P(X \geq c) \leq e^{-\frac{1}{2}c^2} \qquad \text{for any constant } c > 0.$$

To see this result, note that the minimizing value of t in the Chernoff bound $e^{-ct}e^{\frac{1}{2}t^2}$ follows by putting the derivative of $\frac{1}{2}t^2 - ct$ equal to zero. This gives $t = c$ for any positive value of the constant c. Substituting $t = c$ into the bound yields the desired result. The Chernoff bound is much sharper than the Chebyshev bound $\frac{1}{2c^2}$. For example, $P(X \geq c)$ with $c = 5$ has the exact value 2.87×10^{-7} and the Chernoff bound is 3.73×10^{-6}.

Problem 14.17 Prove that $P(X \leq c) \leq \min_{t<0}[e^{-ct}M_X(t)]$ for any constant c, assuming that $M_X(t)$ exists.

Problem 14.18 Let the random variable X be the number of successes in n independent Bernoulli trials with success probability p. Choose any $\delta > 0$ such that $(1+\delta)p < 1$. Use the Chernoff bound to verify that

$$P(X \geq (1+\delta)np) \leq \left[\left(\frac{p}{a}\right)^a \left(\frac{1-p}{1-a}\right)^{1-a}\right]^n,$$

where $a = (1+\delta)p$. (*Remark*: the upper bound can be shown to be smaller than or equal to $e^{-2p^2\delta^2 n}$).

14.2.2 The strong law of large numbers

The strong law of large numbers is one of the pillars of probability theory. Under the assumption that the moment-generating function exists, this law can be derived from the Chernoff bound and the Borel-Cantelli lemma. Let $X_1, X_2, \ldots$ be a sequence of independent random variables that have the same distribution as the random variable X. Denote by μ the expected value of X. The strong law of large numbers states that

$$P\left(\left\{\omega : \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k(\omega) = \mu\right\}\right) = 1,$$

where the symbol ω represents an outcome in the underlying sample space on which the process $\{X_1, X_2, \ldots\}$ is defined. To prove this result, fix $\epsilon > 0$ and note that

$$P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right| \geq \epsilon\right) = P\left(\frac{1}{n}\sum_{k=1}^{n} X_k \geq \mu + \epsilon\right) + P\left(\frac{1}{n}\sum_{k=1}^{n} X_k \leq \mu - \epsilon\right).$$

Let us now assume that the moment-generating function $M_X(t)$ of the random variable X exists. The moment-generating function of $X_1 + \cdots + X_n$ is given by $[M_X(t)]^n$. Using the Chernoff bound, we find

$$P\left(\frac{1}{n}\sum_{k=1}^{n} X_k \geq \mu + \epsilon\right) \leq \min_{t>0} e^{-n(\mu+\epsilon)t}[M_X(t)]^n = \min_{t>0}\left[e^{-(\mu+\epsilon)t}M_X(t)\right]^n$$

for $n = 1, 2, \ldots$. Letting $\lambda = \min_{t>0} e^{-(\mu+\epsilon)t}M_X(t)$, we prove below that $0 \leq \lambda < 1$. This result implies that

$$P\left(\frac{1}{n}\sum_{k=1}^{n} X_k \geq \mu + \epsilon\right) \leq \lambda^n \qquad \text{for } n = 1, 2, \ldots.$$

To verify that $0 \leq \lambda < 1$, let $G(t) = e^{-(\mu+\epsilon)t}M_X(t)$. Obviously, $G(0) = 1$. The derivative of $G(t)$ is given by $-(\mu + \epsilon)e^{-(\mu+\epsilon)t}M_X(t) + e^{-(\mu+\epsilon)t}M_X'(t)$. Since $M_X(0) = 1$ and $M_X'(0) = \mu$, we see that $G'(0) = -(\mu + \epsilon) + \mu = -\epsilon < 0$. This proves that the nonnegative function $G(t)$ is decreasing in $t = 0$ and so $G(t_0) < 1$ for some $t_0 > 0$, showing that $0 \leq \lambda < 1$. In the same way it can be verified that $0 \leq \eta < 1$, where $\eta = \min_{t<0} e^{-(\mu-\epsilon)t}M_X(t)$. Using this result together with the Chernoff bound from Problem 14.17, we obtain

$$P\left(\frac{1}{n}\sum_{k=1}^{n} X_k \leq \mu - \epsilon\right) \leq \eta^n \qquad \text{for } n = 1, 2, \ldots.$$

Thus, letting $A_n = \{\omega : |\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) - \mu| \geq \epsilon\}$, we have

$$\sum_{n=1}^{\infty} P(A_n) \leq \sum_{n=1}^{\infty}(\lambda^n + \eta^n) = \frac{\lambda}{1-\lambda} + \frac{\eta}{1-\eta} < \infty.$$

Invoking now the Borel-Cantelli lemma (see Problem 7.11 in Chapter 7), we find that $P(C_\epsilon) = 0$, where

$$C_\epsilon = \left\{\omega : \left|\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) - \mu\right| \geq \epsilon \text{ for infinitely many values of } n\right\}.$$

In other words, letting

$$\overline{C_\epsilon} = \left\{\omega : \left|\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) - \mu\right| < \epsilon \text{ for all } n \text{ sufficiently large}\right\},$$

we have $P(\overline{C_\epsilon}) = 1$. This result holds for any $\epsilon > 0$. The set $\overline{C_\epsilon}$ decreases as ϵ gets smaller. Taking a decreasing sequence $(\epsilon_k, k \geq 1)$ with $\lim_{k\to\infty} \epsilon_k = 0$

and using Rule 7.2 from Chapter 7, we find that

$$P(\lim_{\epsilon \to 0} \overline{C_\epsilon}) = \lim_{\epsilon \to 0} P(\overline{C_\epsilon}).$$

Since $\lim_{\epsilon \to 0} \overline{C_\epsilon}$ is the set $\{\omega : \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} X_k(\omega) = \mu\}$, we obtain that $P(\{\omega : \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} X_k(\omega) = \mu\}) = 1$, as was to be proved. The proof used the assumption that the moment-generating function of X exists. However, this assumption can be weakened to the assumption that the expected value of X exists.

As the above proof clearly demonstrates, we can conclude from the *strong* law of large numbers that, now matter how small $\epsilon > 0$ is chosen, eventually the sample mean $\frac{1}{n} \sum_{k=1}^{n} X_k$ gets within a distance ϵ from μ and stays within this bandwidth. This conclusion cannot be drawn from the so-called *weak* law of large numbers, which says that

$$\lim_{n \to \infty} P\left(\left| \frac{1}{n} \sum_{k=1}^{n} X_k - \mu \right| \geq \epsilon \right) = 0$$

for any $\epsilon > 0$. The weak law states only that for any specified large value of n the random variable $\frac{1}{n} \sum_{k=1}^{n} X_k$ is likely to be near μ.[†] The proof of the weak law of large numbers is much simpler than that of the strong law. The weak law can be directly obtained from Chebyshev's inequality when it is assumed that the variance of the random variables X_k is finite (verify!).

14.2.3 The central limit theorem revisited

We cannot end this book without offering at least a glimpse of the steps involved in the proof of the central limit theorem, which plays such a prominent role in probability theory. The mathematical formulation of the central limit theorem is as follows. Suppose that $X_1, X_2, \ldots$ are independent and identically distributed random variables with expected value μ and standard deviation σ. Then

$$\lim_{n \to \infty} P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma \sqrt{n}} \leq x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} \, dy \qquad \text{for all } x.$$

We make this result plausible for the case that the moment-generating function of the X_i exists and is finite for all t in some neighborhood of $t = 0$. To do so, consider the standardized variables

$$U_i = \frac{X_i - \mu}{\sigma}, \qquad i = 1, 2, \ldots.$$

[†] The central limit theorem sharpens this result: $P(|\frac{1}{n} \sum_{k=1}^{n} X_k - \mu| \geq \epsilon) \approx 2[1 - \Phi(\epsilon\sqrt{n}/\sigma)]$ for any specified large value of n, where σ is the standard deviation of the X_k.

Then $E(U_i) = 0$ and $\sigma(U_i) = 1$. Letting

$$Z_n = \frac{U_1 + \cdots + U_n}{\sqrt{n}},$$

we have $Z_n = (X_1 + \cdots + X_n - n\mu)/\sigma\sqrt{n}$. Denoting by $M_{Z_n}(t) = E(e^{tZ_n})$ the moment-generating function of Z_n, it will be proved in a moment that

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{\frac{1}{2}t^2}$$

for all t in a neighborhood of $t = 0$. In other words

$$\lim_{n\to\infty} M_{Z_n}(t) = E(e^{tZ})$$

when Z is a standard normal random variable. From this result we can conclude that

$$\lim_{n\to\infty} P(Z_n \le x) = P(Z \le x) \qquad \text{for all } x,$$

using a deep continuity theorem for moment-generating functions. This theorem linking the convergence of moment-generating functions to convergence of probability distribution functions must be taken for granted by the reader.

To verify that the moment-generating function of Z_n converges to the moment-generating function of the standard normal random variable, let $M_U(t)$ be the moment-generating function of the U_i. Using the assumption that $U_1, \ldots, U_n$ are independent and identically distributed, it follows that

$$E(e^{tZ_n}) = E\left(e^{t(U_1+\cdots+U_n)/\sqrt{n}}\right) = E\left(e^{(t/\sqrt{n})U_1}\right)\cdots E\left(e^{(t/\sqrt{n})U_n}\right)$$

and so

$$M_{Z_n}(t) = \left[M_U(t/\sqrt{n})\right]^n, \qquad n = 1, 2, \ldots.$$

Since $M_U(t) = 1 + t\,E(U_1) + \frac{t^2}{2!}\,E(U_1^2) + \frac{t^3}{3!}\,E(U_1^3) + \cdots$ in some neighborhood of $t = 0$ and using the fact that $E(U_1) = 0$ and $\sigma(U_1) = 1$, it follows that

$$M_U(t) = 1 + \frac{1}{2}t^2 + \epsilon(t)$$

in a neighborhood of $t = 0$, where $\epsilon(t)$ tends faster to zero than t^2 as $t \to 0$. That is

$$\lim_{t\to 0} \frac{\epsilon(t)}{t^2} = 0.$$

Now fix t and let $\epsilon_n = \epsilon(t/\sqrt{n})$. Then

$$M_{Z_n}(t) = \left(1 + \frac{1}{2}\frac{t^2}{n} + \frac{n\epsilon_n}{n}\right)^n, \qquad n = 1, 2, \ldots.$$

Since $\lim_{u\to 0}\epsilon(u)/u^2 = 0$, we have that $\lim_{n\to\infty} n\epsilon_n = 0$. Using the fact that $\lim_{n\to\infty}(1+\frac{a}{n})^n = e^a$ for any constant a, it is now a matter of standard manipulation in analysis to conclude that

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{\frac{1}{2}t^2},$$

as was to be proved.

15
Markov chains

In previous chapters we have dealt with sequences of independent random variables. However, many random systems evolving in time involve sequences of dependent random variables. Think of the outside weather temperature on successive days, or the prize of IBM stock at the end of successive trading days. For many such systems it is reasonable to assume that the probability of going from one state to another state depends only on the current state of the system and thus is not influenced by additional information about past states. The probability model with this feature is called a Markov chain. The concepts of state and state transition are at the heart of Markov chain analysis. The line of thinking through the concepts of state and state transition is very useful for analyzing many practical problems in applied probability.

Markov chains are named after the Russian mathematician Andrey Markov (1856-1922), who first developed this probability model in order to analyze the alternation of vowels and consonants in Pushkin's poem "Eugine Onegin." His work helped to launch the modern theory of stochastic processes (a *stochastic process* is a collection of random variables, indexed by an ordered time variable). The characteristic property of a Markov chain is that its memory goes back only to the most recent state. Knowledge of the current state only is sufficient to describe the future development of the process. A Markov model is the simplest model for random systems evolving in time when the successive states of the system are not independent. But this model is no exception to the rule that simple models are often the most useful models for analyzing practical problems. The theory of Markov chains has applications to a wide variety of fields, including biology, physics, engineering, and computer science.

In this chapter we only consider Markov chains with a finite number of states. We first present techniques to analyze the transient behavior of Markov chains. In particular, we give much attention to Markov chains with one or more absorbing states. Such Markov chains have interesting applications in the analysis of success runs. Finally, we deal with the long-run behavior of Markov chains and give solution methods to answer questions such as: what is the long-run proportion of time that the system will be in any given subset of states.

15.1 Markov model

A Markov chain deals with a collection of random variables, indexed by an ordered time parameter. The Markov model is the simplest conceivable generalization of a sequence of independent random variables. A Markov chain is a sequence of trials having the property that the outcome of each last trial provides enough information to predict the outcome of any future trial. Despite its very simple structure, the Markov model is extremely useful in a wide variety of practical probability problems. The beginning student often has difficulties in grasping the concept of the Markov chain when a formal definition is given. Let's begin with an example that illustrates the essence of what a Markov process is.

Example 15.1 A drunkard wanders about a town square. At each step he no longer remembers the direction of his previous step. Each step is a unit distance in a randomly chosen direction and has equal probability $\frac{1}{4}$ of going north, south, east or west as long as the drunkard has not reached the edge of the square (see Figure 15.1). The drunkard never leaves the square. Should he reach the boundary of the square, his next step is equally likely to be in one of the three remaining directions if he is not at a corner point, and is equally likely to be in one of the two remaining directions otherwise. The drunkard starts in the middle of the square. What stochastic process describes the drunkard's walk?

Solution. To answer this question, we define the random variable X_n as

$$X_n = \text{the position of the drunkard just after the } n\text{th step}$$

for $n = 0, 1, \ldots$ with the convention $X_0 = (0, 0)$. We say that the drunkard is in state (x, y) when the current position of the drunkard is described by the point (x, y). The collection $\{X_0, X_1, \ldots\}$ of random variables is a stochastic process with discrete time-parameter and finite state space

$$I = \{(x, y) : x, y \text{ integer and } - L \leq x, y \leq L\},$$

where L is the distance from the middle of the square to its boundary. The successive states of the drunkard are not independent of each other, but the

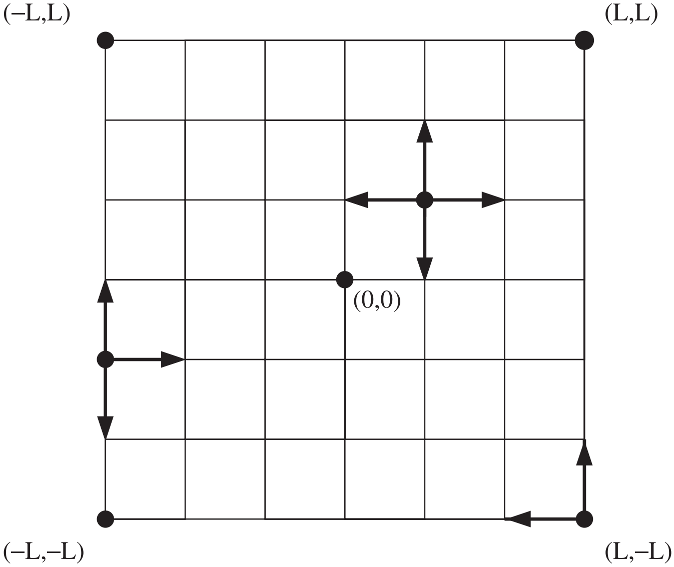


Fig. 15.1. The drunkard's walk.

next position of the drunkard depends only on his current position and is not influenced by the earlier positions in his path. That is, the process $\{X_0, X_1, \ldots\}$ has the so-called *Markovian property*, which says that the state at any given time summarizes everything about the past that is relevant to the future.

Many random systems evolving over time can be modeled to satisfy the Markovian property. Having this property introduced informally, we are now ready to give a formal definition of a Markov chain. Let $X_0, X_1, \ldots$ be a sequence of random variables. It is helpful to think of X_n as the state of a dynamic system at time $t = n$. In the sequel, the set of possible values of the random variables X_n is assumed to be *finite* and is denoted by I. The set I is called the *state space* of the stochastic process $\{X_0, X_1, \ldots\}$.

Definition 15.1 *The stochastic process* $\{X_n, n = 0, 1, \ldots\}$ *with state space* I *is said to be a (discrete-time) Markov chain if it possesses the Markovian property, that is, for each time point* $n = 0, 1, \ldots$ *and all possible values of the states* $i_0, i_1, \ldots, i_{n+1} \in I$, *the process has the property*

$$\begin{aligned}P(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n)\\ = P(X_{n+1} = i_{n+1} \mid X_n = i_n).\end{aligned}$$

The term $P(X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n)$ should be read as follows: it is the conditional probability that the system will be in state i_{n+1} at the *next* time point $t = n + 1$ if the system is in state i_n at the *current* time $t = n$ and has reached the current state i_n via the states $i_0, i_1, \ldots, i_{n-1}$ at the *past* time points $t = 0, 1, \ldots, n - 1$. The Markovian property says that this conditional probability depends only on the current state i_n and is not altered by knowledge of the past states $i_0, i_1, \ldots, i_{n-1}$. The current state summarizes everything about the past that is relevant to the future.

In Example 15.1 the Markovian property was satisfied in a natural way by choosing the state of the process as the position of the drunkard on the square. However, in other applications the choice of the state variable(s) may require more thought in order to satisfy the Markovian property. To illustrate this, consider Example 15.1 again and assume now that the drunkard never chooses the same direction as was chosen in the previous step. Then, we need an extra state variable in order to satisfy the Markovian property. Let's say that the drunkard is in state (x, y, N) when the position of the drunkard on the square is (x, y) and he moved northward in his previous step. Similarly, the states (x, y, E), (x, y, S) and (x, y, W) are defined. Letting X_n be the state of the drunkard after the nth step (with the convention $X_0 = (0, 0)$), the stochastic process $\{X_0, X_1, \ldots\}$ satisfies the Markovian property and thus is a Markov chain. The transition probabilities are easy to give. For example, if the current state of the process is (x, y, S) with (x, y) an interior point of the square, the next state of the process is equally likely to be one of the three states $(x + 1, y, E)$, $(x - 1, y, W)$, and $(x, y + 1, N)$. In the drunkard's walk the concepts of *state* and *state transition* come up in a natural way. These concepts are at the heart of Markov chain analysis.

In the following, we will restrict our attention to time-homogeneous Markov chains. For such chains the transition probability $P(X_{n+1} = j \mid X_n = i)$ does not depend on the value of the time parameter n and so $P(X_{n+1} = j \mid X_n = i) = P(X_1 = j \mid X_0 = i)$ for all n. We write

$$p_{ij} = P(X_{n+1} = j \mid X_n = i).$$

The probabilities p_{ij} are called the *one-step transition probabilities* of the Markov chain and are the same for all time points n. They satisfy

$$p_{ij} \geq 0 \qquad \text{for } i, j \in I \qquad \text{and} \qquad \sum_{j \in I} p_{ij} = 1 \qquad \text{for all } i \in I.$$

The notation p_{ij} is sometimes confusing for the beginning student: p_{ij} is not a joint probability, but a conditional probability. However, the notation p_{ij} rather than the notation $p(j \mid i)$ has found widespread acceptance.

A Markov chain $\{X_n, n = 0, 1, \ldots\}$ is completely determined by the probability distribution of the initial state X_0 and the one-step transition probabilities p_{ij}. In applications of Markov chains the art is:

(a) to choose the state variable(s) such that the Markovian property holds
(b) to determine the one-step transition probabilities p_{ij}.

How to formulate a Markov chain model for a concrete problem is largely an art that is developed with practice. Putting yourselves in the shoes of someone who has to write a simulation program for the problem in question may be helpful in choosing the state variable(s). Once the (difficult) modeling step is done, the rest is simply a matter of applying the theory that will be developed in the next sections. The student cannot be urged strongly enough to try the problems at the end of this section to acquire skills to model new situations. In order to help students develop intuition into how practical situations can be modeled as a Markov chain, we give three examples. The first example deals with the Ehrenfest model for gas diffusion. In physics the Ehrenfest model resolved at the beginning of the twentieth century a seeming contradiction between the second law of thermodynamics and the laws of mechanics.

Example 15.2 Two compartments A and B together contain r particles. With the passage of every time unit, one of the particles is selected at random and is removed from its compartment to the other. What stochastic process describes the contents of the compartments?

Solution. Let us take as state of the system the number of particles in compartment A. If compartment A contains i particles, then compartment B contains $r - i$ particles. Define the random variable X_n as

$$X_n = \text{the number of particles in compartment } A \text{ after the } n\text{th transfer.}$$

By the physical construction of the model with independent selections of a particle, the process $\{X_n\}$ satisfies the Markovian property and thus is a Markov chain. The state space is $I = \{0, 1, \ldots, r\}$. The probability of going from state i to state j in one step is zero unless $|i - j| = 1$. The one-step transition probability $p_{i,i+1}$ translates into the probability that the randomly selected particle belongs to compartment B and $p_{i,i-1}$ translates into the probability that the randomly selected particle belongs to compartment A. Thus, for $1 \leq i \leq r - 1$

$$p_{i,i+1} = \frac{r-i}{r} \quad \text{and} \quad p_{i,i-1} = \frac{i}{r}.$$

Further, $p_{01} = p_{r,r-1} = 1$. The other p_{ij} are zero.

Example 15.3 An absent-minded professor drives every morning from his home to the office and at the end of the day from the office to home. At any given time, his driver's license is located at his home or at the office. If his driver's license is at his location of departure, he takes it with him with probability 0.5. What stochastic process describes whether the professor has the driver's license with him when driving his car to home or to the office?

Solution. Your first thought might be to define two states 1 and 0, where state 1 describes the situation that the professor has his driver's license with him when driving his car and state 0 describes the situation that he does not have his driver's license with him when driving his car. However, these two states do not suffice for a Markov model: state 0 does not provide enough information to predict the state at the next drive. In order to give the probability distribution of this next state, you need information about the current location of the driver's license of the professor. You get a Markov model by simply inserting this information into the state description. Let's say that the system is in state 1 if the professor is driving his car and has his driver's license with him, in state 2 if the professor is driving his car and his driver's license is at the point of departure, and in state 3 if the professor is driving his car and his driver's license is at his destination. Define the random variable X_n as

$$X_n = \text{the state at the } n\text{th drive to home or to the office.}$$

The process $\{X_n\}$ has the property that any present state contains sufficient information for predicting future states. Thus, the process $\{X_n\}$ is a Markov chain with state space $I = \{1, 2, 3\}$. Next, we determine the p_{ij}. For example, p_{32} translates into the probability that the professor will not have his driver's license with him at the next drive given that his driver's license is at the point of departure for the next drive. This gives $p_{32} = 0.5$. Also, $p_{31} = 0.5$. Similarly, $p_{23} = 1$ and $p_{11} = p_{12} = 0.5$. The other p_{ij} are zero.

The next example deals with an inventory problem and the modeling of this problem is more involved than that of the previous three examples.

Example 15.4 The Johnson hardware shop, a family business since 1888, carries adjustable pliers as a regular stock item. The demand for this tool is stable over time. The total demand during a week has a Poisson distribution with expected value $\lambda = 4$. The demands in the successive weeks are independent of each other. Each demand that occurs when the shop is out of stock is lost. The owner of the shop uses a so-called periodic review (s, S) control rule with $s = 5$ and $S = 10$ for stock replenishment of the item. Under this rule the inventory position is only reviewed at the beginning of each week.

If upon review the stock on hand is less than the reorder point s, the inventory is replenished to the order-up-point S; otherwise, no ordering is done. The replenishment time is negligible. What stochastic process describes the stock on hand?

Solution. In this application we take as state variable the stock on hand just prior to review (another possible choice would have been the stock on hand just after review). Let the random variable X_n be defined as

$$X_n = \text{the stock on hand at the beginning of the } n\text{th week just prior to review.}$$

It will be immediately clear that the stochastic process $\{X_n\}$ satisfies the Markovian property: the stock on hand at the beginning of the current week and the demand in the coming week determine the stock on hand at the beginning of the next week. It is not relevant how the stock fluctuated in the past. Hence, the process $\{X_n\}$ is a Markov chain. Its state space is finite and is given by $I = \{0, 1, \ldots, S\}$. How do you find the one-step transition probabilities p_{ij}? In any application the simple but useful advice to you is to translate $P(X_{n+1} = j \mid X_n = i)$ in terms of the concrete situation you are dealing with. For example, how to find p_{0j} in the present application? If state 0 is the current state, then the inventory is replenished to level S, and the stock at the beginning of next week just prior to review will be j only if the demand in the coming week will be equal to $S - j$, provided that $j \neq 0$. The next state will be $j = 0$ only if the demand in the coming week will be S or more. Armed with this argument, we now specify the p_{ij}. We distinguish between the cases (a) $i < s$ and (b) $i \geq s$. In case (a) the stock on hand just after review is S, while in case (b) the stock on hand just after review is i.
Case (a): $i < s$. Then

$$\begin{aligned} p_{ij} &= P(\text{the demand in the next week will be equal to } S - j) \\ &= e^{-\lambda} \frac{\lambda^{S-j}}{(S-j)!} \qquad \text{for } 1 \leq j \leq S, \end{aligned}$$

regardless of the value of $i < s$. Further

$$\begin{aligned} p_{i0} &= P(\text{the demand in the next week will be } S \text{ or more}) \\ &= \sum_{k=S}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!}. \end{aligned}$$

Note that the expression for p_{i0} is in agreement with $p_{i0} = 1 - \sum_{j \neq 0} p_{ij}$. Case (b): $i \geq s$. Then

$$\begin{aligned} p_{ij} &= P(\text{the demand in the next week will be equal to } i - j) \\ &= e^{-\lambda} \frac{\lambda^{i-j}}{(i-j)!} \qquad \text{for } 1 \leq j \leq i \end{aligned}$$

and $p_{ij} = 0$ for $j > i$. Further

$$\begin{aligned} p_{i0} &= P(\text{ the demand in the next week will be } i \text{ or more}) \\ &= \sum_{k=i}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!}. \end{aligned}$$

Problem 15.1 Two compartments A and B each contain r particles. Of these $2r$ particles, r are of type 1 and r are of type 2. With the passing of every time unit, one particle is selected at random from each of the compartments, and each of these two particles is transferred from its compartment to the other one. What stochastic process describes the numbers of type 1 and type 2 particles in each of the two compartments?

Problem 15.2 Consider the following modification of Example 15.3. In the case that the driver's license of the professor is at his point of departure, he takes it with him with probability 0.75 when departing from home and with probability 0.5 when departing from the office. Define a Markov chain that describes whether the professor has the driving license with him when driving his car. Specify the one-step transition probabilities.

Problem 15.3 Let $\{X_n,\ n = 0, 1, \ldots\}$ be a Markov chain. Define the random variables Y_n and U_n by $Y_n = X_{2n}$ and $U_n = |X_n|$. Do you think the processes $\{Y_n\}$ and $\{U_n\}$ are always Markov chains?

Problem 15.4 Every day, it is either sunny or rainy on Rainbow Island. The weather for the next day depends only on today's weather and yesterday's weather. If the last two days were sunny, it will be sunny on the next day with probability 0.9. This probability is 0.45 if the last two days were rainy. The next day will be sunny with probability 0.7 if today's weather is sunny and yesterday's weather was rainy. If today's weather is rainy and yesterday's weather was sunny, it will be sunny on the next day with probability 0.5. Define a Markov chain describing the weather on Rainbow Island and specify its one-step transition probabilities.

Problem 15.5 To improve the reliability of a production system, two identical production machines are connected parallel to one another. For the production process, only one of the two machines is needed. The other machine (if available) takes over when the machine currently in use needs revision. At the end of each production day the used machine is inspected. The probability of an inspection revealing the necessity of a revision of the machine is $\frac{1}{10}$, regardless how long the inspected machine has already been in uninterrupted use. A revision takes exactly two days. There are ample repair facilities so that the revision of a machine can start immediately. The production process must be stopped when both machines are in revision. Formulate an appropriate Markov chain to describe the functioning of the production system and specify the one-step transition probabilities. *Hint*: use an auxiliary state variable to indicate the remaining duration of a revision.

Problem 15.6 A control device contains two parallel circuit boards. Both circuit boards are switched on. The device operates properly as long as at least one of the circuit boards functions. Each circuit board is subject to random failure. The failure rate increases with the age of the circuit board. The circuit boards are identical and their lifetimes are independent. Let r_i denote the probability of a circuit board failing during the next week if the circuit board has functioned for the past i weeks. Past records of circuit boards give the failure function $r_0 = 0, r_1 = 0.05, r_2 = 0.07, r_3 = 0.12, r_4 = 0.25, r_5 = 0.50,$ and $r_6 = 1$. Any failed circuit board is replaced at the beginning of the following week. Also, any six-week-old circuit board is replaced. Formulate an appropriate Markov chain for the failure analysis of the device and specify the one-step transition probabilities.

Problem 15.7 A communication channel transmits messages one at a time, and transmission of a message can only start at the beginning of a time slot. The transmission time of any message is one time slot. However, each transmission can fail with a given probability $f = 0.05$. A failed transmission is tried again at the beginning of the next time slot. Newly arriving messages for transmission are temporarily stored in a finite buffer. The buffer has capacity for only $K = 10$ messages (excluding any message in transmission). The number of new messages arriving during any given time slot has a Poisson distribution with mean $\lambda = 5$. If a newly arriving message finds the buffer full, the message is lost. Formulate an appropriate Markov chain to describe the content of the buffer at the beginning of the time slots and specify its one-step transition probabilities.

15.2 Transient analysis of Markov chains

As said before, a Markov chain $\{X_n, n = 0, 1, \ldots\}$ is completely determined by its one-step transition probabilities p_{ij} and the probability distribution of the initial state X_0. The transient analysis of a Markov chain concerns the calculation of the so-called n-step transition probabilities. The probability of going from state i to state j in the next n transitions of the Markov chain is easily calculated from the one-step transition probabilities. For any $n = 1, 2, \ldots,$ the n-step transition probabilities $p_{ij}^{(n)}$ are defined by

$$p_{ij}^{(n)} = P(X_n = j \mid X_0 = i) \qquad \text{for } i, j \in I.$$

Note that $p_{ij}^{(1)} = p_{ij}$. A basic result is given in the following rule.

Rule 15.1 (**Chapman-Kolmogorov equations**) *For any* $n \geq 2$

$$p_{ij}^{(n)} = \sum_{k \in I} p_{ik}^{(n-1)} p_{kj} \qquad \textit{for all } i, j \in I.$$

This rule states that the probability of going from state i to state j in n transitions is obtained by summing the probabilities of the mutually exclusive events of going from state i to some state k in the first $n-1$ transitions and then going from state k to state j in the nth transition. A formal proof proceeds as follows. Using the law of conditional probabilities and invoking the Markovian property, we obtain

$$\begin{aligned}
p_{ij}^{(n)} &= P(X_n = j \mid X_0 = i)\\
&= \sum_{k \in I} P(X_n = j \mid X_0 = i, X_{n-1} = k) P(X_{n-1} = k \mid X_0 = i)\\
&= \sum_{k \in I} P(X_n = j \mid X_{n-1} = k) P(X_{n-1} = k \mid X_0 = i) = \sum_{k \in I} p_{kj} p_{ik}^{(n-1)},
\end{aligned}$$

where the last equality uses the assumption of time homogeneity.

It is convenient to write the result of Rule 15.1 in terms of matrices. Let

$$\mathbf{P} = (p_{ij})$$

be the matrix having the one-step transition probabilities p_{ij} as entries. If we let $\mathbf{P}^{(n)}$ denote the matrix of the n-step transition probabilities $p_{ij}^{(n)}$, Rule 15.1 asserts that $\mathbf{P}^{(n)} = \mathbf{P}^{(n-1)} \times \mathbf{P}$ for all $n \geq 2$. By iterating this formula and using the fact that $\mathbf{P}^{(1)} = \mathbf{P}$, we obtain

$$\mathbf{P}^{(n)} = \mathbf{P} \times \mathbf{P} \times \ldots \times \mathbf{P} = \mathbf{P}^n.$$

This gives us the following important result:

Rule 15.2 *The n-step transition probabilities $p_{ij}^{(n)}$ can be calculated as the entries in the matrix product* $\mathbf{P}^n$*, which is obtained by multiplying the matrix* $\mathbf{P}$ *by itself n times.*

Example 15.5 On the Island of Hope the weather each day is classified as sunny, cloudy, or rainy. The next day's weather depends only on today's weather and not on the weather of the previous days. If the present day is sunny, the next day will be sunny, cloudy or rainy with probabilities 0.70, 0.10 and 0.20. The transition probabilities for the weather are 0.50, 0.25 and 0.25 when the present day is cloudy and they are 0.40, 0.30 and 0.30 when the present day is rainy. What is the probability that it will be sunny three days from now if it is rainy today? What are the proportions of time the weather will be sunny, cloudy and rainy over a long period?

Solution. These questions can be answered by using a three-state Markov chain. Let's say that the weather is in state 1 if it is sunny, in state 2 if it is cloudy and in state 3 if it is rainy. Define the random variable X_n as the state of the weather on day n. The stochastic process $\{X_0, X_1, \ldots\}$ is then a Markov chain with state space $I = \{1, 2, 3\}$. The matrix $\mathbf{P}$ of one-step transition probabilities is given by

$$\begin{array}{c} \text{from}\backslash\text{to} \\ 1 \\ 2 \\ 3 \end{array} \quad \begin{array}{c} \begin{array}{ccc} \;1\; & \quad 2\quad & \;3\; \end{array} \\ \begin{pmatrix} 0.70 & 0.10 & 0.20 \\ 0.50 & 0.25 & 0.25 \\ 0.40 & 0.30 & 0.30 \end{pmatrix} \end{array}.$$

To find the probability of having sunny weather three days from now, we need the matrix product $\mathbf{P}^3$:

$$\mathbf{P}^3 = \begin{pmatrix} 0.6015000 & 0.1682500 & 0.2302500 \\ 0.5912500 & 0.1756250 & 0.2331250 \\ 0.5855000 & 0.1797500 & 0.2347500 \end{pmatrix}.$$

From this matrix you read off that the probability of having sunny weather three days from now is $p_{31}^{(3)} = 0.5855$ if it is rainy today. What is the probability distribution of the weather after many days? Intuitively, you expect that this probability distribution does not depend on the present state of the weather. This is indeed confirmed by the following calculations:

$$\mathbf{P}^5 = \begin{pmatrix} 0.5963113 & 0.1719806 & 0.2317081 \\ 0.5957781 & 0.1723641 & 0.2318578 \\ 0.5954788 & 0.1725794 & 0.2319419 \end{pmatrix}$$

$$\mathbf{P}^{12} = \begin{pmatrix} 0.5960265 & 0.1721854 & 0.2317881 \\ 0.5960265 & 0.1721854 & 0.2317881 \\ 0.5960265 & 0.1721854 & 0.2317881 \end{pmatrix} = \mathbf{P}^{13} = \mathbf{P}^{14} = \cdots .$$

That is, after 12 matrix multiplications the entries agree row-to-row to seven decimal places. You see that the weather after many days will be sunny, cloudy or rainy with probabilities 0.5960, 0.1722 and 0.2318, respectively. It will be clear that these limiting probabilities also give the proportions of time that the weather will be sunny, cloudy and rainy over a long period. In this example we have answered the question about the long-run behavior of the weather by computing sufficiently high powers of $\mathbf{P}^n$. A computationally better approach for the long-run behavior of the system will be discussed in Section 15.4.

An interesting and useful result is the following:

Rule 15.3 *For any two states* $i, j \in I$

$$E(\textit{number of visits to state } j \textit{ over the time points } t = 1, \ldots, n \mid X_0 = i)$$
$$= \sum_{t=1}^{n} p_{ij}^{(t)} \qquad \textit{for } n = 1, 2, \ldots .$$

The proof of this result is instructive. Fix $i, j \in I$. For any $t \geq 1$, let

$$I_t = \begin{cases} 1 & \text{if } X_t = j \\ 0 & \text{otherwise.} \end{cases}$$

The number of visits to state j over the time points $t = 1, \ldots, n$ is then given by the random variable $\sum_{t=1}^{n} I_t$. Using the observation that

$$\begin{aligned} E(I_t \mid X_0 = i) &= 1 \times P(I_t = 1 \mid X_0 = i) + 0 \times P(I_t = 0 \mid X_0 = i) \\ &= P(X_t = j \mid X_0 = i) = p_{ij}^{(t)}, \end{aligned}$$

we obtain $E(\sum_{t=1}^{n} I_t \mid X_0 = i) = \sum_{t=1}^{n} E(I_t \mid X_0 = i) = \sum_{t=1}^{n} p_{ij}^{(t)}$, proving the desired result.

As an illustration, consider Example 15.5 again. What is the expected value of the number of sunny days in the coming seven days when it is cloudy today? The answer is that this expected value is equal to $\sum_{t=1}^{7} p_{21}^{(t)}$ days. The value of this sum is calculated as 4.049.

Problem 15.8 A car rental agency rents cars at four locations. A rented car can be returned to any of the four locations. A car rented at location 1 will be returned to location 1 with probability 0.8, to location 2 with probability 0.1, to location 3 with probability 0, and to location 4 with probability 0.1. These probabilities have values 0.1, 0.7, 0.2, and 0 for cars rented at location 2, values 0.2, 0.1, 0.5, and 0.2 for cars rented at location 3, and the values 0, 0.2, 0.1, and 0.7 for cars rented at location 4. A particular car is currently at location 3. What is the probability that this car is back at location 3 after being rented out five times? What is the long-run frequency with which any given car is returned to location i for $i = 1, 2, 3, 4$?

Problem 15.9 Consider Problem 15.4 again. What is the probability of having sunny weather five days from now if it rained today and yesterday? What is the proportion of time it will be sunny over a very long period? What is the expected number of days it will be sunny in the next 14 days given that it rained the last two days?

Problem 15.10 A communication system is either in the on-state (state 1) or the off-state (state 0). Every millisecond the state of the system may change. An off-state is changed into an on-state with probability α and an on-state is changed into an off-state with probability β, where $0 < \alpha, \beta < 1$. Use induction to verify that the n-step transition probabilities of the Markov chain describing the state of the system satisfy

$$p_{00}^{(n)} = \frac{\beta}{\alpha + \beta} + \frac{\alpha(1 - \alpha - \beta)^n}{\alpha + \beta} \quad \text{and} \quad p_{11}^{(n)} = \frac{\alpha}{\alpha + \beta} + \frac{\beta(1 - \alpha - \beta)^n}{\alpha + \beta},$$

where $p_{01}^{(n)} = 1 - p_{00}^{(n)}$ and $p_{10}^{(n)} = 1 - p_{11}^{(n)}$. *Remark*: the reader familiar with linear algebra may verify this result from the eigenvalues and eigenvectors of the matrix of one-step transition probabilities.

Problem 15.11 A faulty digital video conferencing system has a clustering error pattern. If a bit is received correctly, the probability of receiving the next bit correctly is 0.999. This probability is only 0.1 if the last bit was received incorrectly. Suppose that the first transmitted bit is received correctly. What is the expected value of the number of incorrectly received bits among the next 5,000 bits?

Problem 15.12 Trees in a forest are assumed to fall into four age groups: baby trees (0-10 years of age), young trees (11-20 years of age), middle-aged trees (21-30 years of age), and old trees (older than 30 years of age). The length of one time period is 10 years. In each time period a certain percentage of trees in each age group dies. These percentages are 20%, 5%, 10%, and 25% for the four age groups. Lost trees are replaced by baby trees. Surviving trees enter the next age group, where old trees remain in the fourth age group. Suppose that the forest is newly planted with 10,000 trees. What is the age distribution of the forest after 50 years? What is the age distribution of the forest in the equilibrium situation?

Problem 15.13 Let $\{X_n\}$ be any Markov chain. For any i and j, define the random variable $V_{ij}(n)$ as the number of visits to state j over the time points $t = 1, 2, \ldots, n$ if the starting state is i. Verify the result

$$\sigma^2[V_{ij}(n)] = \sum_{t=1}^{n} p_{ij}^{(t)}\big(1 - p_{ij}^{(t)}\big) + 2\sum_{t=1}^{n} \sum_{u=t+1}^{n} \big[p_{ij}^{(t)} p_{jj}^{(u-t)} - p_{ij}^{(t)} p_{ij}^{(u)}\big].$$

Hint: use the fact that $P(I_t = 1, I_u = 1|X_0 = i) = p_{ij}^{(t)} p_{jj}^{(u-t)}$ for $u > t \geq 1$, where I_t is defined as in the proof of Rule 15.3.[†] Next, apply the result to Example 15.5 to approximate the probability of having more than 240 sunny days in the next 365 days given that it is rainy today.

15.3 Absorbing Markov chains

Markov chains can also be used to analyze systems in which some states are "absorbing." Once the system reaches an absorbing state, it remains in that state permanently. The Markov chain model with absorbing states has many interesting applications. Examples include stochastic models of biological populations where the absorbing state is extinction and gambling models where the absorbing state is ruin.

Let $\{X_n\}$ be a Markov chain with one-step probabilities p_{ij}. State i is said to be an *absorbing* state if $p_{ii} = 1$. The Markov chain $\{X_n\}$ is said to be an absorbing Markov chain if it has one or more absorbing states and the set of absorbing states is accessible from the other states. Interesting questions are: (a) How long will it take before the system hits an absorbing state, and (b) If there are multiple absorbing states, what is the probability that the system will end up in each of those absorbing states? We address these questions in the two examples below. The first example deals with a variant of the coupon collector's problem. This problem was discussed before in Sections 3.2 and 14.1. It is instructive to demonstrate how the probability distribution of the number of trials needed to collect all of the different types of coupons can be calculated through an absorbing Markov chain. The line of thinking through the concepts of state and state transition is very useful for analyzing this problem (and many other problems in applied probability!). It leads to an algorithmic

[†] It can be shown that for any i and j the random variable $V_{ij}(n)$ is approximately normally distributed for n sufficiently large when the Markov chain has the property that any state is accessible from any other state. A state k is said to be *accessible* from another state j if $p_{jk}^{(n)} > 0$ for some $n \geq 1$.

solution which tends to be at a more intuitive level than a neat closed-form solution.

Example 15.6 A fair die is rolled until each of the six possible outcomes $1, 2, \ldots, 6$ has appeared. How to calculate the probability mass function of the number of rolls needed?

Solution. Let's say that the system is in state i if i different outcomes have appeared so far. Define the random variable X_n as the state of the system after the nth roll. State 6 is taken as an absorbing state. The process $\{X_n\}$ is an absorbing Markov chain with state space $I = \{0, 1, \ldots, 6\}$. The matrix $\mathbf{P} = (p_{ij})$ of one-step transition probabilities is given by

$$p_{01} = 1, \quad p_{ii} = \frac{i}{6} \text{ and } p_{i,i+1} = 1 - \frac{i}{6} \text{ for } i = 1, \ldots, 5, \quad p_{66} = 1,$$

and $p_{ij} = 0$ otherwise. The starting state of the process is state 0. Let the random variable R denote the number of rolls of the die needed to obtain all of the six possible outcomes. The random variable R takes on a value larger than r only if the Markov chain has not visited the absorbing state 6 in the first r transitions. Hence

$$P(R > r) = P(X_k \neq 6 \text{ for } k = 1, \ldots, r \mid X_0 = 0).$$

However, since state 6 is absorbing, it automatically holds that $X_k \neq 6$ for any $k < r$ if $X_r \neq 6$. Hence

$$P(X_k \neq 6 \text{ for } k = 1, \ldots, r \mid X_0 = 0) = P(X_r \neq 6 \mid X_0 = 0).$$

Noting that $P(X_r \neq 6 \mid X_0 = 0) = 1 - P(X_r = 6 \mid X_0 = 0)$, we obtain

$$P(R > r) = 1 - p_{06}^{(r)} \qquad \text{for } r = 1, 2, \ldots.$$

In other words, the desired probability $P(R > r)$ can be calculated by multiplying the matrix $\mathbf{P}$ by itself r times. For example, $P(R > r)$ has the values 0.7282, 0.1520, and 0.0252 for r=10, 20, and 30. It is worthwhile to point out that $p_{0j}^{(n)}$ for $j = 1, \ldots, 6$ represents the probability of having j different outcomes after n rolls of the die. For example, $p_{0j}^{(10)}$ has the values 0.0000, 0.0003, 0.0185, 0.2031, 0.5064, and 0.2718 for j=1, 2, 3, 4, 5, and 6.

The next example shows that absorbing Markov chains are also very useful for analyzing success runs.

Example 15.7 Chess grandmaster Boris Karparoff has entered a competition with chess computer Deep Blue. The competition will continue until either Karparoff or Deep Blue has won two consecutive matches. For the first match as well as any match ending in a draw, it holds that the next match will be won by Karparoff with probability 0.4, by Deep Blue with probability 0.3, and will end in a draw with probability 0.3. After a win by Karparoff, the probabilities of these outcomes for the next match will have the values 0.5, 0.25, and 0.25, while after a loss by Karparoff the probabilities will have the values 0.3, 0.5, and 0.2. What is the probability that the competition will last for longer than ten games? What is the probability that Karparoff will be the final winner, and what is the expected value of the duration of the competition?

Solution. To answer these questions, we use an absorbing Markov chain with two absorbing states. Let's say that the system is in state $(1, K)$ if Karparoff has won the last game but not the game before, and in state $(2, K)$ if Karparoff has won the last two games. Similarly, the states $(1, D)$ and $(2, D)$ are defined. The system is said to be in state 0 if the match is about to begin or the last game is a draw. We take the states $(2, K)$ and $(2, D)$ as absorbing states. Define the random variable X_n as the state of the system after the nth game. The process $\{X_n\}$ is an absorbing Markov chain with five states. Its matrix $\mathbf{P}$ of one-step transition probabilities is given by

$$\begin{array}{c} \text{from /to} \\ 0 \\ (1, K) \\ (1, D) \\ (2, K) \\ (2, D) \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & (1, K) & (1, D) & (2, K) & (2, D) \end{array} \\ \begin{pmatrix} 0.3 & 0.4 & 0.3 & 0 & 0 \\ 0.25 & 0 & 0.25 & 0.5 & 0 \\ 0.2 & 0.3 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Let the random variable L denote the duration of the match. The random variable L takes on a value larger than r only if the Markov chain does not visit either of the states $(2, K)$ and $(2, D)$ in the first r steps. Hence

$$\begin{aligned} P(L > r) &= P(X_k \neq (2, K), (2, D) \text{ for } k = 1, \ldots, r \mid X_0 = 0) \\ &= P(X_r \neq (2, K), (2, D) \mid X_0 = 0), \end{aligned}$$

where the last equality uses the fact that the states $(2, K)$ and $(2, D)$ are absorbing so that $X_k \neq (2, K), (2, D)$ for any $k < r$ if $X_r \neq (2, K), (2, D)$. Noting that $P(X_r \neq (2, K), (2, D) \mid X_0 = 0) = 1 - P(X_r = (2, K) \mid X_0 = 0) - P(X_r = (2, D) \mid X_0 = 0)$, we obtain

$$P(L > r) = 1 - p^{(r)}_{0,(2,K)} - p^{(r)}_{0,(2,D)}.$$

Hence, the value of $P(L > r)$ can be calculated by multiplying the matrix $\mathbf{P}$ by itself r times. We give the matrix product $\mathbf{P}^r$ for $r = 10$ and 30:

$$\mathbf{P}^{10} = \begin{pmatrix} 0.0118 & 0.0109 & 0.0092 & 0.5332 & 0.4349 \\ 0.0066 & 0.0061 & 0.0051 & 0.7094 & 0.2727 \\ 0.0063 & 0.0059 & 0.0049 & 0.3165 & 0.6663 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$$

$$\mathbf{P}^{30} = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.5506 & 0.4494 \\ 0.0000 & 0.0000 & 0.0000 & 0.7191 & 0.2809 \\ 0.0000 & 0.0000 & 0.0000 & 0.3258 & 0.6742 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix} = \mathbf{P}^{31} = \dots .$$

In particular, $P(L > 10) = 1 - 0.5332 - 0.4349 = 0.0319$. The numerical calculations show that by $r = 30$ all of the entries of the matrix product $\mathbf{P}^r$ have converged up to four decimal places. The probability that the system will ultimately be absorbed in state $(2, K)$ is given by $\lim_{r\to\infty} p^{(r)}_{0,(2,K)}$ (why?). Thus we can read off from the matrix $\mathbf{P}^{30}$ that with probability 0.5506 Karparoff will be the final winner.

Instead of computing the absorption probability by calculating sufficiently high powers of $\mathbf{P}^r$, it can be more efficiently computed by solving a system of linear equations. To write down these equations, we use a parametrization idea. The idea is to define f_s as the probability that Karparoff will be the final winner when the starting point is state s, where s is any of the states $0, (1, K), (2, K), (1, D), (2, D)$. The probability f_0 is of main interest, but we need the other probabilities f_s to write down the linear equations. Obviously, $f_{(2,K)} = 1$ and $f_{(2,D)} = 0$. In general, how do we find f_s? Either the absorbing state $(2, K)$ is reached directly from state s , or it is reached from some other state v. The joint probability of the independent events of passing from state s to state v and then proceeding from state v to the absorbing state $(2, K)$ is $p_{sv} f_v$. Applying next the law of conditional probabilities, the equation $f_s = \sum_v p_{sv} f_v$ is obtained. In this way, we find

$$\begin{aligned} f_0 &= 0.3 f_0 + 0.4 f_{(1,K)} + 0.3 f_{(1,D)} \\ f_{(1,K)} &= 0.25 f_0 + 0.25 f_{(1,D)} + 0.5 f_{(2,K)} \\ f_{(1,D)} &= 0.2 f_0 + 0.3 f_{(1,K)} + 0.5 f_{(2,D)}, \end{aligned}$$

where $f_{(2,K)} = 1$ and $f_{(2,D)} = 0$. The solution of this system of three linear equations in three unknowns is given by $f_0 = 0.5506$, $f_{(1,K)} = 0.7191$ and $f_{(1,D)} = 0.3258$, in agreement with the entries of the matrix $\mathbf{P}^r$ for r large.

In order to find the expected value of the duration of the match, we use again the approach of setting up a system of linear equations through first-step analysis. Define μ_s as the expected value of the remaining duration of the match when the starting point is state s. The goal is to find μ_0. Given that the system begins in state s, the system will be in state v after the first step with probability p_{sv}, and the additional number of steps from state v until the process enters an absorbing state has expected value μ_v. Hence, by the law of conditional expectations, we have the general formula $\mu_s = \sum_v (1 + \mu_v) p_{sv}$. This leads to the linear equations

$$\begin{aligned} \mu_0 &= 1 + 0.3\mu_0 + 0.4\mu_{(1,K)} + 0.3\mu_{(1,D)} \\ \mu_{(1,K)} &= 1 + 0.25\mu_0 + 0.25\mu_{(1,D)} + 0.5\mu_{(2,K)} \\ \mu_{(1,D)} &= 1 + 0.2\mu_0 + 0.3\mu_{(1,K)} + 0.5\mu_{(2,D)}, \end{aligned}$$

where $\mu_{(2,K)} = \mu_{(2,D)} = 0$. The solution of this system of three linear equations in three unknowns is given by $\mu_0 = 4.079$, $\mu_{(1,K)} = 2.674$ and $\mu_{(1,D)} = 2.618$. In this way we find that the expected value of the duration of the match is 4.079 games. Isn't this approach much more elegant and simpler than the approach of calculating μ_0 as $\mu_0 = \sum_{r=1}^{\infty} r P(L = r)$?

An absorbing Markov chain may also be useful to calculate so-called tabu probabilities. A *tabu probability* is the probability of avoiding some given set of states during a certain number of transitions. To illustrate this, we consider Example 15.5 again and ask the following question. What is the probability of no rain in the next five days given that it is sunny today? The trick is to make state 3 (rainy weather) absorbing. The Markov matrix **P** in Example 15.5 is adjusted by replacing the third row corresponding to state 3 by the row vector (0, 0, 1). This gives the Markov matrix

$$\mathbf{Q} = \begin{pmatrix} 0.70 & 0.10 & 0.20 \\ 0.50 & 0.25 & 0.25 \\ 0 & 0 & 1 \end{pmatrix}.$$

Some reflection shows that the probability of no rain in the next five days, given that is rainy today, equals $1 - q_{13}^{(5)}$. The matrix product $\mathbf{Q}^5$ is

$$\mathbf{Q}^5 = \begin{pmatrix} 0.2667 & 0.0492 & 0.6841 \\ 0.2458 & 0.0454 & 0.7087 \\ 0 & 0 & 1 \end{pmatrix}.$$

Hence, $1 - q_{13}^{(5)} = 1 - 0.6841 = 0.3159$. Suppose we had asked the question of what is the probability of no rain during the coming five days given that it is rainy today? The answer to this question requires the matrix product $\mathbf{Q}^4$ rather than $\mathbf{Q}^5$. By conditioning on the state of tomorrow's weather, it is readily seen that the probability called for is given by $p_{31}(1 - q_{13}^{(4)}) + p_{32}(1 - q_{23}^{(4)})$. The value of this probability is 0.2698.

Problem 15.14 A theater buff has attended 150 performances at a theater with 49 seats. At the start of each performance, the theater buff has been randomly directed to one of the 49 seats. Calculate the probability that this person has occupied every seat in the theater at least one time.

Problem 15.15 In each drawing of the Lotto 6/45 six different numbers are drawn from the numbers $1, 2, \ldots, 45$. Calculate for $r = 15, 25, 35$, and 50 the probability that more than r drawings are needed until each of the numbers $1, 2, \ldots, 45$ has been drawn.

Problem 15.16 Calculate the probability of a run of five heads or five tails occurring in 20 tosses of a fair coin. What is the probability of a run of five heads occurring in 20 tosses of a fair coin?

Problem 15.17 The Bubble Company offers a picture of one of 25 popstars in a pack of chewing gum. John and Peter each buy one pack every week. They pool the pictures of the popstars. Assuming equal chances of getting any of the 25 pictures with one purchase, denote by the random variable N the number of weeks until John and Peter have collected two complete sets of 25 pictures. Calculate the expected value of N and calculate the probability $P(N > n)$ for $n = 50, 75, 100, 125$, and 150.

Problem 15.18 A fair coin is tossed until the last three tosses either show the combination TTH or the combination THH. Here H stands for heads and T stands for tails. What is the probability that the combination TTH appears before the combination THH? Can you explain why this probability is larger than 0.5? Also, consider the following game. A fair coin is tossed until heads appears three times in a row. You pay \$1 for each toss of the coin, but you get \$12.50 as soon as heads has appeared three times in a row. Is this a fair game?

Problem 15.19 In European roulette the wheel is divided in 37 sections, numbered as $1, 2, \ldots, 36$ and 0. Of the sections numbered from 1 to 36, 18 are red and 18 are black. The section marked 0 is winning for the house. Use an absorbing Markov chain to calculate the probability that in the next 1,000,000 spins of the wheel the same color will come up 26 or more times in a row. *Remark*: you can reduce the computational effort by using the relation $\mathbf{P}^n = \mathbf{P}^{\frac{1}{2}n} \times \mathbf{P}^{\frac{1}{2}n}$ if $n = 2^r$ for a positive integer r.

Problem 15.20 Joe Dalton desperately wants to raise his current bankroll of \$800 to \$1,000 in order to pay his debts before midnight. He enters a casino and decides to play for high stakes at European roulette. He bets on red each time. The stake is \$200 if his bankroll is \$200 or \$800 and is \$400 if his bankroll is \$400 or \$600. Joe quits as soon as he has either reached his goal or lost everything. For $r = 1, 2, \ldots, 10$, calculate the probability that he will place exactly r bets. What is the probability that he will reach his goal? Also, calculate the expected value and the standard deviation of the total number of bets. *Hint*: define X_i as the number of remaining bets if Joe's current bankroll is \$200i and use the relation $E(X_i^2) = \frac{19}{37}E[(1 + X_j)^2] + \frac{18}{37}E[(1 + X_k)^2]$ for appropriate j and k (e.g., X_3 is distributed as $1 + X_1$ with probability $\frac{19}{37}$ and is distributed as $1 + X_5$ with probability $\frac{18}{37}$).

Problem 15.21 You start out with 25 coins in small change in your pocket. Each time a beggar asks you for money you give him a random number of the coins left in your pocket. Use an absorbing Markov chain to calculate the probability mass function of the number of beggars who are favored by you. Also, calculate the expected value and the standard deviation of the number of favored beggars.

Problem 15.22 Consider Problem 2.40 from Chapter 2 again. Use a Markov chain to find the probability of the first passenger in line changing seats r or more times before getting to his assigned seat for $r = 1, 2, \ldots, 10$. What is the expected number of times the passenger will change seats? Also, solve the Problems 2.33 and 2.42 from Chapter 2 by using an absorbing Markov chain.

Problem 15.23 Consider Problem 15.8 again. A certain car is now at location 4. As soon as this car returns to location 1, it will be overhauled. What is the probability that the car will be rented out more than five times before it returns to location 1? What is this probability if the car is originally at location 1?

Problem 15.24 Consider Problem 15.4 again. Use an absorbing Markov chain to calculate the probability of having no rain on two consecutive days during the next seven days given that it was sunny during the last two days. What is the value of this probability if the last two days were rainy?

15.4 Long-run analysis of Markov chains

In Example 15.5, the long-run behavior of a Markov chain describing the state of the weather was analyzed by taking sufficiently high powers of the matrix of one-step transition probabilities. It was empirically found that the n-step transition probabilities $p_{ij}^{(n)}$ have a limit as n becomes very large. Moreover, it turned out the limit was independent of the starting state i. The limiting probabilities in the weather example also had a natural interpretation in terms of long-run frequencies. In this section these results will be put in a general framework. In particular, it will be seen that the long-run behavior of a Markov chain can be more efficiently analyzed than by taking high powers of the matrix of one-step transition probabilities.

The long-run (or equilibrium) analysis of Markov chains only makes sense for Markov chains without absorbing states. In the sequel we restrict ourselves to Markov chains with no two or more disjoint closed sets of states. A closed set of states is naturally defined as follows:

Definition 15.2 *A nonempty set C of states is said to be a closed set for the Markov chain $\{X_n\}$ if*

$$p_{ij} = 0 \qquad \textit{for } i \in C \textit{ and } j \notin C,$$

that is, the process cannot leave the set C once the process is in the set C.

The assumption of no two disjoint closed sets is necessary in order to produce the situation in which the effect of the starting state fades away after a sufficiently long period of time. To illustrate this, we consider the following example. Take a Markov chain with state space $I = \{1, 2, 3, 4\}$ and one-step transition probabilities p_{ij} with $p_{11} = p_{21} = 0.7, p_{12} = p_{22} = 0.3, p_{33} = p_{43} = 0.2$, $p_{34} = p_{44} = 0.8$, and the other $p_{ij} = 0$. In this example, the Markov chain has the two disjoint closed sets $C_1 = \{1, 2\}$ and $C_2 = \{3, 4\}$, and so, for any state j, $\lim_{n\to\infty} p_{ij}^{(n)}$ depends on the starting state i. In most applications of Markov chains the assumption of no two disjoint closed sets is naturally satisfied. In a Markov chain with multiple disjoint closed sets, each closed set can be separately analyzed as an independent chain.

In the following analysis, the basic assumption that the system has a *finite* state space I is important. The long-run analysis of infinite-state Markov chains involves subtleties which are beyond the scope of this book.

Rule 15.4 *Suppose that the n-step transition probability $p_{ij}^{(n)}$ of the Markov chain $\{X_n\}$ has a limit as $n \to \infty$ for all $i, j \in I$ such that for each $j \in I$ the limit is independent of the starting state i. Denote the limit by $\pi_j = \lim_{n\to\infty} p_{ij}^{(n)}$ for any $j \in I$. Then, the limiting probabilities π_j are the unique solution to the linear equations*

$$\pi_j = \sum_{k\in I} \pi_k p_{kj} \qquad \textit{for } j \in I \quad \textit{and} \quad \sum_{j\in I} \pi_j = 1.$$

The proof of Rule 15.4 is based on the Chapman-Kolmogorov equations in Rule 15.1. Letting n tend to infinity in these equations, we obtain

$$\pi_j = \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_{k\in I} p_{ik}^{(n-1)} p_{kj} = \sum_{k\in I} \lim_{n\to\infty} p_{ik}^{(n-1)} p_{kj} = \sum_{k\in I} \pi_k p_{kj}.$$

The interchange of the order of limit and summation in the third equality is justified by the finiteness of the state space I. Letting $n \to \infty$ in $\sum_{j\in I} p_{ij}^{(n)} = 1$, we obtain $\sum_{j\in I} \pi_j = 1$. It remains to prove that the above system of linear equations has a unique solution. To verify this, let $(x_j,\ j \in I)$ be any solution to the linear equations $x_j = \sum_{k\in I} x_k p_{kj}$. It is helpful to use matrix notation. Define the row vector $\mathbf{x} = (x_j)$ and the matrix $\mathbf{P} = (p_{ij})$. Then $\mathbf{x} = \mathbf{xP}$. Multiplying both sides of this equation by $\mathbf{P}$, we obtain $\mathbf{xP} = \mathbf{xP}^2$. Hence, by $\mathbf{xP} = \mathbf{x}$, we have $\mathbf{x} = \mathbf{xP}^2$. Applying this argument repeatedly, we find $\mathbf{x} = \mathbf{xP}^n$ for all $n = 1, 2, \ldots$. Componentwise, for each $j \in I$

$$x_j = \sum_{k\in I} x_k p_{kj}^{(n)} \qquad \text{for all } n = 1, 2, \ldots.$$

This implies that $x_j = \lim_{n\to\infty} \sum_{k\in I} x_k p_{kj}^{(n)}$. Interchanging the order of limit and summation, we obtain $x_j = \sum_{k\in I} x_k \pi_j = \pi_j(\sum_{k\in I} x_k)$ for all $j \in I$. Hence, $x_j = c\pi_j$ for all $j \in I$ with the constant $c = \sum_{k\in I} x_k$. Since the x_k also satisfy the normalizing equation $\sum_{k\in I} x_k = 1$, we have $c = 1$ and so $x_j = \pi_j$ for all $j \in I$, proving the desired uniqueness result.

The limiting probabilities π_j in Rule 15.4 constitute a probability distribution, that is, $\pi_j \geq 0$ for all j and $\sum_{j\in I} \pi_j = 1$. This is not always true for an infinite-state Markov chain. In the counterexample with $I = \{1, 2, \ldots\}$ and $p_{i,i+1} = 1$ for all i, we have $\lim_{n\to\infty} p_{ij}^{(n)} = 0$ for all i, j.

The result of Rule 15.4 motivates the concept of *equilibrium distribution.*

Definition 15.3 *A probability distribution $\{\eta_j,\ j \in I\}$ is called an equilibrium distribution of the Markov chain $\{X_n\}$ if*

$$\eta_j = \sum_{k\in I} \eta_k p_{kj} \qquad \textit{for all } j \in I.$$

The terms *invariant distribution* and *stationary distribution* are also often used. The name equilibrium distribution can be explained as follows. If $P(X_0 = j) = \eta_j$ for all $j \in I$, then, for any time point $n \geq 1$, $P(X_n = j) = \eta_j$ for all $j \in I$. This result should be understood as follows. Suppose that you are going to inspect the state of the process at any time $t = n$ having *only* the information that the starting state of the process was determined according to the probability distribution $\{\eta_j\}$. Then the probability of finding the process in state s is η_s for any $s \in I$. The proof is simple. Suppose it has been verified for

$t = 0, 1, \ldots, n-1$ that $P(X_t = j) = \eta_j$ for all $j \in I$. Then

$$P(X_n = j) = \sum_{k \in I} P(X_n = j \mid X_{n-1} = k) P(X_{n-1} = k)$$

gives that $P(X_n = j) = \sum_{k \in I} p_{kj} \eta_k = \eta_j$ for all $j \in I$, as was to be verified.

It will be seen below that a Markov chain without two disjoint closed sets has a unique equilibrium distribution. Such a Markov chain is said to have reached *statistical equilibrium* if its state is distributed according to the equilibrium distribution. Under the assumption that $\lim_{n\to\infty} p_{ij}^{(n)}$ exists for all $i, j \in I$ and is independent of the starting state i, Rule 15.4 states that the Markov chain has a unique equilibrium distribution. The limiting probabilities $\pi_j = \lim_{n\to\infty} p_{ij}^{(n)}$ then constitute the equilibrium probabilities. Three obvious questions are: Does $\lim_{n\to\infty} p_{ij}^{(n)}$ always exist? Does any Markov chain have an equilibrium distribution? If an equilibrium distribution exists, is it unique? It will be seen below that the answer to the last two questions is positive if the Markov chain has no two or more disjoint closed sets. The answer, however, to the first question is negative. A counterexample is provided by the two-state Markov chain with state space $I = \{1, 2\}$ and one-step transition probabilities $p_{12} = p_{21} = 1$ and $p_{11} = p_{22} = 0$. In this example the system alternates between the states 1 and 2. This means that, as a function of the time parameter n, the n-step transition probability $p_{ij}^{(n)}$ is alternately 0 and 1 and thus has no limit as n becomes very large. The periodicity of the Markov chain is the reason that $\lim_{n\to\infty} p_{ij}^{(n)}$ does not exist in this example. Periodicity of a Markov chain is defined as follows:

Definition 15.4 *A Markov chain $\{X_n\}$ is said to be periodic if there are multiple disjoint sets $R_1, \ldots, R_d$ with $d \geq 2$ such that a transition from a state in R_k always occurs to a state in R_{k+1}* for $k = 1, \ldots, d$ *with $R_{d+1} = R_1$. Otherwise, the Markov chain is said to be aperiodic.*

In general, the existence of $\lim_{n\to\infty} p_{ij}^{(n)}$ requires an aperiodicity condition. However, it is not necessary to impose an aperiodicity condition on the Markov chain in order to have the existence of an equilibrium distribution. To work this out, we need the concept of Cesàro-limit. A sequence $(a_1, a_2, \ldots)$ of real numbers is said to have a *Cesàro-limit* if $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} a_k$ exists. The Cesàro-limit is more general than the ordinary limit. A basic result from calculus is that $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} a_k$ exists and is equal to $\lim_{n\to\infty} a_n$ if the latter limit exists. A beautiful and useful result from Markov chain theory is that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)}$$

always exists! A heuristic explanation of this result is as follows. Think of a reward structure imposed on the process with reward 1 in one of the states and reward 0 in the other states. Fix state $j = r$ and imagine that a reward 1 is earned each time the process makes a transition to state r and a reward 0 is earned in any other state. Then, by Rule 15.3, $\sum_{k=1}^{n} p_{ir}^{(k)}$ is the total expected reward earned up to time n when the starting state is i. It is plausible that the long-run average expected reward per unit time is well defined. In other words, $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ir}^{(k)}$ exists. This limit gives also the long-run frequency at which the process visits state r.

We now come to the main result of this section. This result will be stated without proof.

Rule 15.5 *Let $\{X_n\}$ be a finite-state Markov chain with no two or more disjoint closed sets. The Markov chain then has a unique equilibrium distribution $\{\pi_j\}$:*

(a) *The equilibrium probabilities π_j are given by*

$$\pi_j = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} \qquad \textit{for all } j \in I,$$

with the averaging limit being independent of the starting state i.

(b) *The π_j are the unique solution to the linear equations*

$$\pi_j = \sum_{k\in I} \pi_k p_{kj} \qquad \textit{for } j \in I \quad \textit{and} \quad \sum_{j\in I} \pi_j = 1.$$

(c) *If the Markov chain is aperiodic, then* $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ *for all $i, j \in I$.*

The equations $\pi_j = \sum_{k\in I} \pi_k p_{kj}$ for $j \in I$ are called the *equilibrium equations* and the equation $\sum_{j\in I} \pi_j = 1$ is called the *normalizing equation*. In a similar way as in the proof of Rule 15.4, it can be shown that any solution (x_j) to the equilibrium equations alone is uniquely determined up to a multiplicative constant; that is, for some constant c, $x_j = c\pi_j$ for all $j \in I$. The size of the system of linear equations in part (b) of Rule 15.5 is one more than the number of unknowns. However, it is not difficult to see that one of the equilibrium equations is redundant (summing both sides of the equilibrium equations over j gives "$1 = 1$" after an interchange of the order of summation). Thus, by deleting one of the equilibrium equations, one obtains a square system of linear equations which uniquely determine the unknowns π_j.

An easy way to memorize the equilibrium equations is to note that the equilibrium equations are obtained by multiplying the row vector $\vec{\pi}$ of the equilibrium probabilities with the *column* vectors of the matrix $\mathbf{P}$ of one-step transition probabilities ($\vec{\pi} = \vec{\pi}\mathbf{P}$).

Example 15.5 (continued) The Markov chain describing the state of the weather has no two disjoint closed sets. Thus, the unique equilibrium probabilities of the Markov chain are found from the equilibrium equations

$$\begin{aligned}\pi_1 &= 0.70\pi_1 + 0.50\pi_2 + 0.40\pi_3\\ \pi_2 &= 0.10\pi_1 + 0.25\pi_2 + 0.30\pi_3\\ \pi_3 &= 0.20\pi_1 + 0.25\pi_2 + 0.30\pi_3\end{aligned}$$

together with $\pi_1 + \pi_2 + \pi_3 = 1$. One of the equilibrium equations (say, the first one) can be omitted to obtain a square system of three linear equations in three unknowns. Solving these equations gives

$$\pi_1 = 0.5960,\ \pi_2 = 0.1722,\ \pi_3 = 0.2318.$$

Noting that the Markov chain in this example is aperiodic, this result agrees with the earlier calculated matrix product $\mathbf{P}^n$ for n sufficiently large. The equilibrium probability π_j can be given two interpretations in this example. First, it can be stated that the weather after many days will be sunny, cloudy and rainy with probabilities 0.5960, 0.1722 and 0.2318, respectively. Secondly, these probabilities also give the long-run proportions of time during which the weather will be sunny, cloudy and rainy.

The next example deals with a Markov chain in which the equilibrium distribution cannot be seen as the state distribution at a time point in the far distant future.

Example 15.2 (continued) The equilibrium equations for the Ehrenfest model are given by

$$\pi_j = \frac{r-j+1}{r}\pi_{j-1} + \frac{j+1}{r}\pi_{j+1} \qquad \text{for } j = 1, \ldots, r-1$$

with $\pi_0 = \frac{1}{r}\pi_1$ and $\pi_r = \frac{1}{r}\pi_{r-1}$. Intuitively, any marked particle is to be found equally likely in either of the two compartments after many transitions. This suggests the binomial distribution for the equilibrium probabilities. Indeed, by substitution into the equilibrium equations, it is readily verified that $\pi_j = \binom{r}{j}(\frac{1}{2})^r$ for $j = 0, 1, \ldots, r$. The equilibrium distribution is unique, since the Markov chain has no two disjoint closed sets. However, the Markov chain is periodic: a transition from any state in the subset of even-numbered states leads to a state in the subset of odd-numbered states, and vice versa. Thus, $\lim_{n\to\infty} p_{ij}^{(n)}$ does not exist and the proper interpretation of π_j is the interpretation as the long-run proportion of time during which compartment A contains j particles.

In each of the above two examples the equilibrium probabilities could be interpreted as long-run frequencies. This interpretation is generally valid.

Rule 15.6 *Let $\{\pi_j\}$ be the unique equilibrium distribution of a finite-state Markov chain $\{X_n\}$ that has no two or more disjoint closed sets. Then, for any state $j \in I$*

the long-run proportion of time the process will be in state $j = \pi_j$

with probability one, independently of the starting state $X_0 = i$.

The term "with probability one" is subtle and should be interpreted as follows: for any fixed state j, $P(\{\omega : \lim_{n\to\infty}(1/n)\sum_{k=1}^{n} I_k(\omega) = \pi_j\}) = 1$ when the random variable I_k equals 1 if $X_k = j$ and 0 otherwise, and ω represents a possible outcome of the infinite sequence $X_0, X_1, \ldots$. In other words, the set of outcomes ω for which the values of $(1/n)\sum_{k=1}^{n} I_k(\omega)$ do not converge to π_j has probability zero. A mathematical proof of this strong law of large numbers for Markov chains is beyond the scope of this book.

In the case that the Markov chain is aperiodic, π_j can also be interpreted as the probability of finding the system in state j at a point of time in the far distant future. One should understand this interpretation as follows: if you inspect the process after it has been running for a very long time and you have no information about recently visited states, then you will find the process in state j with probability π_j. In the case that you have information, probabilities change. The interpretation of π_j as a long-run frequency is much more concrete and is often more useful from a practical point of view.

Also, a physical interpretation can be given to the equilibrium equations. In physical terms, $\pi_k p_{kj}$ is the long-run average rate at which the process goes from state k to state j. Thus, the equation $\pi_j = \sum_{k\in I} \pi_k p_{kj}$ expresses in mathematical terms the physical principle:

> **the average rate at which the process makes a transition *from* state j is equal to the average rate at which the process makes a transition *to* state j.**

Remark 15.1 For a finite-state Markov chain having no two disjoint closed sets, it can be shown that the equilibrium probability $\pi_j = 0$ if state j is transient and $\pi_j > 0$ if state j is recurrent. A state j is said to be *transient* if $\sum_{n=1}^{\infty} p_{jj}^{(n)} < \infty$ and is said to be *recurrent* if $\sum_{n=1}^{\infty} p_{jj}^{(n)} = \infty$. The rationale for this definition is the fact that $\sum_{n=1}^{\infty} p_{jj}^{(n)}$ represents the expected value of the number of returns of the process to state j over the time points $n = 1, 2, \ldots$ given that the process starts in state j (see Rule 15.3). Loosely speaking, a recurrent state is one to which the process keeps coming back and a transient state is one which the process eventually leaves forever. Also, for a recurrent state j, it can be proved from the strong law of large numbers that $\pi_j = 1/\mu_{jj}$, where the mean recurrence time μ_{jj} is defined as the expected value of the number of transitions needed to return from state j to itself.

The above definition of transient state and recurrent state applies to any Markov chain. A finite-state Markov chain can be shown to have the following properties: (a) the set of recurrent states is not empty, and (b) the mean recurrence time μ_{jj} is finite for any recurrent state j. These properties do not necessarily hold for an infinite-state Markov chain. We give two counterexamples:

(a) The Markov chain has state space $I = \{0, 1, 2, \ldots\}$ and $p_{i,i+1} = 1$ for all $i \in I$. Then, all states are transient.

(b) The Markov chain has state space $I = \{0, \pm 1, \pm 2, \ldots\}$ and $p_{i,i-1} = p_{i,i+1} = 0.5$ for all $i \in I$. Then, it can be shown that any state j is recurrent with mean recurrence time $\mu_{jj} = \infty$.

The phenomena in (a) and (b) cannot occur in infinite-state Markov chains satisfying the regularity condition that some state r exists such that state r will ultimately be reached with probability one from any starting state i and the mean recurrence time μ_{rr} is finite. Under this regularity condition it can be shown that the equilibrium results of Section 15.4 also hold for infinite-state Markov chains.

In many applications a cost structure is imposed on a Markov chain. We conclude this chapter with a useful ergodic theorem for such Markov chains.

Rule 15.7 *Let $\{\pi_j\}$ be the unique equilibrium distribution of a finite-state Markov chain $\{X_n\}$ that has no two or more disjoint closed sets. Assume that a cost $c(j)$ is incurred at each visit of the Markov chain to state j for any $j \in I$. Then, with probability one*

$$\textit{the long-run average cost per unit time } = \sum_{j \in I} c(j)\pi_j$$

independently of the starting state $X_0 = i$.

This result is obvious from the interpretation of the π_j in Rule 15.6.

Problem 15.25 The much feared professor Frank N. Stone gives varying versions of an oral examination in assembly line fashion, with students taking the exam one after the other. Each version of the exam may be categorized as difficult, normal or easy. After a difficult exam, the next exam will be difficult with probability 0.2, will be normal with probability 0.5, and will be easy with probability 0.3. After normal and easy exams, these probabilities are 0.5, 0.25 and 0.25. Let's say you take the exam without any knowledge of the difficulty factor of the preceding exams. What is the probability that you will get a difficult exam? What is this probability if you know that your friend had an easy exam, five exams previously?

Problem 15.26 Consider Problem 15.4 again. Calculate the equilibrium probabilities of the Markov chain describing the weather. What is the long-run proportion of days it will be sunny? What is the probability that it will be rainy on a given Sunday many days from now?

Problem 15.27 Consider Example 5.3 again. What is the long-run proportion of time the professor has his license with him? Also, answer this question for Problem 15.2.

Problem 15.28 Consider Example 15.7 again. It is now assumed that Boris Karparoff and Deep Blue play infinitely often against each other. What is the long-run proportion of games won by Boris? How often will Boris win a game after having won the previous game?

Problem 15.29 Let $\{X_n\}$ be a Markov chain with no two disjoint closed sets and state space $I = \{1, 2, \ldots, N\}$. Suppose that the Markov chain is *doubly stochastic*; that is, for each of the columns of the matrix of one-step transition probabilities the column elements sum to one. Verify that the Markov chain has the unique equilibrium distribution $\pi_j = \frac{1}{N}$ for all j.

Problem 15.30 Consider Problem 2.44 from Chapter 2 with Parrondo's paradox again. For each of the two strategies described in this problem, use a Markov chain to calculate the long-run win probability. *Hint*: use a Markov chain with three states and a Markov chain with 12 states.

Problem 15.31 Consider Example 15.4 again. What is the long-run average stock on hand at the end of the week? What is the long-run average ordering frequency and what is the long-run amount of demand lost per week?

Problem 15.32 Consider Problem 15.4 again. The local entrepeneur Jerry Woodside has a restaurant on the island. On every sunny day, his turnover (in dollars) has an $N(\mu_1, \sigma_1^2)$ distribution with $\mu_1 = 1{,}000$ and $\sigma_1 = 200$, while on rainy days his turnover is $N(\mu_2, \sigma_2^2)$ distributed with $\mu_2 = 500$ and $\sigma_2 = 75$. What is the long-run average sales per day?

Problem 15.33 Consider Problem 15.6 again. Suppose that a cost of \$750 is incurred each time the device fails and that each circuit board replaced costs

\$100. What is the long-run proportion of weeks the device operates properly? What is the long-run average weekly cost?

Problem 15.34 A transport firm has effected an insurance contract for a fleet of vehicles. The premium payment is due at the beginning of each year. There are four possible premium classes with a premium payment of P_i in class i, where $P_{i+1} < P_i$ for $i = 1, 2, 3$. If no damage is claimed in the year just ended and the last premium charged is P_i, the next premium payment is P_{i+1} (with $P_5 = P_4$); otherwise, the highest premium P_1 is due. The transport firm has obtained the option to decide only at the end of the year whether the accumulated damage during that year should be claimed or not. In the case that a claim is made, the insurance company compensates the accumulated damage minus an own risk which amounts to r_i for premium class i. The sizes of the damages in successive years are independent random variables that are exponentially distributed with mean $1/\eta$. The claim strategy of the firm is characterized by four given numbers $\alpha_1, \ldots, \alpha_4$ with $\alpha_i > r_i$ for all i. If the current premium class is i, then the firm claims at the end of the year only damages larger than α_i; otherwise, nothing is claimed. How do you calculate the long-run fraction of time the firm is in premium class i? Also, give an expression for the long-run average yearly cost.

Problem 15.35 Consider Problem 15.1 again. Let the Markov chain $\{X_n\}$ describe the number of type-1 particles in compartment A. Prove that the equilibrium probabilities satisfy the recurrence relation

$$p_{k,k-1}\pi_k = p_{k-1,k}\pi_{k-1} \qquad \text{for } k = 1, 2, \ldots, r.$$

Use this result to verify that $\pi_j = \binom{r}{j}\binom{r}{r-j}/\binom{2r}{r}$ for $j = 0, 1, \ldots, r$. *Remark*: the recurrence relation for the π_k expresses that the system has the following property when it has reached statistical equilibrium. Conditionally upon being in state k, the probability of coming from state $k-1$ is the same as the probability of going to state $k-1$ ($\frac{p_{k-1,k}\pi_{k-1}}{\pi_k} = p_{k,k-1}$), and the probability of coming from state $k+1$ is the same as the probability of going to state $k+1$ ($\frac{p_{k+1,k}\pi_{k+1}}{\pi_k} = p_{k,k+1}$). In other words, two outside observers using clocks in opposite directions will see probabilistically identical evolutions of the system when the system is in statistical equilibrium. A Markov chain having this property is said to be a *time-reversible* Markov chain.

Problem 15.36 Let $\{r_j, j \in I\}$ be a given probability distribution, where I is a finite set with N elements. Define the Markov chain $\{X_n\}$ on the state space I by the one-step transition probabilities $p_{ij} = \frac{1}{N-1}\min(\frac{r_j}{r_i}, 1)$ for $j \neq i$ and

$p_{ii} = 1 - \sum_{j:j \neq i} p_{ij}$.[†] Prove that $\{r_j\}$ is the equilibrium distribution of the process $\{X_n\}$. *Hint*: verify the reversibility property $r_j p_{jk} = r_k p_{kj}$ for all j, k, which implies that $r_j = \sum_{k \in I} r_k p_{kj}$ for all j.

Problem 15.37 Let $c(i)$ be a given function on a finite set I. For any $i \in I$, a local neighborhood $N(i)$ of other points from I is given such that $k \in N(j)$ if $j \in N(k)$. For ease, it is assumed that each set $N(i)$ contains the same number of points. The following Markov chain is defined on I. If the current state is i, a candidate state j is chosen at random from $N(i)$. The next state of the process is always j if $c(j) < c(i)$; otherwise, the process moves to j with probability $e^{-c(j)/T}/e^{-c(i)/T}$ and stays in i with probability $1 - e^{-c(j)/T}/e^{-c(i)/T}$. Here $T > 0$ is a control parameter. It is assumed that the sets $N(i)$ are such that the Markov chain is irreducible (that is, any state is accessible from any other state). Then, the unique equilibrium probabilities of the Markov chain are given by $\pi_i = e^{-c(i)/T} / \sum_{k \in I} e^{-c(k)/T}$ for $i \in I$. Prove this result by verifying the reversibility condition $e^{-c(j)/T} p_{jk} = e^{-c(k)/T} p_{kj}$ for all j, k, where the p_{jk} are the one-step transition probabilities of the Markov chain. *Remark*: if the function $c(i)$ assumes its absolute minimum in a unique point m, then $\pi_m \to 1$ as $T \to 0$ (verify). This fact is exploited in the simulated annealing algorithm that is often used to find the minimum of a function on a finite but very large set.

[†] In physical terms, this Markov chain operates as follows. If the current state is i, you choose a candidate state j at random from the $N - 1$ other states. The candidate state j is always the next state of the process if it is more likely than state i (i.e., $r_j > r_i$); otherwise, the process moves to j with probability r_j/r_i and stays in i with probability $1 - r_j/r_i$. The idea of this construction underlies the famous Metropolis-Hastings algorithm from statistics and physics. This algorithm generalizes the acceptance-rejection method discussed in Problem 13.7.

Appendix
Counting methods and e^x

This appendix first gives some background material on counting methods. Many probability problems require counting techniques. In particular, these techniques are extremely useful for computing probabilities in a chance experiment in which all possible outcomes are equally likely. In such experiments, one needs effective methods to count the number of outcomes in any specific event. In counting problems, it is important to know whether the order in which the elements are counted is relevant or not. After the discussion on counting methods, the Appendix summarizes a number of properties of the famous number e and the exponential function e^x both playing an important role in probability.

Permutations

How many different ways can you arrange a number of different objects such as letters or numbers? For example, what is the number of different ways that the three letters A, B, and C can be arranged? By writing out all the possibilities ABC, ACB, BAC, BCA, CAB, and CBA, you can see that the total number is six. This brute-force method of writing down all the possibilities and counting them is naturally not practical when the number of possibilities gets large, for example the number of different ways to arrange the 26 letters of the alphabet. You can also determine that the three letters A, B, and C can be written down in six different ways by reasoning as follows. For the first position, there are three available letters to choose from, for the second position there are two letters over to choose from, and only one letter for the third position. Therefore, the total number of possibilities is $3 \times 2 \times 1 = 6$. The general rule should now be evident. Suppose that you have n distinguishable objects. How many ordered arrangements of these objects are possible? Any ordered sequence of the objects is called a *permutation*. Reasoning similar to that described shows that there are n ways for choosing the first object, leaving $n - 1$ choices for the second object, etc. Therefore the total number of ways to order n distinguishable objects is $n \times (n - 1) \times \cdots \times 2 \times 1$. A convenient shorthand for this product is $n!$ (pronounce: n factorial). Thus, for any positive integer n,

$$n! = 1 \times 2 \times \cdots \times (n - 1) \times n.$$

A convenient convention is $0! = 1$. Summarizing

the total number of ordered sequences (permutations) of n distinguishable objects is $n!$.

Example A.1 A scene from the movie "The Quick and the Dirty" depicts a Russian roulette type of duel. Six identical shot glasses of whiskey are set on the bar, one of which is laced with deadly strychnine. The bad guy and the good guy must drink in turns. The bad guy offers $1,000 to the good guy, if the latter will go first. Is this an offer that should not be refused?

Solution. A handy way to think of the problem is as follows. Number the six glasses from 1 to 6 and assume that the glasses are arranged in a random order after strychnine has been put in one of the glasses. There are 6! possible arrangements of the six glasses. If the glass containing strychnine is in the first position, there remain 5! possible arrangements for the other five glasses. Thus, the probability that the glass in the first position contains strychnine is equal to $5!/6! = 1/6$. By the same reasoning, the glass in each of the other five positions contains strychnine with a probability of $1/6$, before any glass is drunk. It is a fair game. Each of the two "duelists" will drink the deadly glass with a probability of $(3 \times 5!)/6! = 1/2$. The good guy will do well to accept the offer of the bad guy. If the good guy survives the first glass after having drunk it, the probability that the bad guy will get the glass with strychnine becomes $(3 \times 4!)/5! = 3/5$.

Example A.2 Eight important heads of state, including the U.S. President and the British Premier, are present at a summit conference. For the perfunctory group photo, the eight dignitaries are lined up randomly next to one other. What is the probability that the U.S. President and the British Premier will stand next to each other?

Solution. Number the eight heads of state as $1, \ldots, 8$, where the number 1 is assigned to the U.S. President and number 2 to the British Premier. The eight statesmen are put in a random order in a row. There are 8! possible arrangements. If the positions of the U.S. President and the British Premier are fixed, there remain 6! possible arrangements for the other six statesmen. The U.S. President and the British Premier stand next to each other if they take up the positions i and $i + 1$ for some i with $1 \leq i \leq 7$. In the case that these two statesmen take up the positions i and $i + 1$, there are 2! possibilities for the order among them. Thus, there are $6! \times 7 \times 2!$ arrangements in which the U.S. President and the British Premier stand next to each other, and so the sought probability equals $(6! \times 7 \times 2!)/8! = 1/4$.

Combinations

How many different juries of three persons can be formed from five persons A, B, C, D, and E? By direct enumeration you see that the answer is ten: $\{A, B, C\}$, $\{A, B, D\}$, $\{A, B, E\}$, $\{A, C, D\}$, $\{A, C, E\}$, $\{A, D, E\}$, $\{B, C, D\}$, $\{B, C, E\}$, $\{B, D, E\}$, $\{C, D, E\}$. In this problem, the order in which the jury members are chosen is not relevant. The answer ten juries could also have been obtained by a basic principle of counting. First, count how many juries of three persons are possible when attention is paid to the order. Then determine how often each group of three persons has been counted. Thus, the reasoning is as follows. There are five ways to select the first jury member, four ways to then select the next member, and three ways to select the final member. This would give $5 \times 4 \times 3$ ways of forming the jury when the order in which the members are chosen would be relevant. However, this order makes no difference. For example, for the jury consisting of the persons A, B, and C, it is not relevant which of the $3!$ ordered sequences ABC, ACB, BAC, BCA, CAB, CBA has led to the jury. Hence the total number of ways a jury of three persons can be formed from a group of five persons is equal to $\frac{5\times4\times3}{3!}$. This expression can be rewritten as

$$\frac{5\times4\times3\times2\times1}{3!\times2!} = \frac{5!}{3!\times2!}.$$

In general, you can calculate that the total number of possible ways to choose a jury of k persons out of a group of n persons is equal to

$$\begin{aligned}&\frac{n\times(n-1)\times\cdots\times(n-k+1)}{k!}\\ &\quad=\frac{n\times(n-1)\times\cdots\times(n-k+1)\times(n-k)\times\cdots\times1}{k!\times(n-k)!}\\ &\quad=\frac{n!}{k!\times(n-k)!}.\end{aligned}$$

For nonnegative integers n and k with $k \leq n$, we define

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

The quantity $\binom{n}{k}$ (pronounce: n over k) has the interpretation:

> $\binom{n}{k}$ **is the total number of ways to choose** k **different objects out of** n **distinguishable objects, paying no attention to their order.**

The numbers $\binom{n}{k}$ are referred to as the *binomial coefficients*. The binomial coefficients arise in numerous counting problems.

Example A.3 Is the probability of winning the jackpot with a single ticket in Lotto 6/45 larger than the probability of getting 22 heads in a row when tossing a fair coin 22 times?

Solution. In Lotto 6/45, six different numbers are drawn out of the numbers $1, \ldots, 45$. The total number of ways the winning six numbers can be drawn is equal to $\binom{45}{6}$. Hence, the probability of hitting the jackpot with a single ticket is

$$\frac{1}{\binom{45}{6}} = 1.23\times10^{-7}.$$

This probability is smaller than the probability $\left(\frac{1}{2}\right)^{22} = 2.38\times10^{-7}$ of getting 22 heads in a row.

Example A.4 In the Powerball lottery, five distinct white balls are drawn out of a drum with 53 white balls, and one red ball is drawn from a drum with 42 red balls. The white balls are numbered $1, \ldots, 53$ and the red balls are numbered $1, \ldots, 42$. You have filled in a single ticket with five different numbers for the white balls and one number for the red ball (the Powerball number). What is the probability that you match only the Powerball number?

Solution. There are $42 \times \binom{53}{5}$ ways to choose your six numbers. Your five white numbers must come from the 48 white numbers not drawn by the lottery. This can happen in $\binom{48}{5}$ ways. There is only one way to match the Powerball number. Hence the probability that you match the red Powerball alone is

$$\frac{1 \times \binom{48}{5}}{42 \times \binom{53}{5}} = 0.0142.$$

Example A.5 What is the probability that a bridge player's hand of 13 cards contains exactly k aces for $k = 0, 1, 2, 3, 4$?

Solution. There are $\binom{4}{k}$ ways to choose k aces from the four aces and $\binom{48}{13-k}$ ways to choose the other $13 - k$ cards from the remaining 48 cards. Hence, the desired probability is

$$\frac{\binom{4}{k}\binom{48}{13-k}}{\binom{52}{13}}.$$

This probability has the values 0.3038, 0.4388, 0.2135, 0.0412, and 0.0026 for $k = 0, 1, 2, 3$, and 4.

Example A.6 The following question is posed in the sock problem from Chapter 1. What are the probabilities of seven and four matching pairs of socks remaining when six socks are lost during the washing of ten different pairs of socks?

Solution. There are $\binom{20}{6}$ possible ways to choose six socks out of ten pairs of socks. You are left with seven complete pairs of socks only if both socks of three pairs are missing. This can happen in $\binom{10}{3}$ ways. Hence, the probability that you are left with seven complete pairs of socks is equal to

$$\frac{\binom{10}{3}}{\binom{20}{6}} = 0.0031.$$

You are left with four matching pairs of socks only if exactly one sock of each of six pairs is missing. These six pairs can be chosen in $\binom{10}{6}$ ways. There are two possibilities for how to choose one sock from a given pair. This means that there are $\binom{10}{6}2^6$ ways to choose six socks so that four matching pairs of socks are left. Hence, the probability of four matching pairs of socks remaining is equal to

$$\frac{\binom{10}{6}2^6}{\binom{20}{6}} = 0.3467.$$

It is remarkable that the probability of the worst case of four matching pairs of socks remaining is more than hundred times as large as the probability of the best case of seven matching pairs of socks remaining. When things go wrong, they really go wrong.

Exponential function

The history of the number e begins with the discovery of logarithms by John Napier in 1614. At this time in history, international trade was experiencing a period of strong growth, and, as a result, there was much attention given to the concept of compound interest. At that time, it was already noticed that $(1 + \frac{1}{n})^n$ tends to a certain limit if n is allowed to increase without bound

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e,$$

where e is the famous number $e = 2.71828\ldots$.[†] The *exponential function* is defined by e^x, where the variable x runs through the real numbers. A fundamental property of e^x is that this function has itself as derivative. That is

$$\frac{de^x}{dx} = e^x.$$

This property is easy to explain. Consider the function $f(x) = a^x$ for some constant $a > 0$. It then follows from $f(x+h) - f(x) = a^{x+h} - a^x = a^x(a^h - 1)$ that

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = cf(x)$$

for the constant $c = \lim_{h\to 0}(a^h - 1)/h$. The proof is omitted that this limit always exists. Next, one might wonder for what value of a the constant $c = 1$ so that $f'(x) = f(x)$. Noting that the condition $(a^h - 1)/h = 1$ can be written as $a = (1+h)^{1/h}$, it can easily be shown that $\lim_{h\to 0}(a^h - 1)/h = 1$ boils down to $a = \lim_{h\to 0}(1+h)^{1/h}$, yielding $a = e$.

How do we calculate the function e^x? The generally valid relation

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x \qquad \text{for each real number } x$$

is not useful for this purpose. The calculation of e^x is based on the power series expansion

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots.$$

In a compact notation

$$e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!} \qquad \text{for each real number } x.$$

The proof of this power series expansion requires Taylor's theorem from calculus. The fact that e^x has itself as derivative is crucial in the proof. Note that term-by-term differentiation of the series $1 + x + \frac{x^2}{2!} + \cdots$ leads to the same series, in agreement with the

[†] A wonderful account of the number e and its history can be found in E. Maor, *e: The Story of a Number*, Princeton University Press, 1994.

fact that e^x has itself as derivative. The series expansion of e^x shows that $e^x \approx 1 + x$ for x close to 0. In other words

$$1 - e^{-\lambda} \approx \lambda \qquad \text{for } \lambda \text{ close to } 0.$$

This approximation formula is very useful in probability theory.

Geometric series

For any nonnegative integer n

$$\sum_{k=0}^{n} x^k = \frac{1 - x^{n+1}}{1 - x} \qquad \text{for each real number } x \neq 1.$$

This useful result is a direct consequence of

$$\begin{aligned}(1 - x)\sum_{k=0}^{n} x^k &= \sum_{k=0}^{n} x^k - \sum_{k=0}^{n} x^{k+1}\\ &= (1 + x + \cdots + x^n) - (x + x^2 + \cdots + x^n + x^{n+1})\\ &= 1 - x^{n+1}.\end{aligned}$$

The term x^{n+1} converges to 0 for $n \to \infty$ if $|x| < 1$. This leads to the important result

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1 - x} \qquad \text{for each real number } x \text{ with } |x| < 1.$$

This series is called the *geometric series* and is frequently encountered in probability problems. The series $\sum_{k=1}^{\infty} kx^{k-1}$ may be obtained by differentiating the geometric series $\sum_{k=0}^{\infty} x^k$ term by term and using the fact that the derivative of $1/(1 - x)$ is given by $1/(1 - x)^2$. The operation of term-by-term differentiation is justified by a general theorem for the differentiation of power series and leads to the result

$$\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1 - x)^2} \qquad \text{for each real number } x \text{ with } |x| < 1.$$

Recommended reading

There are many fine books on probability theory available. The following more applied books are recommended for further reading.

1. W. Feller, *Introduction to Probability Theory and its Applications*, Vol I, third edition, Wiley, New York, 1968.
   This classic in the field of probability theory is still up-to-date and offers a rich assortment of material. Intended for the somewhat advanced reader.
2. S.M. Ross, *Introduction to Probability Models*, eighth edition, Academic Press, New York, 2002.
   A delightfully readable book that makes a good companion to Feller, noted above. Provides a clear introduction to many advanced topics in applied probability.
3. H.C. Tijms, *A First Course in Stochastic Models*, Wiley, Chichester, 2003.
   This is an advanced textbook on stochastic processes and gives particular attention to applications and solution tools in computational probability.

Answers to odd-numbered problems

Chapter 2

2.1 The answer is yes. Use the sample space for this conclusion.

2.3 Take $\Omega = \{(i_1, i_2, i_3, i_4) | i_k = 0, 1 \text{ for } k = 1, \ldots, 4\}$ as sample space and assign a probability of $\frac{1}{16}$ to each element of Ω. The probability of three puppies of one gender and one of the other is $\frac{8}{16}$. The probability of two puppies of each gender is $\frac{6}{16}$.

2.5 Take the set of all 10! permutations of the integers $1, \ldots, 10$ as sample space. The number of permutations having the winning number in any given position i is 9! for each $i = 1, \ldots, 10$. In both cases your probability of winning is $9!/10! = 1/10$.

2.7 Take $\Omega = \{(i, j) | i, j = 1, \ldots, 6\}$ as sample space and assign a probability of $\frac{1}{36}$ to each element of Ω. The expected payoff is $\$2 \times \frac{15}{36} + \$0 \times \frac{21}{36} = \$\frac{30}{36}$ for both bets.

2.9 Invest 35.5% of your bankroll in the risky project each time. The effective rate of return is 6.6%.

2.11 The probabilities are 0.1646 and 0.6703. Simulation leads to the estimate 9.58 for the expected value.

2.13 This problem is a variant on the daughter-son problem from Chapter 1. Your probability of winning is $\frac{1}{3}$. The bet is not fair.

2.15 The probability is 0.875.

2.17 For the case of random numbers from the interval $(-q, q)$, the probability has the value 0.627, independently of q (dividing A, B and C by q gives random numbers between -1 and 1). For the case of nonzero random integers between $-q$ and q, the probability has the values 0.500, 0.603, 0.624, 0.627, and 0.627 for $q = 1$, 10, 100, 1,000, and 10,0000.

2.19 For the triangle OAB, the probability has the value 0.750 for the circle and the value 0.625 for the sphere. The simulated values of the other two probabilities are 0.720 and 0.529.

2.21 The expected values in parts (a), (b), (c), and (d) are 0.333, 0.521, 0.905, and 0.365, respectively.

2.23 The expected values of your loss and the total amount you bet are \$0.942 and \$34.86.

2.25 The probabilities are 0.058 and 0.045.

2.27 The probability of the bank winning is 0.81, and the average number of points collected by the bank is 9.4.

2.29 The optimal value of L is 50 and the maximal probability of candidate A winning is 0.482.

2.31 The expected payoff is \$0.60 and the probability of getting 25 or more points is 0.693.

2.33 The probability is 0.257.

2.35 The game is not fair. The expected number of tosses required is 14.

2.37 For $n = 25$ and 100, simulation leads to the values 6.23 and 12.52 for the expected value of the distance between the starting and ending points. The simulated values of the desired probability are 0.396 and 0.335 for $n = 25$ and 100.

2.39 The probabilities are 0.587, 0.312, 0.083, and 0.015, respectively.

2.41 The probabilities are 0.143, 0.858, 0.833, 0.800, 0.750, 0.667, and 0.500.

2.43 The probability is 0.60.

2.45 The probabilities are 0.329 and 0.536.

Chapter 3

3.1 Yes.

3.3 No, the probability is $\frac{1}{10{,}000}$.

3.5 The probability is 0.01.

3.7 The proposition is unfavorable for the friends who stay behind. Their leaving friend wins on average $1 \times \frac{65}{81} - 4 \times \frac{16}{81} = \frac{1}{81}$ drink per round.

3.9 Your probability of winning is $1 - (100 \times 99 \times \cdots \times 86)/100^{15} = 0.6687$.

3.11 The probabilities in parts (a) and (b) are $1 - (25 \times 24 \times \cdots \times 19)/25^7 = 0.6031$ and $1 - \frac{24^6}{25^6} = 0.2172$.

3.13 Substituting $n = 78{,}000$ and $c = \binom{52}{13}$ into the formula in part (b) of Problem 3.12 gives the probability 0.0048.

3.15 Substituting $n = 500$ and $c = 2{,}400{,}000$ into the formula in part (b) of Problem 3.12 gives the probability 0.051.

3.17 The probabilities in parts (a) and (b) are 0.764 and 0.006.

3.19 Let A be the event that the sports car is won and let B_i be the event that the contestant selects i keys. Then, $P(A) = \sum_{i=0}^{2} P(A \mid B_i)P(B_i)$ with $P(A \mid B_0) = 0$, $P(A \mid B_1) = \frac{2}{5}$, and $P(A \mid B_2) = 1 - \frac{3}{10} = \frac{7}{10}$. This leads to $P(A) = 0 \times \frac{1}{4} + \frac{2}{5} \times \frac{1}{2} + \frac{7}{10} \times \frac{1}{4} = \frac{3}{8}$.

3.21 The main point will be equal to m with probability $r_m = p_m / \sum_{k=5}^{9} p_k$ for $5 \leq m \leq 9$, where the p_i are the same as in the craps example. Using the law of conditional probabilities, we obtain that the probability of the player winning on the main point is $p_5r_5 + (p_6 + p_7)r_6 + (p_7 + p_{11})r_7 + (p_8 + p_{12})r_8 + p_9r_9 = 0.1910$. Similarly, the conditional probability of the player winning on a chance point given that the main point is m equals $\sum_{i \notin A_m} p_i \frac{p_i}{p_i + p_m}$, where $A_5 = \{5, 11, 12\}$, $A_6 = \{6, 7, 11\}$, $A_7 = \{7, 11, 12\}$, $A_8 = \{8, 11, 12\}$, and $A_9 = \{9, 11, 12\}$. Finally, it follows that the probability of the player winning on a chance point is 0.3318. The house percentage is 5%.

3.23 The house percentage is 2.88%.

3.25 The house percentage is 7.45%.

3.27 The probabilities are 0.0515, 0.1084, 0.1592, and 0.1790.

3.29 The probability is 0.7853. The expected values are \$57.64 and \$2,133.

3.31 Let the random variable N_i denote the number of times that number i will be drawn in the next 250 draws of Lotto 6/45. Using computer simulation, we find that $P(\sum_{i=1}^{45} |N_i - 33.333| > 202) = 0.333$.

Chapter 4

4.1 Poisson distribution.

4.3 Apply the binomial distribution with $n = 1{,}500$ and $p = \frac{1}{1000}$. The desired probability is 0.7770.

4.5 Apply the binomial distribution with $n = 125$ and $p = \left(\frac{1}{2}\right)^7$. The desired probabilities are 0.625 and 0.075.

4.7 The second method. The binomial probabilities are 0.634 and 0.640.

4.9 The expected payoff is $\$5 \times 0.1820 = \$0.91$ per dollar staked.

4.11 Let E be the expected payoff for any newly purchased ticket. The equation $E = 5{,}000 \times (3.5 \times 10^{-5}) + 50 \times 0.00168 + 5 \times 0.03025 + 0.2420 \times E$ gives $E = \$0.541$. The house percentage is 45.9%.

4.13 Apply the multinomial distribution with $n = 5$, $p_1 = \frac{1}{6}$, $p_2 = \frac{1}{6}$, and $p_3 = \frac{2}{3}$. The probabilities are 0.5355 and 0.4930.

4.15 Use the multinomial distribution to find the house percentage of 12.3%.

4.17 The number of winners is approximately Poisson distributed with an expected value of $\lambda = 200 \times \frac{25}{2{,}500{,}000}$. The monthly amount the corporation will have to give away is zero with probability 0.9980 and \$25,000 with probability 0.002.

4.19 This is a birthday problem with a group of 100,000 persons and $\binom{45}{6} = 8{,}145{,}060$ birthdays. Letting $\lambda = \binom{100{,}000}{8} \frac{1}{(8{,}145{,}060)^7}$, the Poisson approximation for the desired probability is $1 - e^{-\lambda} = 1.04 \times 10^{-13}$.

4.21 Using the approach from Section 4.2.3 for the birthday problem, a trial is associated with each sample of three persons from the group of 25 persons. This leads to the Poisson approximations $1 - e^{-\lambda_0} = 0.0171$ and $1 - e^{-\lambda_1} = 0.1138$, where $\lambda_0 = \binom{25}{3} \times \left(\frac{1}{365}\right)^2$ and $\lambda_1 = \binom{25}{3} \times \left(7 \times \left(\frac{1}{365}\right)^2\right)$. The simulated values of the desired probabilities are 0.0164 and 0.1030.

4.23 Letting $\lambda = \binom{25}{2} \times \left(\frac{14}{365} \times \frac{1}{365}\right)$, a Poisson approximation for the desired probability is $1 - e^{-\lambda} = 0.0310$. The simulated value is 0.0299.

4.25 Letting $\lambda_0 = 44 \times \left[\binom{43}{4}/\binom{45}{6}\right]$ and $\lambda_1 = 43 \times \left[\binom{42}{3}/\binom{45}{6}\right]$, the Poisson approximations are 0.487 and 0.059. The simulated values are 0.529 and 0.056.

4.27 A remarkably accurate approximation is given by the Poisson distribution with expected value $\lambda = 8 \times \frac{1}{15} = \frac{8}{15}$ (see also the answer to Problem 2.39). In the Poisson approximation approach, the ith trial concerns your prediction of the ith match. The success probability of each trial is given by $p = \frac{8 \times 2 \times 14!}{16!} = \frac{1}{15}$.

4.29 Letting $\lambda = 365 \times \left[1 - \left(\frac{354}{365}\right)^{75} - 75 \times \frac{1}{365} \times \left(\frac{364}{365}\right)^{74}\right] = 6.6603$, a Poisson approximation for the probability of having at least 7 days on which two or more employees have birthdays is $1 - \sum_{k=0}^{6} e^{-\lambda}\lambda^k/k! = 0.499$. The simulated value is 0.516.

4.31 The probability is $1 - e^{-\alpha} \times e^{-\alpha} = 1 - e^{-2\alpha}$. Use the property that the Poisson process has independent increments.

4.33 The simulated value is 0.203.

4.35 The number of illegal parking customers is binomially distributed with parameters $n = 75$ and $p = 5/(45 + 5) = \frac{1}{10}$. The desired probability is 0.0068.

4.37 Apply the hypergeometric model with $R = 122$, $W = 244$ and $n = 31$. The desired probability is 0.0083.

4.39 The probabilities are 7.15×10^{-9}, 6.44×10^{-8}, 4.29×10^{-7}, 1.80×10^{-5}, 4.51×10^{-5}, 9.23×10^{-4}, 0.00123, and 0.01667. The probability of not winning the jackpot in the coming m years is $1 - e^{-52\times 12\times m\times p}$ with $p = \frac{1}{10} \times \frac{1}{\binom{49}{6}}$. Putting this probability equal to 0.5 yields $m = 155{,}334$ years.

Chapter 5

5.1 Statement (b).

5.3 $\Phi(\frac{550-799.5}{121.4}) = 0.0199$.

5.5 $1 - \Phi(\frac{20}{16}) = 0.1056$.

5.7 If Y is distributed as $2X$, then $\sigma(Y) = 2\sigma(X)$.

5.9 (a) The correlation coefficient is -1. (b) Invest $\frac{1}{2}$ of your capital in stock A and $\frac{1}{2}$ in stock B. The expected value of the rate of return is 7% and the standard deviation is zero. In other words, the portfolio has a guaranteed rate of return of 7%.

5.11 For the case of $p = 0.5$ and $f = 0.2$ the simulated probability mass function is given by (0.182, 0.093, 0.047, 0.029, 0.022, 0.016, 0.014, 0.012, 0.011, 0.011, 0.010, 0.011, 0.011, 0.012, 0.012, 0.014, 0.013, 0.015, 0.029, 0.001, 0.433).

5.13 For the case of $p = 0.8$ and $f = 0.1$, the simulated values of the expected value and the standard deviation of the investor's capital after 20 years are about \$270,000 and \$71,000. For the case of $p = 0.5$ and $f = 0.2$, the simulated values are about \$430,000 and \$2,150,000.

5.15 This value converges to $\frac{1}{2}$, since $P(X \geq \mu) = \frac{1}{2}$ for any $N(\mu, \sigma^2)$ random variable X.

5.17 The Poisson model is applicable. A Poisson distribution with mean 81 can be approximated by an $N(81, 81)$ distribution. The observed value of 117 lies 4 standard deviations above the expected value of 81. This is difficult to explain as a chance variation.

5.19 The observed value of 70 lies 3 standard deviations below the expected value of 70. This is difficult to explain as a chance variation.

5.21 This can hardly be explained as a chance variation. In 1,000 rolls of a fair die the average number of points per roll is approximately $N(\mu, \sigma^2)$ distributed with $\mu = 3.5$ and $\sigma = \frac{1.708}{\sqrt{1000}} = 0.054$. The reported value of 3.25 lies 4.63 standard deviations above the expected value of 3.5.

5.23 The outcome can hardly be explained as a chance variation. It lies 3.93 standard deviations above the expected value of 17.83.

5.25 The probability is about $1 - \Phi(2.8) = 0.093$.

5.27 The 95% confidence interval is $0.52 \pm 1.96\frac{\sqrt{0.52\times 0.48}}{\sqrt{400}} = 0.52 \pm 0.049$. The enlarged sample size must be about 2,400 students.

5.29 Under the hypothesis that the aspirin has no effect, the observed value of 104 for the aspirin group lies $\left(\frac{313}{2} - 104\right)/\left(\frac{1}{2}\sqrt{313}\right) = 5.93$ standard deviations below the expected value of 156.5. This is strong evidence against the null hypothesis.

5.31 Under the hypothesis that the generator produces true random numbers, the number of runs is distributed as $1 + R$, where R has a binomial distribution with parameters $n = 99{,}999$ and $p = \frac{1}{2}$. The observed value of 49,487 runs lies 3.25 standard deviations below the expected value. This is a strong indication that the new random number generator is a bad one.

5.33 Use the fact that the process $\{\ln(S_t/S_0)\}$ is a Brownian motion process with drift parameter $\mu - \frac{1}{2}\sigma^2$ and variance parameter σ^2. Next, apply the formula from Problem 5.32 with $c = \ln(a)$ and $d = -\ln(b)$.

Chapter 6

6.1 Disagree.

6.3 A chance tree leads to the probability $\frac{1}{5}$.

6.5 A chance tree leads to the probability $4 \times (0.2 \times 0.5) = 0.4$.

6.7 A chance tree leads to the probability $\frac{0.375}{0.375+0.075} = 0.8333$.

6.9 A chance tree gives $P(\text{not drunk} \mid \text{positive}) = \frac{0.0475}{0.0475+0.045} = 0.5135$.

6.11 A chance tree gives $P(\text{white cab} \mid \text{white cab seen}) = \frac{0.12}{0.12+0.17} = 0.4138$.

6.13 A chance tree leads to the probability $\frac{1/3}{1/3+1/6} = \frac{2}{3}$.

6.15 The probability is $\frac{1}{1+9\cdot 9} = 0.092$.

6.17 Pick three marbles out of the vase. Guess the dominant color among these three marbles. Under this strategy you win \$8,500 with probability 0.7407.

Chapter 7

7.1 The probability is $\frac{2}{3}$.

7.3 (a) Take the sample space with as elements the 15 teams other than the Johnson team. The desired probability is $\frac{1}{15}$. (b) Take the sample space with as elements the $\binom{16}{2} = 120$ unordered pairs of two teams. The desired probability is $\frac{1}{120}$.

7.5 The sample space is $\Omega = \{(x, y) | 0 \le x \le a, 0 \le y \le a\}$. The probability $P(A) =$ (area of $A)/a^2$ is assigned to each subset A of Ω. The desired probability is $(a - d)^2/a^2$.

7.7 For the case of $q = 1$, take the set $\Omega = \{(x, y) : -1 < x, y < 1\}$ as sample space. Let the subset A consist of the points $(x, y) \in \Omega$ satisfying $y \le \frac{1}{4}x^2$. The desired probability equals $P(A) = \frac{1}{4}(2 + \int_{-1}^{1} \frac{1}{4}x^2 dx) = 0.5417$. In general, the probability of the quadratic equation $x^2 + Bx + C = 0$ having two real roots is $\frac{1}{2} + \frac{q}{24}$ for $0 < q < 4$ and $1 - \frac{2}{3\sqrt{q}}$ for $q \ge 4$.

7.9 The first probability is $(h - d)^2/h^2$. The second probability is $\frac{1}{2} + \frac{\pi}{6\sqrt{3}} = 0.8023$.

7.11 Let $B_n = \cup_{k=n}^{\infty} A_k$ for $n \ge 1$, then $B_1, B_2, \ldots$ is a nonincreasing sequence of sets. Note that $\omega \in C$ if and only if $\omega \in B_n$ for all $n \ge 1$. This implies that set C equals the intersection of all sets B_n. Using the continuity of probabilities, $P(C) =$

$\lim_{n\to\infty} P(B_n)$. This gives $P(C) = \lim_{n\to\infty} P(\cup_{k=n}^{\infty} A_k) \le \lim_{n\to\infty} \sum_{k=n}^{\infty} P(A_k)$. The latter limit is zero, since $\sum_{k=1}^{\infty} P(A_k) < \infty$.

7.13 The desired probability is equal to $\sum_{k=0}^{\infty} \left(\frac{7}{10}\right)^{2k} \frac{3}{10} = 0.5882$.

7.15 The desired probability is $1 - P(A \cup B) = 1 - 0.7 - 0.5 + 0.3 = 0.1$.

7.17 The probability is 0.45.

7.19 Let $A = \{3k \mid 1 \le k \le 333\}$, $B = \{5k \mid 1 \le k \le 200\}$, and $C = \{7k \mid 1 \le k \le 142\}$. The desired probabilities are $P(A \cup B) = \frac{333}{1000} + \frac{200}{1000} - \frac{66}{1000} = 0.467$ and $P(A \cup B \cup C) = \frac{333}{1000} + \frac{200}{1000} + \frac{142}{1000} - \frac{66}{1000} - \frac{47}{1000} - \frac{28}{1000} + \frac{9}{1000} = 0.543$.

7.21 The inclusion-exclusion formula leads to the probability 0.051.

7.23 Take as sample space the collection of all unordered sets of 13 numbers from the numbers $1, \ldots, 52$. The desired probability is $\binom{13}{1} \times \left[\binom{48}{9}/\binom{52}{13}\right] - \binom{13}{2} \times \left[\binom{44}{5}/\binom{52}{13}\right] + \binom{13}{3} \left[\binom{40}{1}/\binom{52}{13}\right] = 0.0342$.

Chapter 8

8.1 Take as sample space the set of four pairs (G, G), (G, F), (F, G), and (F, F), where G stands for a "correct prediction" and F stands for a "false prediction," and the first and second components of each pair refer to the predictions of weather station 1 and weather station 2. The probabilities $0.9 \times 0.8 = 0.72$, $0.9 \times 0.2 = 0.18$, $0.1 \times 0.8 = 0.08$, and $0.1 \times 0.2 = 0.02$ are assigned to the elements (G, G), (G, F), (F, G), and (F, F). The desired probability is $P(\{(G, F)\} | \{(G, F), (F, G)\}) = 0.18/0.26 = 0.692$.

8.3 The probabilities are $\frac{0.4388}{0.6962} = 0.630$ and $\frac{0.1097}{0.25} = 0.439$.

8.5 The desired probability is $\frac{2}{7} \times \frac{2}{6} \times \frac{2}{5} \times \frac{2}{4} \times \frac{2}{3} = 0.0127$. *Remark*: using clever counting arguments, it can be reasoned that the desired probability is $2^6/7!$; however, the derivation using conditional probabilities is simpler.

8.7 Yes: $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2}$.

8.9 The probability is $\frac{5}{6}$.

8.11 Condition on the number of matches among the one million tickets. As shown in Example 7.11, the probability of j matches is $e^{-1}/j!$ for $j \ge 0$. The desired probability is $1 - \sum_j \left[\binom{1{,}000{,}000-j}{500{,}000}\Big/\binom{1{,}000{,}000}{500{,}000}\right] \frac{e^{-1}}{j!} = 0.3935$.

8.13 The recursion is $a_k = \frac{1}{2} a_{k-1} + \frac{1}{4} a_{k-2}$ for $k \ge 2$ with the boundary conditions $a_0 = a_1 = 1$. This leads to $a_5 = 0.4063$, $a_{10} = 0.1406$, $a_{25} = 5.85 \times 10^{-3}$, and $a_{50} = 2.93 \times 10^{-5}$.

8.15 Denote by $p(n)$ the probability that the last passenger will get his/her own seat when the plane has n seats. For the case of $n = N$, let $a_k = \frac{1}{N}$ for $1 \le k \le N$ be the probability that the first passenger takes seat k. Then

$$p(N) = a_1 + \sum_{k=2}^{N-1} p(N-k+1) a_k = \frac{1}{N} \left(1 + \sum_{k=2}^{N-1} p(N-k+1)\right)$$

with $p(2) = \frac{1}{2}$. Applying this recursion, it follows that $p(N) = \frac{1}{2}$ for all $N \ge 2$. *Remark*: the probability that the jth passenger will get his/her own seat can be shown to be equal to $\frac{N-j+1}{N-j+2}$ for $j = 1, 2, \ldots, N$.

8.17 It follows from $\frac{p}{1-p} = \frac{1/5}{4/5} \frac{(0.75)^3}{(0.5)^3} = 0.84375$ that $p = 0.4576$.

Chapter 9

9.1 Take $\Omega = \{(i, j, k, l) | i, j, k, l = 1, \dots, 6\}$ as sample space. The expected payoff is $\$100 \times \frac{6}{1{,}296} + \$10 \times \frac{54}{1{,}296} = \$\frac{95}{108}$.

9.3 The optimal strategy is to stop after the first spin if this spin gives a score of more than 414 points.

9.5 Denote by the random variable X the payoff of the game. Then, $E(X) = 2(\frac{1}{2} - \frac{1}{3}) \times 1 + 2(\frac{1}{3} - \frac{1}{4}) \times 2 + \cdots + 2(\frac{1}{m} - \frac{1}{m+1}) \times (m-1) + \frac{2}{m+1} \times m$.

9.7 $\sum_{k=0}^{\infty} P(X > k) = \sum_{k=0}^{\infty} \sum_{j=k+1}^{\infty} P(X = j)$. Interchanging the order of summation gives $\sum_{k=0}^{\infty} P(X > k) = \sum_{j=0}^{\infty} \sum_{k=0}^{j-1} P(X = j) = \sum_{j=0}^{\infty} j P(X = j)$, proving the desired result.

9.9 Let X_i be equal to 1 if there is a birthday on day i and 0 otherwise. For each i, $P(X_i = 0) = \left(\frac{364}{365}\right)^{100}$ and $P(X_i = 1) = 1 - P(X_i = 0)$. The expected number of distinct birthdays is $365 \times \left(1 - \left(\frac{364}{365}\right)^{100}\right) = 87.6$.

9.11 Let the random variable X_i be equal to 1 if the numbers i and $i+1$ appear in the lotto drawing and 0 otherwise. Then, $P(X_i = 1) = \binom{43}{4} / \binom{45}{6}$ for all i and so $E(X_1 + \cdots + X_{44}) = 44 \times [\binom{43}{4} / \binom{45}{6}] = \frac{2}{3}$.

9.13 The standard deviation is $\sqrt{4/9} = 2/3$.

9.15 The expected value is \$281.00 and the standard deviation is \$555.85.

9.17 The random variables X and Y are dependent (e.g. $P(X = 2, Y = 1)$ is not equal to $P(X = 2)P(Y = 1)$). The values of $E(XY)$ and $E(X)E(Y)$ are given by $\frac{1232}{36}$ and $7 \times \frac{161}{36}$.

9.19 The random variables $X_2, \dots, X_{10}$ are independent, where X_i has the discrete uniform distribution on $0, 1, \dots, i-1$. The expected value and the variance of the sum $X_2 + \cdots + X_{10}$ are 22.5 and 31.25.

9.21 Let X_i denote the number of draws needed to go from i different integers to $i+1$ different integers for $i = 1, 2$. Using the convolution formula, it follows that r draws are needed with probability $P(X_1 + X_2 = r - 1) = \sum_{j=1}^{r-2} \frac{9}{10} (\frac{1}{10})^{j-1} \frac{8}{10} (\frac{2}{10})^{r-2-j}$ for $j \geq 3$.

9.23 Ten dice.

9.25 The probability is $\sum_{k=4}^{7} \binom{k-1}{3} (0.45)^4 (0.55)^{k-4} = 0.3917$. The expected value and the standard deviation are 5.783 and 1.020.

9.27 The binomial distribution with parameters $n = 4 \times 6^{r-1}$ and $p = \frac{1}{6^r}$ converges to a Poisson distribution with expected value $\frac{2}{3}$ as $r \to \infty$.

9.29 The probability is 0.8675.

9.31 The hypergeometric model with $R = W = 25$ and $n = 25$ is applicable under the hypothesis that the psychologist blindly guesses which 25 persons are left-handed. Then, the probability of identifying correctly 18 or more of the 25 left-handers is 2.1×10^{-3}. This small probability provides evidence against the hypothesis.

9.33 The probability is $\binom{2m-k}{m} p^{m+1} (1-p)^{m-k} + \binom{2m-k}{m} (1-p)^{m+1} p^{m-k}$.

Chapter 10

10.1 The probability density function of Y is $\frac{f(\sqrt{y})}{2\sqrt{y}}$ for $y > 0$ and is 0 otherwise. Using the fact that $P(c \leq V \leq d) = \frac{d-c}{2a}$ for any c and d with $-a \leq c \leq d \leq a$, it follows that the density function of W is $\frac{1}{2a\sqrt{w}}$ for $0 < w < a^2$ and is 0 otherwise.

10.3 The random variable V satisfies $P(V \leq v) = P(X \leq v/(1+v)) = \frac{v}{1+v}$ for $v \geq 0$. Its density function is equal to $\frac{1}{(1+v)^2}$ for $v > 0$ and 0 otherwise. The random variable W satisfies $P(W \leq w) = 1 - \sqrt{1-4w}$ for $0 \leq w \leq \frac{1}{4}$ and its density function is equal to $2(1-4w)^{-1/2}$ for $0 < w < \frac{1}{4}$ and 0 otherwise.

10.5 The random variable W has probability density function $g(w) = 2w$ for $0 < w < 1$ and $g(w) = 0$ otherwise. This follows from

$$\begin{aligned} P(W \leq w) &= P(U_1 \leq w,\ U_2 \leq w) = \\ &= P(U_1 \leq w)P(U_2 \leq w) = w^2, \qquad 0 \leq w \leq 1. \end{aligned}$$

The random variable V has probability density function $h(v) = 2(1-v)$ for $0 < v < 1$ and $h(v) = 0$ otherwise. This follows from

$$\begin{aligned} P(V \leq v) &= 1 - P(V > v) = 1 - P(U_1 > v,\ U_2 > v) \\ &= 1 - P(U_1 > v)P(U_2 > v) = 1 - (1-v)^2, \qquad 0 \leq v \leq 1. \end{aligned}$$

10.7 The expected value is $73\frac{1}{3}$ meters.

10.9 Let X be the distance from the point to the origin. Then $P(X \leq a) = \frac{1}{4}\pi a^2$ for $0 \leq a \leq 1$ and $P(X \leq a) = \frac{1}{4}\pi a^2 - 2\int_1^a \sqrt{a^2 - x^2}\, dx = \frac{1}{4}\pi a^2 - a^2 \arccos(\frac{1}{a}) + \sqrt{a^2 - 1}$ for $1 < a \leq \sqrt{2}$. The density function $f(x)$ of X satisfies $f(x) = \frac{1}{2}\pi x$ for $0 < x \leq 1$ and $f(x) = \frac{1}{2}\pi x - 2x \arccos(\frac{1}{x})$ for $1 < x < \sqrt{2}$. Numerical integration leads to $E(X) = \int_0^{\sqrt{2}} x f(x)\, dx = 0.765$.

10.11 The expected value is $\int_0^h x \frac{2(h-x)}{h^2}\, dx = \frac{1}{3}h$.

10.13 The expected value is $\frac{3}{4}r$ and the standard deviation is $0.194r$.

10.15 (a) $E\left[(X-c)^2\right] = E(X^2) - 2cE(X) + c^2$. This expression is minimal for $c = E(X)$. The minimal value is the variance of X. (b) $E(|X - c|) = \int_{-\infty}^c (c - x) f(x) dx + \int_c^\infty (x - c) f(x) dx$. The derivative of this function of c is $2F(c) - 1$, where $F(c) = P(X \leq c)$. Putting the derivative equal to 0 gives $F(c) = \frac{1}{2}$.

10.17 Let the random variable X represent the total demand during the lead time of the replenishment order. Define the function $g(x)$ by $g(x) = x - s$ for $x > s$ and $g(x) = 0$ for $0 \leq x \leq s$. Simple calculations lead to $E\left[g(X)\right] = e^{-\lambda s}/\lambda$ and $E[\left(g(X)\right)^2] = 2e^{-\lambda s}/\lambda^2$. Hence, the expected value and standard deviation of the shortage are given by $\frac{1}{\lambda}e^{-\lambda s}$ and $\frac{1}{\lambda}\left[e^{-\lambda s}(2 - e^{-\lambda s})\right]^{1/2}$.

10.19 The median is 3.

10.21 If X is gamma(α, λ) distributed, then

$$E(X) = \int_0^\infty x \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\, dx = \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} \int_0^\infty \frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)} x^\alpha e^{-\lambda x}\, dx,$$

and so $E(X) = \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} = \frac{\alpha\Gamma(\alpha)}{\lambda\Gamma(\alpha)} = \frac{\alpha}{\lambda}$. Similarly, $E(X^2) = \frac{\Gamma(\alpha+2)}{\lambda^2\Gamma(\alpha)} = \frac{(\alpha+1)\alpha}{\lambda^2}$.

10.23 Letting $S_n = \sum_{i=1}^n D_i$ with $D_i = X_i - \text{round}(X_i)$, the desired probability is given by $P(-\frac{1}{2} \leq S_n < \frac{1}{2})$. The random variables $D_1, \ldots, D_n$ are independent and uniformly distributed on $(-\frac{1}{2}, \frac{1}{2})$.

10.25 Denoting by the random variable F_n the factor at which the size of the population changes in the nth generation, the size of the population after n generations is distributed as $(F_1 \times \cdots \times F_n)s_0$. By the central limit theorem, $\ln(S_n) = \sum_{i=1}^n \ln(F_i) + \ln(s_0)$ has approximately a normal distribution with mean

$n\mu_1 + \ln(s_0)$ and standard deviation $\sigma_1\sqrt{n}$ for n large, where $\mu_1 = 0.5\ln(1.25) + 0.5\ln(0.8) = 0$ and $\sigma_1 = \sqrt{0.5[\ln(1.25)]^2 + 0.5[\ln(0.8)]^2} = 0.22314$. Thus, the probability distribution of S_n can be approximated by a lognormal distribution with parameters $\mu = \ln(s_0)$ and $\sigma = 0.22314\sqrt{n}$.

10.27 (a) $P(I(U) \leq x) = P(U \leq F(x)) = F(x)$ for all x. (b) Apply twice the substitution rule.

Chapter 11

11.1 Let X denote the low points rolled and Y the high points rolled. Then $P(X = i, Y = i) = \frac{1}{36}$ for $1 \leq i \leq 6$ and $P(X = i, Y = j) = \frac{2}{36}$ for $1 \leq i < j \leq 6$.

11.3 The area of the triangle is $\frac{1}{2}$. The joint density function $f(x, y)$ of X and Y equals 2 for (x, y) inside the triangle and 0 otherwise. The random variable $V = X + Y$ satisfies $P(V \leq v) = 2\int_0^v dx \int_0^{v-x} dy$ and so $P(V \leq v) = v^2$ for $0 \leq v \leq 1$. The density function of V is $2v$ for $0 < v < 1$ and 0 otherwise. The random variable $W = \max(X, Y)$ satisfies

$$P(W \leq w) = P(X \leq w, Y \leq w) = 2\int_0^w dx \int_0^{\min(1-x,w)} dy,$$

yielding that $P(W \leq w) = 2w^2$ for $0 \leq w \leq \frac{1}{2}$ and $P(W \leq w) = 4w - 2w^2 - 1$ for $\frac{1}{2} \leq w \leq 1$. The density function of W equals $4w$ for $0 < w < \frac{1}{2}$, $4 - 4w$ for $\frac{1}{2} \leq w < 1$ and 0 otherwise.

11.5 Let X and Y denote the smallest and the largest of the two random numbers. Then, $P(x \leq X \leq x + \Delta x, y \leq Y \leq y + \Delta y) = 2\Delta x \Delta y$ for all $0 < x < y < 1$, showing that the joint density function of X and Y is given by $f(x, y) = 2$ for $0 < x < y < 1$. Denote by V the length of the middle interval. Then, by $P(V \leq v) = \int_0^1 2\,dx \int_x^{\min(1,x+v)} dy$, we have $P(V \leq v) = 2v - v^2$ for $0 \leq v \leq 1$, showing that V has the density function $2 - 2v$ for $0 < v < 1$. The probability that the smallest of the three resulting intervals is larger than a is given by $P(X > a, Y - X > a, 1 - Y > a)$. This probability can be evaluated as $\int_a^{1-2a} 2\,dx \int_{x+a}^{1-a} dy = (1 - 3a)^2$ for $0 \leq a \leq 1/3$.

11.7 The joint density function $f(x, y)$ of X and Y is equal to $4/\sqrt{3}$ for points (x, y) inside the triangle and 0 otherwise. The marginal density function $f_X(x)$ is equal to $\int_0^{x\sqrt{3}} f(x, y)\,dy = 4x$ for $0 < x < \frac{1}{2}$, $\int_0^{(1-x)\sqrt{3}} f(x, y)\,dy = 4(1 - x)$ for $\frac{1}{2} < x < 1$ and 0 otherwise. The marginal density function $f_Y(y)$ is equal to $\int_{y/\sqrt{3}}^{1-y/\sqrt{3}} f(x, y)\,dx = 4/\sqrt{3} - 8y/3$ for $0 < y < \frac{1}{2}\sqrt{3}$ and 0 otherwise.

11.9 The range of Z is $(2c, 2d)$. Using the result from Problem 11.8, we obtain that Z has a triangular density with parameters $a = 2c$, $b = 2d$ and $m = c + d$. The density function of V is $f_V(v) = \frac{1}{4}\int_0^v \frac{1}{\sqrt{v-y}}\frac{1}{\sqrt{y}}dy$ for $0 < v < 2$. Using numerical integration, the expected distance is calculated as 0.752.

11.11 The inverse functions are given by the functions $x = \frac{v}{\sqrt{v^2+w^2}}e^{-\frac{1}{4}(v^2+w^2)}$ and $y = \frac{w}{\sqrt{v^2+w^2}}e^{-\frac{1}{4}(v^2+w^2)}$, and the Jacobian is $\frac{1}{2}e^{-\frac{1}{2}(v^2+w^2)}$.

11.13 Using the marginal densities $f_X(x) = \frac{4}{3}(1 - x^3)$ for $0 < x < 1$ and $f_Y(y) = 4y^3$ for $0 < y < 1$, we obtain $E(X) = \frac{2}{5}$, $E(Y) = \frac{4}{5}$, $\sigma^2(X) = \frac{14}{225}$, $\sigma^2(Y) = \frac{2}{75}$, and $E(XY) = \frac{1}{3}$. This leads to $\rho(X, Y) = 0.3273$.

11.15 E(annual rainfall) $= 779.5$ mm and σ(annual rainfall) $= 121.4$ mm. The desired probability is $1 - \Phi(\frac{1{,}000-799.5}{121.4}) = 0.049$.

11.17 Use the relations $P(X_i = 1) = \binom{1}{1}\binom{R+W-1}{n-1}/\binom{R+W}{n} = \frac{n}{R+W}$ for all i and $P(X_i = 1, X_j = 1) = \binom{2}{2}\binom{R+W-2}{n-2}/\binom{R+W}{n} = \frac{n(n-1)}{(R+W)(R+W-1)}$ for all i, j with $j \neq i$.

11.19 $\rho(X, Y) = \frac{441/36-(91/36)(161/36)}{(1.40408)^2} = 0.479$.

Chapter 12

12.1 It suffices to prove the result for the standard bivariate normal distribution. Also it is no restriction to take $b > 0$. Let $W = aX + bY$. Differentiating

$$P(W \leq w) = \frac{1/b}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} dx\Big[\int_{-\infty}^{(w-ax)/b} e^{-\frac{1}{2}(x^2-2\rho xy+y^2)/(1-\rho^2)}\, dy\Big]$$

yields that the density function of W is given by

$$f_W(w) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[x^2-2\rho x(w-ax)/b+(w-ax)^2/b^2]/(1-\rho^2)}\, dx.$$

This expression for $f_W(w)$ can be reduced to $(\eta\sqrt{2\pi})^{-1}\exp(-\frac{1}{2}w^2/\eta^2)$ with $\eta = \sqrt{a^2 + b^2 + 2ab\rho}$. Since $X - Y$ is $N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$ distributed, it follows that $P(X > Y) = 1 - \Phi(-(\mu_1 - \mu_2)/(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)^{1/2})$.

12.3 $P(Z \leq z) = \int_0^{\infty} dy \int_{-\infty}^{yz} f(x, y)\, dx + \int_{-\infty}^{0} dy \int_{yz}^{\infty} f(x, y)\, dx$. Differentiation leads to $f_Z(z) = \int_0^{\infty} y f(yz, y)\, dy - \int_{-\infty}^{0} y f(yz, y)\, dy$. Hence $f_Z(z) = \int_{-\infty}^{\infty} |y|\, f(yz, y)\, dy$. Inserting the standard bivariate normal density for $f(x, y)$ and using the results of Example 10.7, the desired result follows.

12.5 Any linear combination of V and W is a linear combination of X and Y and thus is normally distributed. This shows that (V, W) has a bivariate normal distribution.

12.7 Any linear combination of $X + Y$ and $X - Y$ is a linear combination of X and Y and thus is normally distributed. Hence, the random vector $(X + Y, X - Y)$ has a bivariate normal distribution. The components $X + Y$ and $X - Y$ are independent if $\text{cov}(X + Y, X - Y) = 0$. We have $\text{cov}(X + Y, X - Y) = \text{cov}(X, X) - \text{cov}(X, Y) + \text{cov}(X, Y) - \text{cov}(Y, Y)$ and so $\text{cov}(X + Y, X - Y) = \sigma^2(X) - \sigma^2(Y) = 0$.

12.9 Go through the path of length n in opposite direction and next continue this path with m steps.

12.11 The vector (X_1, X_2) has a bivariate normal distribution. Use the fact that $aX_1 + bX_2$ is normally distributed for all constants a and b.

12.13 The observed value of test statistic D is 0.470. The probability $P(\chi_3^2 > 0.470) = 0.925$. The agreement with the theory is very good.

12.15 The value of the test statistic D is 20.848. The probability $P(\chi_6^2 > 20.848) = 0.00195$. This is a strong indication that the tickets are not randomly filled in.

Chapter 13

13.1 Condition on the unloading time. The probability of no breakdown is given by $\int_{-\infty}^{\infty} e^{-\lambda y} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2/\sigma^2}\, dy = e^{-\mu\lambda+\frac{1}{2}\sigma^2\lambda^2}$.

13.3 Using a conditioning argument, the probability of having k red balls among the r selected balls is given by $\sum_{n=0}^{B}[\binom{n}{k}\binom{B-n}{r-k}/\binom{B}{r}]\binom{B}{n}p^n(1-p)^{B-n} = \binom{r}{k}p^k(1-p)^{r-k}$. This result can be understood as follows. Suppose that the B balls are originally noncolored, r balls are chosen, and each of these r balls acquires the color red with probability p.

13.5 The desired probability $P(B^2 \geq 4AC)$ can be calculated as

$$\int_0^1 P\left(AC \leq \frac{b^2}{4}\right) db = \int_0^1 db \left[\int_0^1 P\left(C \leq \frac{b^2}{4a}\right) da\right]$$

$$= \int_0^1 db \left[\int_0^{\frac{b^2}{4}} 1\, da + \int_{\frac{b^2}{4}}^1 \frac{b^2}{4a}\, da\right] = \int_0^1 db \left[\frac{b^2}{4} - \frac{b^2}{4}\ln\left(\frac{b^2}{4}\right)\right].$$

This leads to $P\left(B^2 \geq 4AC\right) = \frac{5}{36} + \frac{1}{6}\ln(2) = 0.2544$.

13.7 By conditioning on Y and using the fact that $P(U \leq u) = u$ for $0 < u < 1$,

$$P\left(U \leq \frac{f(Y)}{cg(Y)}\right) = \int_{-\infty}^{+\infty} \frac{f(y)}{cg(y)} g(y)dy = \frac{1}{c}.$$

Also, $P\left(Y \leq x, U \leq \frac{f(Y)}{cg(Y)}\right) = \int_{-\infty}^{x} \frac{f(y)}{cg(y)} g(y)dy = \frac{1}{c}\int_{-\infty}^{x} f(y)dy$.

13.9 The desired probability is $\frac{2}{3}(1-\Phi(\frac{Q-\mu_1}{\sigma_1})) + \frac{1}{3}(1-\Phi(\frac{Q-\mu_2}{\sigma_2}))$. The expected value of the shortage is $\frac{2}{3}\sigma_1 I(\frac{Q-\mu_1}{\sigma_1}) + \frac{1}{3}\sigma_2 I(\frac{Q-\mu_2}{\sigma_2})$, where $I(k)$ is the so-called normal loss integral $\frac{1}{\sqrt{2\pi}}\int_k^{\infty}(x-k)e^{-\frac{1}{2}x^2}\, dx (= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}k^2} - k[1-\Phi(k)])$. The expected value of the number of gallons left over is equal to the expected value of the shortage minus $\frac{2}{3}\mu_1 + \frac{1}{3}\mu_2 - Q$.

13.11 Condition on the spread. The probability of a spread of i points is given by $\alpha_i = [(12-i)\times 4\times 4\times 2]/(52\times 51)$ for $i = 0, 1, \ldots, 11$. Define the constants $\gamma_1 = 5,\ \gamma_2 = 4,\ \gamma_3 = 2$ and $\gamma_i = 1$ for $i \geq 4$. Then

$$\begin{aligned} E(\text{stake}) &= 10 + \sum_{i=7}^{11} \alpha_i \times 10 = \$11.81 \\ E(\text{payoff}) &= \sum_{i=1}^{6} \alpha_i \times \frac{4i}{50} \times \gamma_i \times 10 + \sum_{i=7}^{11} \alpha_i \times \frac{4i}{50} \times \gamma_i \times 20 \\ &\quad + \alpha_0 \times 10 + 13 \times \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times 120 = \$10.93. \end{aligned}$$

The house percentage is 7.45%.

13.13 For fixed n, let $u_k(i) = E\,[X_k(i)]$. The goal is to find $u_n(0)$. Apply the recursion $u_k(i) = \frac{1}{2}u_{k-1}(i+1) + \frac{1}{2}u_{k-1}(i)$ for i satisfying $\frac{i}{n-k} \leq \frac{1}{2}$, and use the boundary conditions $u_0(i) = \frac{i}{n}$ and $u_k(i) = \frac{i}{n-k}$ for $i > \frac{1}{2}(n-k)$ and $1 \leq k \leq n$. The desired probability $u_n(0)$ has the values 0.7083, 0.7437, 0.7675, and 0.7761 for $n = 5$, 10, 25, and 50. *Remark*: $u_n(0)$ tends to $\frac{\pi}{4}$ as n increases without bound.

13.15 The optimal strategy is to stop after the first spin if this spin gives a score larger than $\sqrt{2}-1$. Your expected payoff is \$609.48 under the optimal strategy.

Chapter 14

14.1 The Bernoulli distribution with parameter p has $1 - p + pz$ as generating function and so the generating function of the binomially distributed random variable X with parameters n and p is given by $G_X(z) = (1 - p + pz)^n$. A negative binomially distributed random variable X with parameters r and p has generating function $G_X(z) = [pz/(1 - (1 - p)z)]^r$.

14.3 The number of record draws is distributed as $R = X_1 + \cdots + X_r$, where X_i equals 1 if the ith draw is a record draw and 0 otherwise. For each i, $P(X_i = 1) = \frac{1}{i}$ and $P(X_i = 0) = 1 - \frac{1}{i}$. The random variables $X_1, \ldots, X_r$ are independent (the proof of this fact is not trivial). This leads to $G_R(z) = z(\frac{1}{2} + \frac{1}{2}z) \cdots (1 - \frac{1}{r} + \frac{1}{r}z)$. The expected value and variance of R are given by $\sum_{i=1}^{r} 1/i$ and $\sum_{i=1}^{r} (i - 1)/i^2$.

14.5 By conditioning on N, we find that $G_S(z) = E(z^S)$ is given by

$$\sum_{n=0}^{\infty} E(S \mid N = n)P(N = n) = z^0 P(N = 0) + \sum_{n=1}^{\infty} E(z^{X_1 + \cdots + X_n})$$
$$P(N = n) = \sum_{n=0}^{\infty} [A(z)]^n e^{-\mu} \frac{\mu^n}{n!} = e^{-\mu[1 - A(z)]}.$$

Differentiating $G_S(z)$ gives the expressions for $E(S)$ and $\text{var}(S)$.

14.7 By $E(z^X) = pzE(z^X) + q + rE(z^X)$, we have $\sum_{k=0}^{\infty} P(X = k)z^k = q/(1 - pz - r)$. Writing $q/(1 - pz - r)$ as $(q/(1 - r))/(1 - pz/(1 - r))$ and using the expansion $1/(1 - pz/(1 - r)) = \sum_{k=0}^{\infty} (\frac{p}{1-r})^k z^k$, we obtain by equating terms that $P(X = k) = \frac{q}{1-r}(\frac{p}{1-r})^k$ for all $k \geq 0$.

14.9 The generating function $G_X(z)$ satisfies $G_X(z) = \frac{1}{2}z + \frac{1}{2}z[G_X(z)]^2$ and thus $G_X(z) = \frac{1}{z} - \frac{1}{z}\sqrt{1 - z^2}$. By $\lim_{z \to 1} G_X'(z) = \infty$, we have $E(X) = \infty$.

14.11 The generating function of the offspring distribution is $P(u) = \frac{1}{3} + \frac{2}{3}u^2$. (a) To find u_3, iterate $u_n = P(u_{n-1})$ starting with $u_0 = 0$. This gives $u_1 = P(0) = \frac{1}{3}$, $u_2 = P(\frac{1}{3}) = \frac{1}{3} + \frac{2}{3}\left(\frac{1}{3}\right)^2 = \frac{11}{27}$, and $u_3 = P(\frac{11}{27}) = \frac{1}{3} + \frac{2}{3}\left(\frac{11}{27}\right)^2 = 0.4440$. (b) The equation $u = \frac{1}{3} + \frac{2}{3}u^2$ has roots $u = 1$ and $u = \frac{1}{2}$. The probability $u_\infty = \frac{1}{2}$. (c) The probabilities are $u_3^2 = 0.1971$ and $u_\infty^2 = 0.25$.

14.13 The function $\left(\frac{\lambda}{\lambda-t}\right)^{\alpha_1} \cdots \left(\frac{\lambda}{\lambda-t}\right)^{\alpha_n} = \left(\frac{\lambda}{\lambda-t}\right)^{\alpha_1 + \cdots + \alpha_n}$ is the moment-generating function of the random variable $X_1 + \cdots + X_n$.

14.15 Let (X, Y) have a bivariate normal density with parameters μ_1, μ_2, σ_1^2, σ_2^2, and ρ. First, consider the special case of a random vector (Z_1, Z_2) having a standard bivariate normal distribution with parameter ρ. It is then a matter of simple algebra to verify the following expression for $E(e^{vZ_1 + wZ_2})$:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{vx + wy} \frac{1}{2\pi\sqrt{1 - \rho^2}} e^{-\frac{1}{2}(x^2 - 2\rho xy + y^2)/(1-\rho^2)} \, dx \, dy$$
$$= \int_{-\infty}^{\infty} e^{vx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \int_{-\infty}^{\infty} e^{wy} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1 - \rho^2}} e^{-\frac{1}{2}(y - \rho x)^2/(1-\rho^2)} dy$$
$$= e^{\frac{1}{2}(v^2 + 2\rho vw + w^2)} \qquad \text{for } -\infty < v, w < \infty,$$

using twice the fact that $E(e^{tU}) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ for an $N(\mu, \sigma^2)$-distributed random variable U. Next, consider the general case. Letting $Z_1 = (X - \mu_1)/\sigma_1$ and $Z_2 =$

$(Y-\mu_2)/\sigma_2$ and noting that

$$E[e^{vX+wY}] = E[e^{v\sigma_1 Z_1 + v\mu_1 + w\sigma_2 Z_2 + w\mu_2}] = e^{v\mu_1 + w\mu_2} E[e^{v\sigma_1 Z_1 + w\sigma_2 Z_2}],$$

we find that the answer to part (a) is

$$M_{X,Y}(v,w) = e^{v\mu_1 + w\mu_2 + \frac{1}{2}(v^2\sigma_1^2 + 2vw\rho\sigma_1\sigma_2 + w^2\sigma_2^2)}.$$

For part (b), let (X,Y) have a joint distribution with $\mu_1 = E(X), \mu_2 = E(Y), \sigma_1^2 = \sigma^2(X), \sigma_2^2 = \sigma^2(Y)$ and $\rho = \rho(X,Y)$. By assumption, the random variable $vX + wY$ is $N(v\mu_1 + w\mu_2, v^2\sigma_1^2 + 2vw\rho\sigma_1\sigma_2 + w^2\sigma_2^2)$ distributed for any constants v and w. Hence, again using the relation $E(e^{tU}) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ for an $N(\mu,\sigma^2)$-distributed random variable U,

$$E(e^{vX+wY}) = e^{v\mu_1 + w\mu_2 + \frac{1}{2}(v^2\sigma_1^2 + 2vw\rho\sigma_1\sigma_2 + w^2\sigma_2^2)}.$$

This proves the desired result with an appeal to the result of part (a) and the uniqueness property of the moment-generating function.

14.17 If $t < 0$, then $P(X \leq c) = P(tX \geq tc) = P(e^{tX} \geq e^{tc})$. Next apply Markov's inequality.

Chapter 15

15.1 Let X_n denote the number of type-1 particles in compartment A after the nth transfer. The process $\{X_n\}$ is a Markov chain with state space $I = \{0, 1, \ldots, r\}$. The one-step transition probabilities are given by $p_{i,i-1} = \frac{i^2}{r^2}$, $p_{ii} = \frac{2i(r-i)}{r^2}$, $p_{i,i+1} = \frac{(r-i)^2}{r^2}$, and $p_{ij} = 0$ otherwise.

15.3 The process $\{Y_n\}$ is always a Markov chain, but the process $\{U_n\}$ is not necessarily a Markov chain. A counterexample is provided by the Markov chain $\{X_n\}$ with state space $I = \{-1, 0, 1\}$ and one-step transition probabilities $p_{00} = 1$, $p_{10} = p_{1,-1} = \frac{1}{2}$, $p_{-1,-1} = 1$, and $p_{ij} = 0$ otherwise.

15.5 Let's say that the system is in state (0, 0) if both machines are good, in state $(0, k)$ if one of the machines is good and the other one is in revision with a remaining repair time of k days for $k = 1, 2$, and in state (1, 2) if both machines are in revision with remaining repair times of one day and two days. Defining X_n as the state of the system at the end of the nth day, the process $\{X_n\}$ is a Markov chain. The one-step transition probabilities are given by $p_{(0,0)(0,0)} = \frac{9}{10}$, $p_{(0,0)(0,2)} = \frac{1}{10}$, $p_{(0,1)(0,0)} = \frac{9}{10}$, $p_{(0,1)(0,2)} = \frac{1}{10}$, $p_{(0,2)(0,1)} = \frac{9}{10}$, $p_{(0,2)(1,2)} = \frac{1}{10}$, $p_{(1,2)(0,1)} = 1$, and $p_{vw} = 0$ otherwise.

15.7 Let's say that the system is in state i if the channel holds i messages (including any message in transmission). If the system is in state i at the beginning of a time slot, the buffer contains $\max(i-1, 0)$ messages. Define X_n as the state of the system at the beginning of the nth time slot. The process $\{X_n\}$ is a Markov chain with state space $I = \{0, 1, \ldots, K+1\}$. In a similar way as in Example 15.4, the one-step transition probabilities are obtained. Let $a_k = e^{-\lambda}\lambda^k/k!$ for $k \geq 0$. Then, $p_{0j} = a_j$ for $0 \leq j \leq K-1$, $p_{0,K} = \sum_{k=K}^{\infty} a_k$, $p_{i,i-1} = (1-f)a_0$ for $1 \leq i \leq K$, $p_{K+1,K} = 1-f$, $p_{ij} = (1-f)a_{j-i+1} + fa_{j-i}$ for $1 \leq i \leq K$ and $i \leq j \leq K$, $p_{i,K+1} = 1 - \sum_{j=i-1}^{K} p_{ij}$ for $1 \leq i \leq K$, and $p_{ij} = 0$ otherwise.

15.9 The probabilities are 0.7440 and 0.7912. The expected value is 10.18.

15.11 The expected value is 49.417.

15.13 $\text{var}(\sum_{t=1}^{n} I_t)$=$\sum_{t=1}^{n}\text{var}(I_t)$+$2\sum_{t=1}^{n}\sum_{u=t+1}^{n}\text{cov}(I_t, I_u)$. The approximate value is calculated as $1-\Phi\left(\frac{240.5-217.294}{12.101}\right)=0.0276$ (the simulated value is 0.0267).

15.15 Let's say that the system is in state i if i different numbers are drawn so far. Define the random variable X_n as the state of the system after the nth drawing. The process $\{X_n\}$ is a Markov chain with state space $I=\{0, 1, \ldots, 45\}$. State 45 is an absorbing state. The one-step transition probabilities are given by $p_{i,i+k} = \binom{45-i}{k}\binom{i}{6-k}/\binom{45}{6}$ for $i=0, 1, \ldots, 44$ and $k=0, 1, \ldots, \min(45-i, 6)$, $p_{45,45}=1$, and $p_{ij}=0$ otherwise. The probability that more than r drawings are needed to obtain all of the numbers $1, 2, \ldots, 45$ is equal to $1-p_{0,45}^{(r)}$. This probability has the values 0.9989, 0.7409, 0.2643, and 0.035 for $r=15$, 25, 35, and 50.

15.17 Let's say that the system is in state (i, j) if i pictures are in the pool once and j pictures are in the pool twice or more. Let X_n be the state of the system in the nth week. The process $\{X_n\}$ is a Markov chain with state space $I=\{(i, j)|i, j \geq 0, i+j \leq 25\}$. State $(0, 25)$ is an absorbing state. The one-step transition probabilities are $p_{(i,j),(i,j)}=\frac{j}{25}\times\frac{j}{25}$, $p_{(i,j),(i+1,j)}=2\times\frac{25-i-j}{25}\times\frac{j}{25}$, $p_{(i,j),(i+2,j)}=\frac{25-i-j}{25}\times\frac{24-i-j}{25}$, $p_{(i,j),(i,j+1)}=\frac{i}{25}\times\frac{25-i-j}{25}+\frac{25-i-j}{25}\times\frac{i+1}{25}$, $p_{(i,j),(i-1,j+1)}=\frac{i}{25}\times\frac{j+1}{25}+\frac{j}{25}\times\frac{i}{25}$, $p_{(i,j),(i-2,j+2)}=\frac{i}{25}\times\frac{i-1}{25}$. The probability $P(N>n)$ has the values 0.9288, 0.3395, 0.0648, 0.0105, and 0.0016 for $n=50$, 75, 100, 125, and 150. The expected value of N is 71.4 weeks.

15.19 Let's say that the system is in state i if the last i spins of the wheel showed the same color for $i=0, 1, \ldots, 26$. State 26 is taken as an absorbing state. Let X_n be the state of the system after the nth spin of the wheel. The process $\{X_n\}$ is a Markov chain with one-step transition probabilities $p_{00}=\frac{1}{37}$, $p_{01}=\frac{36}{37}$, $p_{i0}=\frac{1}{37}$, $p_{i,i+1}=p_{i1}=\frac{18}{37}$ for $i=1, 2, \ldots, 25$, $p_{26,26}=1$, and $p_{ij}=0$ otherwise. The probability that in the next n spins of the wheel the same color will come up 26 or more times in a row is given by $p_{0,26}^{(n)}$. This probability has the value 0.00748 for $n=$ 1,000,000.

15.21 Use an absorbing Markov chain with as state the number of coins that are still in your pocket. The probability that exactly r beggars will be favored by you has the values 0.0400, 0.1510, 0.2531, 0.2538, 0.1726, 0.0853, 0.0320, 0.0094, 0.0022, and 0.0004 for $r=1, 2, \ldots, 10$. The expected value and the standard deviation of the number of favored beggars are given by to 3.816 and 1.487.

15.23 The probabilities are 0.7574 and 0.1436.

15.25 The probabilities are 0.2692 and 0.3836.

15.27 The answers are $\frac{1}{3}$ and 0.4286.

15.29 The Markov chain has a unique equilibrium distribution, since it has no two disjoint closed sets. By $\sum_{k=1}^{N} p_{kj}=1$ for all j, we have $\frac{1}{N}=\sum_{k=1}^{N}\frac{1}{N}p_{kj}$ for all j, proving that $\pi_j=\frac{1}{N}$ satisfies $\pi_j=\sum_{k=1}^{N}\pi_k p_{kj}$ for all j.

15.31 (a) The long-run average stock on hand at the end of the week equals $\sum_{j=0}^{S} j\pi_j$=4.387. (b) The long-run average ordering frequency is $\sum_{j=0}^{s-1}\pi_j=0.5005$. (c) The long-run average amount of demand lost per week is given by $L(S)\sum_{j=0}^{s-1}\pi_j+\sum_{j=s}^{S}L(j)\pi_j=0.0938$, where $L(j)=\sum_{k=j+1}^{\infty}(k-j)e^{-\lambda}\lambda^k/k!$ denotes the expected value of the amount of demand lost in the coming week if the current stock on hand just after review is j.

15.33 A circuit board is said to have status 0 if it has failed and is said to have status i if it functions and has the age of i weeks. Let's say that the system is in state (i, j) with $0 \leq i \leq j \leq 6$ if one of the circuit boards has status i and the other one has status j just before any replacement. The one-step probabilities can be expressed in terms of the failure probabilities r_i. For example, for $0 \leq i < j \leq 5$, $p_{(i,j),(i+1,j+1)} = (1 - r_i)(1 - r_j)$, $p_{(i,j),(0,i+1)} = (1 - r_i)r_j$, $p_{(i,j),(0,j+1)} = r_i(1 - r_j)$, $p_{(i,j),(0,0)} = r_i r_j$, and $p_{(i,j),(v,w)} = 0$ otherwise. (a) The long-run proportion of time the device operates properly is $1 - \pi_{(0,0)} = 0.9814$. (b) The long-run average weekly cost is $750\pi_{(0,0)} + 200[\pi_{(0,0)} + \pi_{(6,6)} + \pi_{(0,6)}] + 100 \sum_{j=1}^{5} [\pi_{(0,j)} + \pi_{(j,6)}] = 52.46$ dollars.

15.35 The equilibrium equations are $\pi_0 = p_{10}\pi_1, \pi_j = p_{j-1,j}\pi_{j-1} + p_{jj}\pi_j + p_{j+1,j}\pi_{j+1}$ for $1 \leq j \leq r - 1$, and $\pi_r = p_{r-1,r}\pi_{r-1}$, where $p_{i,i-1} = \frac{i^2}{r^2}$, $p_{ii} = \frac{2i(r-i)}{r^2}$ and $p_{i,i+1} = \frac{(r-i)^2}{r^2}$ for $0 \leq i \leq r$. The desired recurrence relation can next be proved by using induction. By substitution, the hypergeometric distribution for the π_j can be verified.

15.37 Let K be the common number of points in each of the sets $N(i)$. Fix $j, k \in I$ with $j \neq k$. If $k \notin N(j)$, then $p_{jk} = p_{kj} = 0$ and so $e^{-c(j)/T} p_{jk} = e^{-c(k)/T} p_{kj}$. Otherwise, $e^{-c(j)/T} p_{jk}$ is given by $e^{-c(j)/T} \frac{1}{K} \min\left(1, \frac{e^{-c(k)/T}}{e^{-c(j)/T}}\right) = \frac{1}{K} \min(e^{-c(j)/T}, e^{-c(k)/T}) = e^{-c(k)/T} p_{kj}$.

Bibliography

D.J. Aldous and P. Diaconis. "Shuffling cards and stopping times." *The American Mathematical Monthly* **93** (1986): 333–348.

D.J. Bennett. *Randomness*. Cambridge MA: Harvard University Press, 1999.

D. Bernoulli. "Specimen theoriae novae de mensura sortis," *Commentarii Academiae Scientiarum Imperalis Petropolitanea* **V** (1738): 175–192 (translated and republished as "Exposition of a new theory on the measurement of risk," *Econometrica* **22** (1954): 23–36).

D.A. Berry. *Statistics: A Bayesian Perspective*. Belmont CA: Duxbury Press, 1996.

S. Chu. "Using soccer goals to motivate the Poisson process," *Informs Transactions on Education* **3** (2003): 62–68.

P. Diaconis and F. Mosteller. "Methods for studying coincidences." *Journal of the American Statistical Association* **84** (1989): 853–861.

B. Efron. "Bayesians, frequentists and scientists." *Journal of the American Statistical Association* **100** (2005): 1–5.

G. Gigerenzer. *Calculated Risks*. New York NY: Simon & Schuster, 2002.

J.A. Hanley. "Jumping to coincidences: defying odds in the realm of the preposterous." *American Statistician* **46** (1992): 197–202.

G.P. Harmer and D. Abbott. "Losing strategies can win by Parrondo's paradox." *Nature* **402** (1999, 23/30 December): 864.

N. Henze and H. Riedwyl. *How to Win More*. Natick MA: A.K. Peters, 1998.

T.P. Hill. "The difficulty of faking data." *Chance Magazine* **12** (1999): 27–31.

D. Kadell and D. Ylvisaker. "Lotto play: the good, the fair and the truly awful." *Chance Magazine* **4** (1991): 22–25.

D. Kahneman, P. Slovic and A. Tversky. *Judgment under Uncertainty: Heuristics and Bias*. Cambridge MA: Cambridge University Press, 1982.

E. Maor. *e: The Story of a Number*. Princeton NJ: Princeton University Press, 1994.

R. Matthews. "Ladies in waiting." *New Scientist* **167** (2000, July 29): 40.

K. McKean. "Decisions, decisions, ..." *Discover* (June 1985): 22–31.

J.F. Merz and J.P. Caulkins. "Propensity to abuse-propensity to murder?" *Chance Magazine* **8** (1995): 14.

N. Metropolis and S. Ulam. "The Monte Carlo method." *Journal of the American Statistical Association* **44** (1949): 335–341.

J.I. Nauss. "An extension of the birthday problem." *The American Statistician* **22** (1968): 27–29.

J.A. Paulos. *Innumeracy: Mathematical Illiteracy and its Consequences*. New York NY: Vintage Books, 1988.

J.A. Paulos. *Mathematician Plays the Stock Market*. New York NY: Basic Books, 2003.

S. Savage. "The flaw of averages." *San Jose Mercury News* (October 8, 2000).

G. Székely and D. Richards. "The St. Petersburg paradox and the crash of high-tech stocks in 2000." *American Statistician* **58** (2004): 225–231.

R.H. Thaler. *The Winner's Curse, Paradoxes and Anomalies in Economic Life*. Princeton NJ: Princeton University Press, 1992.

M. vos Savant. *The Power of Logical Thinking*. New York NY: St. Martin's Press, 1997.

M. vos Savant. "Ask Marilyn." *Parade* (February 7, 1999).

G. Weiss. "Random walks and their applications." *American Scientist* **71** (1983, January-February): 65–70.

W.A. Whitworth. *Chance and Choice*. Cambridge: Deighton Bell, 3rd edition, 1886.

Index