A Math Book

کانال A Math Bookبرای استفاده علاقمندان و اساتید و دانشجویان از کتب ریاضی و فیزیک و نشر و انتشار آنهاست.لطفاجهت استفاده همه علاقمندان و دانشجویان لینک کانال را در گروهایتان نشر داده و کتب کانال را فوروارد کنید.
کتابهای آنالیز عددی ، آنالیز ریاضی ، هندسه ، توابع مختلط و نظریه اعداد ، هندسه منیفلد ، ریاضیات کاربردی گرایش تحقیق در عملیات ، جبر ، جبر خطی ، جبر لی ، آموزش ریاضی ، کتابهای المپیاد ریاضی، رمز نگاری و کد گذاری و ریاضیات گسسته و معادلات دیفرانسیل و نظریه گراف و تاریخ ریاضیات و آمار و احتمال و فیزیک کوانتوم،مکانیک و استاتیک و فیزیک حرارتی و فیزیک پلاسما،فیزیک هسته ای و .....منبعی با بیش از 2000 فایل کتابهای ریاضی و فیزیک.

https://telegram.me/AMathbook

Telegram

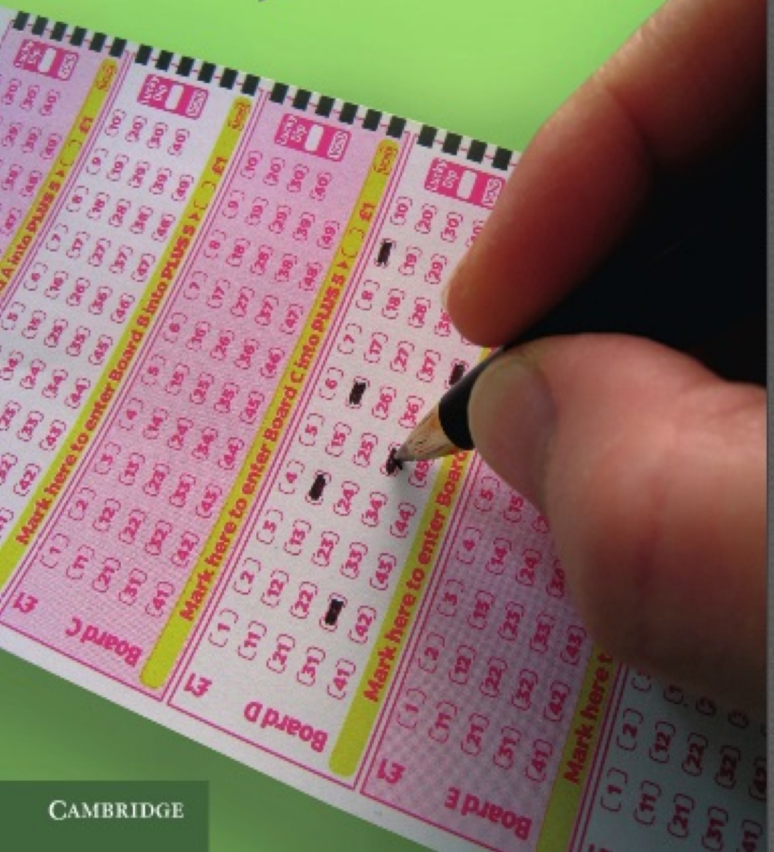
A Math Book

Channel math and physics books

THIRD EDITION

Understanding Probability

Henk Tijms



CAMBRIDGE

Understanding Probability

Understanding Probability is a unique and stimulating approach to a first course in probability. The first part of the book demystifies probability and uses many wonderful probability applications from everyday life to help the reader develop a feel for probabilities. The second part, covering a wide range of topics, teaches clearly and simply the basics of probability.

This fully revised Third Edition has been packed with even more exercises and examples, and it includes new sections on Bayesian inference, Markov chain Monte Carlo simulation, hitting probabilities in random walks and Brownian motion, and a new chapter on continuous-time Markov chains with applications. Here you will find all the material taught in an introductory probability course. The first part of the book, with its easy-going style, can be read by anybody with a reasonable background in high school mathematics. The second part of the book requires a basic course in calculus.

HENK TIJMS is Emeritus Professor at Vrije University in Amsterdam. He is the author of several textbooks and numerous papers on applied probability and stochastic optimization. In 2008 Henk Tijms received the prestigious INFORMS Expository Writing Award for his publications and books.

Understanding Probability

Third Edition

HENK TIJMS
Vrije University, Amsterdam

CAMBRIDGE
UNIVERSITY PRESS

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

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/9781107658561

© H. Tijms 2004, 2007, 2012

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2004
Second edition 2007
Third edition 2012

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data
Tijms, H. C.
Understanding probability / Henk Tijms. – 3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-107-65856-1 (pbk.)
1. Probabilities. 2. Mathematical analysis. 3. Chance. I. Title.
QA273.T48 2012
519.2 – dc23 2012010536

ISBN 978-1-107-65856-1 Paperback

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party internet websites referred to
in this publication, and does not guarantee that any content on such
websites is, or will remain, accurate or appropriate.

Contents

	Preface	*page* ix
	Introduction	1
	PART ONE: PROBABILITY IN ACTION	9
1	**Probability questions**	11
2	**Law of large numbers and simulation**	18
2.1	Law of large numbers for probabilities	19
2.2	Basic probability concepts	28
2.3	Expected value and the law of large numbers	34
2.4	Drunkard's walk	38
2.5	St. Petersburg paradox	41
2.6	Roulette and the law of large numbers	44
2.7	Kelly betting system	46
2.8	Random-number generator	52
2.9	Simulating from probability distributions	57
2.10	Problems	65
3	**Probabilities in everyday life**	75
3.1	Birthday problem	76
3.2	Coupon collector's problem	83
3.3	Craps	86
3.4	Gambling systems for roulette	90
3.5	Gambler's ruin problem	93
3.6	Optimal stopping	95

3.7 The 1970 draft lottery 98
3.8 Problems 102

4 Rare events and lotteries 108
4.1 Binomial distribution 109
4.2 Poisson distribution 111
4.3 Hypergeometric distribution 129
4.4 Problems 136

5 Probability and statistics 142
5.1 Normal curve 144
5.2 Concept of standard deviation 152
5.3 Square-root law 160
5.4 Central limit theorem 162
5.5 Graphical illustration of the central limit theorem 166
5.6 Statistical applications 168
5.7 Confidence intervals for simulations 171
5.8 Central limit theorem and random walks 177
5.9 Brownian motion 186
5.10 Falsified data and Benford's law 196
5.11 Normal distribution strikes again 202
5.12 Statistics and probability theory 203
5.13 Problems 206

6 Chance trees and Bayes' rule 212
6.1 Monty Hall dilemma 213
6.2 Test paradox 218
6.3 Problems 222

PART TWO: ESSENTIALS OF PROBABILITY 227

7 Foundations of probability theory 229
7.1 Probabilistic foundations 230
7.2 Compound chance experiments 239
7.3 Some basic rules 243

8 Conditional probability and Bayes 256
8.1 Conditional probability 256
8.2 Law of conditional probability 266

8.3 Bayes' rule in odds form 270
8.4 Bayesian statistics – discrete case 278

9 Basic rules for discrete random variables 283
9.1 Random variables 283
9.2 Expected value 286
9.3 Expected value of sums of random variables 290
9.4 Substitution rule and variance 293
9.5 Independence of random variables 299
9.6 Important discrete random variables 303

10 Continuous random variables 318
10.1 Concept of probability density 319
10.2 Expected value of a continuous random variable 326
10.3 Substitution rule and the variance 328
10.4 Important probability densities 333
10.5 Transformation of random variables 353
10.6 Failure rate function 357

11 Jointly distributed random variables 360
11.1 Joint probability mass function 360
11.2 Joint probability density function 362
11.3 Marginal probability densities 367
11.4 Transformation of random variables 374
11.5 Covariance and correlation coefficient 377

12 Multivariate normal distribution 382
12.1 Bivariate normal distribution 382
12.2 Multivariate normal distribution 391
12.3 Multidimensional central limit theorem 393
12.4 Chi-square test 399

13 Conditioning by random variables 404
13.1 Conditional distributions 404
13.2 Law of conditional probability for random variables 411
13.3 Law of conditional expectation 418
13.4 Conditional expectation as a computational tool 422
13.5 Bayesian statistics – continuous case 428

14 Generating functions 435
14.1 Generating functions 435
14.2 Moment-generating functions 443
14.3 Chernoff bound 448
14.4 Strong law of large numbers revisited 450
14.5 Central limit theorem revisited 454
14.6 Law of the iterated logarithm 456

15 Discrete-time Markov chains 459
15.1 Markov chain model 460
15.2 Time-dependent analysis of Markov chains 468
15.3 Absorbing Markov chains 472
15.4 Long-run analysis of Markov chains 481
15.5 Markov chain Monte Carlo simulation 490

16 Continuous-time Markov chains 507
16.1 Markov chain model 507
16.2 Time-dependent probabilities 516
16.3 Limiting probabilities 520

Appendix Counting methods and e^x 532
Recommended reading 538
Answers to odd-numbered problems 539
Bibliography 556
Index 558

Preface

Why do so many students find probability difficult? Could it be the way the subject is taught in so many textbooks? When I was a student, a class in topology made a great impression on me. The teacher asked us not to take notes during the first hour of his lectures. In that hour, he explained ideas and concepts from topology in a non-rigorous, 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 is written very much in the same spirit. It first helps you develop a "feel for probabilities" before presenting the more formal mathematics. The book 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. 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. The book is organized into two parts. Part One is informal, using many thought-provoking examples and problems from the real world to help the reader understand what probability really means. Probability can be fun and engaging, but this beautiful branch of mathematics is also indispensable to modern science. The basic theory of probability – including topics that are usually not found in introductory probability books – is covered in Part Two. Designed for an introductory probability course, this can be read independently of Part One. The book can be used in a variety of disciplines ranging from finance and engineering to mathematics and statistics. As well as for introductory courses, the book is also suitable for self-study. The prerequisite knowledge for Part Two is a basic course in calculus, but Part One can be read by anybody with a reasonable background in high school mathematics.

This 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 for clarifying and enlivening the basic concepts of probability. For this reason, computer simulation is used to give 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 lie at the center in the first part of the book. Many of the examples used to illustrate these themes deal with lotteries and casino games.

Good exercises are an essential part of each textbook. Much care has been paid to the collecting of problems that will enhance your understanding of probability. You will be challenged to think about ideas, rather than simply plug numbers into formulas. As you work through them, you may find that probability problems are often harder than they first appear. It is particularly true for the field of probability that a feeling for the subject can only be acquired by going through the process of learning-by-doing.

New to this edition

This edition continues the well-received structure of the book, where Part One is informal and gives many probability applications from daily life, while Part Two teaches clearly and simply the mathematics of probability theory. Revisions and additions are in particular made in Part Two in order to serve better the requirements for classroom use of the book. The most significant additions are:

- more material on coincidences in gambling games (Chapters 3 and 4)
- hitting probabilities in random walks and Brownian motion with applications to gambling (Chapters 3 and 5)
- a new section on optimal stopping (Chapter 3)
- more material on Kelly betting (Chapter 5)
- a new section on Bayesian statistics for the continuous case (Chapter 13)
- a new section on the law of the iterated logarithm (Chapter 14)
- a new section on Markov chain Monte Carlo simulation (Chapter 15)
- a new chapter on continuous-time Markov chains (Chapter 16).

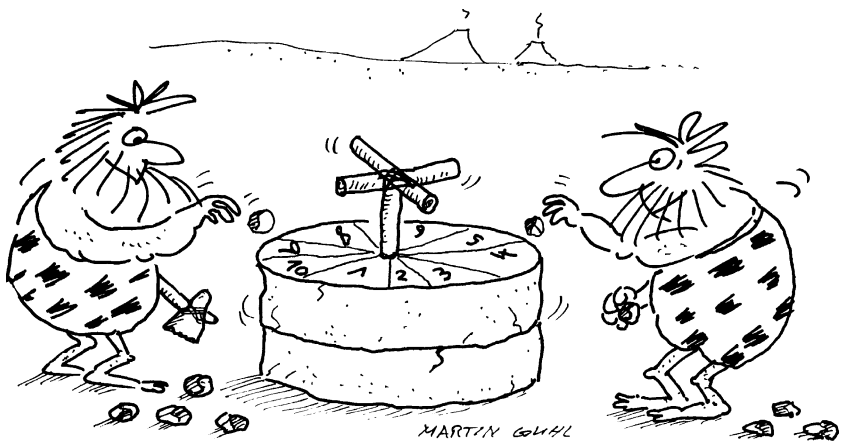
Also, many new exercises are added. A detailed solution manual is available from the author to instructors who adopt this book as a course text.

Finally, thanks to all those who have made valuable comments on the text, particularly Ted Hill (Georgia Institute of Technology), Karl Sigman (Columbia University), Virgil Stokes (Uppsala University) and my colleagues Rein Nobel, Ad Ridder and Aart de Vos at the Vrije University.

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



The cartoons in the book are supplied by www.cartoonstock.com

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 (1629–1695) 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

[†] 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.

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 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. Even for experienced probabilists, it is often difficult to say whether a probability problem is too hard to solve analytically. However, computer simulation always works when you want to get a numerical answer to such a problem. To illustrate this, consider the problem of finding the probability that any two adjacent letters are different when the eleven letters of the word Mississippi are put in random order. Seemingly, a simple probability problem. However, it turns

out that the analytical solution of this problem is very difficult to obtain, whereas an accurate estimate of the answer is easily obtained by simulation, see Section 2.9.4.

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. Also, the theory of probability is used to put coincidences in a broader context. Nearly all coincidences can be explained using probabilistic reasoning.

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

[†] 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.

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. In Chapter 5 we also have a first acquaintance with Bayesian statistics. The English clergyman Thomas Bayes (1702–1761) laid the foundation for this branch of statistics. The Bayesian approach is historically the original approach to statistics, predating what is nowadays called classical statistics by a century. Astronomers have contributed much to the Bayesian approach. In Bayesian inference 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 inference plays a very important role as well. The Bayesian approach is increasingly used in modern science. In the second part of the book we delve more deeply into this approach.

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.

Part Two of the book is designed for an introductory probability course and consists of Chapters 7–16. These chapters can be studied independently of Part One. Chapter 7 states the axioms of probability and derives several basic rules from the axioms. These rules include the inclusion-exclusion rule and are illustrated with many examples. Chapter 8 delves into the concept of conditional probability which lies at the heart of probability theory. The law of conditional probability and Bayes' rule for revising conditional probabilities in light of new information are discussed in depth. Chapter 9 addresses discrete random variables and introduces the concepts of expected value and variance of a random variable. Rules for the expected value and variance of a sum of random variables are treated, including the square-root rule for the standard deviation of the sum of independent random variables. Also, Chapter 9 introduces the most important discrete probability distributions such as the binomial, the Poisson and the hypergeometric distributions. Chapter 10 addresses continuous random variables and explains the concept of probability density, always a difficult concept for the beginner to absorb. Also, this chapter provides insight into the most important probability densities. In particular, the normal density and its role in the central limit theorem are discussed. Whereas Chapter 10 deals

with the probability distribution of a single random variable, Chapter 11 treats joint probability distributions for two or more dependent random variables and introduces the concepts of covariance and correlation. The multivariate normal distribution is the most important joint probability distribution and is the subject of Chapter 12. The multidimensional central limit theorem is treated together with several applications, including the chi-square test. Chapter 13 defines conditional probability densities and states the law of conditional expectation. Practical applications of the theory are given. Also, Bayesian inference for continuous models is discussed. The Bayesian approach is inextricably bound up with conditional probabilities. Chapter 14 deals with the method of moment-generating functions. This useful 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 Chapters 15 and 16 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. Chapter 15 deals with discrete-time Markov chains in which state changes can only occur at fixed times. The basic theory of Markov chains is presented. The theory is illustrated with many examples, including the application of absorbing Markov chains to the computation of success run probabilities in a sequence of independent trials. Also, the powerful method of Markov chain Monte Carlo simulation is treated in Chapter 15. Chapter 16 deals with continuous-time Markov chains in which the times between state changes are continuously distributed. This versatile probability model has numerous applications to queueing systems and several of these applications are discussed. Each of the chapters is accompanied by carefully designed homework problems. The text includes several hundred homework problems. The answers to the odd-numbered problems are given at the end of the book.

PART ONE

Probability in action



1
Probability questions

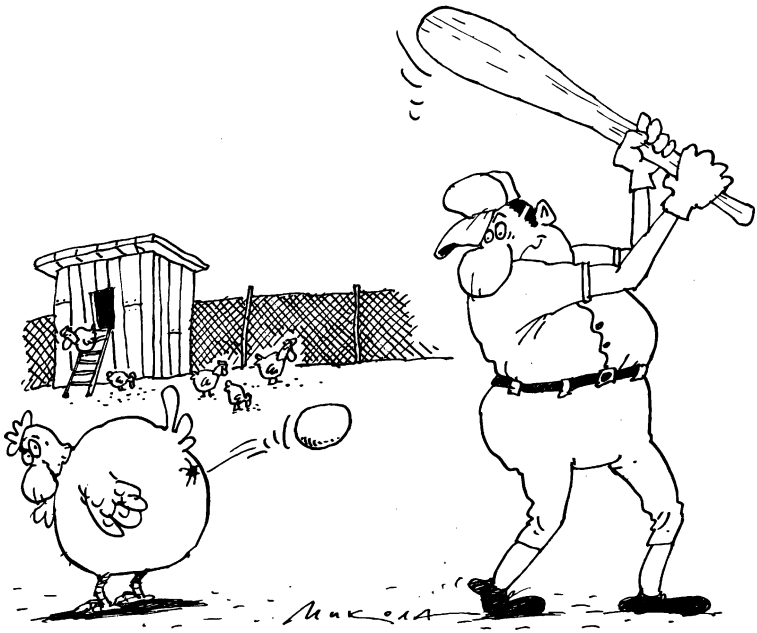
In this chapter, we provide a number of probability problems that challenge the reader to test his or her feeling for probabilities. As stated in the introduction, it is possible to fall wide of the mark when using intuitive reasoning to calculate a probability, or to estimate the order of magnitude of a probability. To find out how you fare in this regard, it may be useful to try one or more of these twelve 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. Solving probability problems usually requires creative thinking, more than technical skills. All of the solutions to the probability questions posed in this chapter 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, §5.10.1)

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? (§8.3)

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 (§12.4)

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, §3.6)

Your friend proposes the following wager: twenty 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 twenty 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. An offer you can't refuse − or can you? (§9.6.3, §10.4.7)

The New York State Lottery offers the game called Quick Draw. The game can be played in bars, restaurants, bowling areas and other places. A player chooses four numbers from 1 to 80. The lottery then randomly choose twenty numbers from 1 to 80. The payoffs on an one-dollar bet are $55 for four matches, $5 for three matches, $2 for two matches and $0 otherwise. In November of 1997, the state lottery offered a promotion "Big Dipper Wednesday" where payoffs on

the game were doubled on the four Wednesdays in that month. Is this a good deal for the player or just a come-on for a sucker bet?

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 Section 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 Oldsmobile cars 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 translation step to the birthday problem is to imagine that each of the 500 Oldsmobile cars gets assigned a "birthday" chosen at random from 2.4 million possible "birthdays".

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

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 dice 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 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 is 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

$$K_n = \text{the total number of heads that will appear in the first } n \text{ tosses}$$

and

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

[†] 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	0.1695
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

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 $\left(p = \frac{1}{2}\right)$ and for an unfair coin $\left(p = \frac{1}{6}\right)$. 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 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

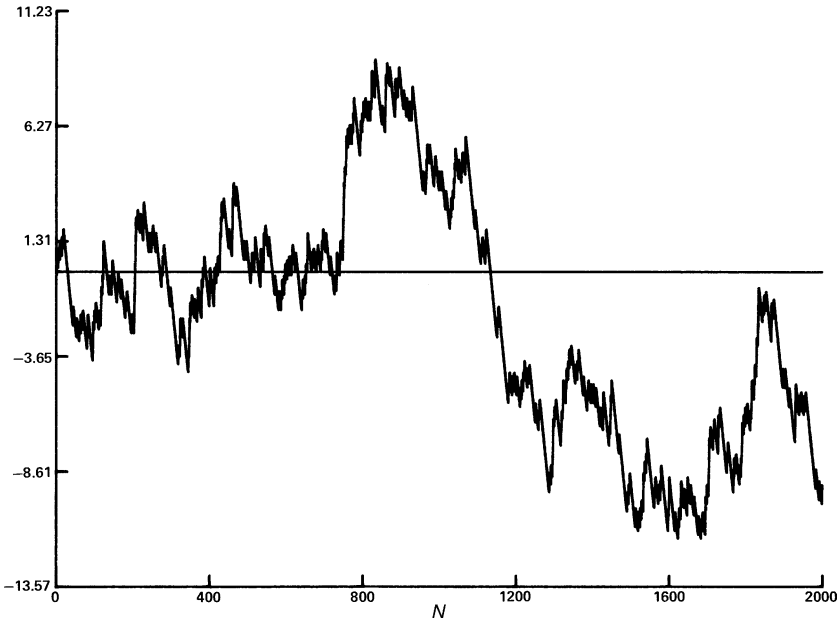


Fig. 2.1. A random walk of 2,000 coin tosses.

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 with the law of large numbers, which says that $\frac{1}{n}\times$ (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 fifty-fifty ratio as n increases and has the property that its spread around this ratio is on the order of $\frac{1}{\sqrt{n}}$.

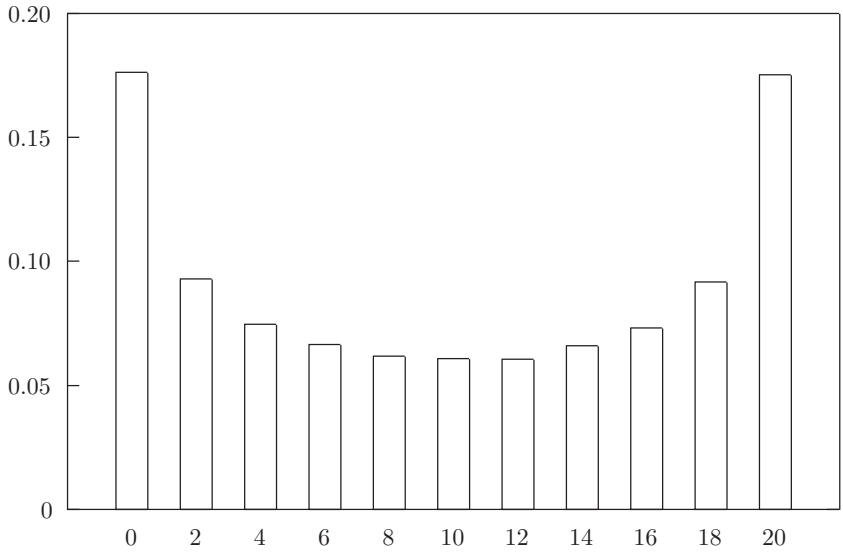


Fig. 2.2. Simulated distribution for 20 tosses of a fair coin.

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 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

† This specialized section may be omitted at first reading.

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 holds true:

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

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 $\frac{4}{\pi}\left\{\arcsin(\sqrt{\beta}) - \arcsin(\sqrt{\alpha})\right\}$. 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 1 of 5 matches, one of the two players is in the lead for more than 97.5% of the time. It also shows that in 1 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.[†]

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

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

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 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 nice illustration 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 $\frac{1}{3} + \frac{1}{6} + \frac{1}{3} = \frac{5}{6}$. 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

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

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 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.

The *union* $A \cup B$ of two events A and B is defined as the set of all outcomes that are either in A or in B, or in both. 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 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

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

[†] 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 $\left(\frac{n(A)+n(B)}{n} = \frac{n(A)}{n} + \frac{n(B)}{n}\right)$.

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 one, "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 9.2). 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 get nothing back). 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 $P(X = 5) = \frac{6}{36}$. 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

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

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 10 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 $20\times 19\times\ldots\times 1$, see also the Appendix. Using the notation

$$n! = 1\times 2\times\ldots\times n,$$

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 10 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 2.9.4.

In Section 3.6 of Chapter 3, we come back to the best-choice problem, and you may be surprised by the solution there. If the number n of papers is sufficiently large (say, $n \geq 100$), then the maximum probability of winning is approximately equal to $\frac{1}{e} = 0.3679$, irrespective of the value of n, where $e = 2.71828\ldots$ is the base of the natural logarithm. The approximately 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 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, etc. 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

$$E(D_m^2) = m$$

for every value of m. A proof of this result will be outlined in Problem 9.26 in Chapter 9. For now, it is worth noting that the result does *not* allow us 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.

[†] see G.H. Weiss, "Random walks and their applications," *American Scientist* **71** (1983): 65–70.

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 approximately given 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} millimeters. 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} millimeters, 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 approximately 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 one both for the drunkard's walk on the line and the drunkard's walk in two dimensions, but it is less than one 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 one 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 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 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

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

[†] 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).

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 \$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

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

[†] 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.

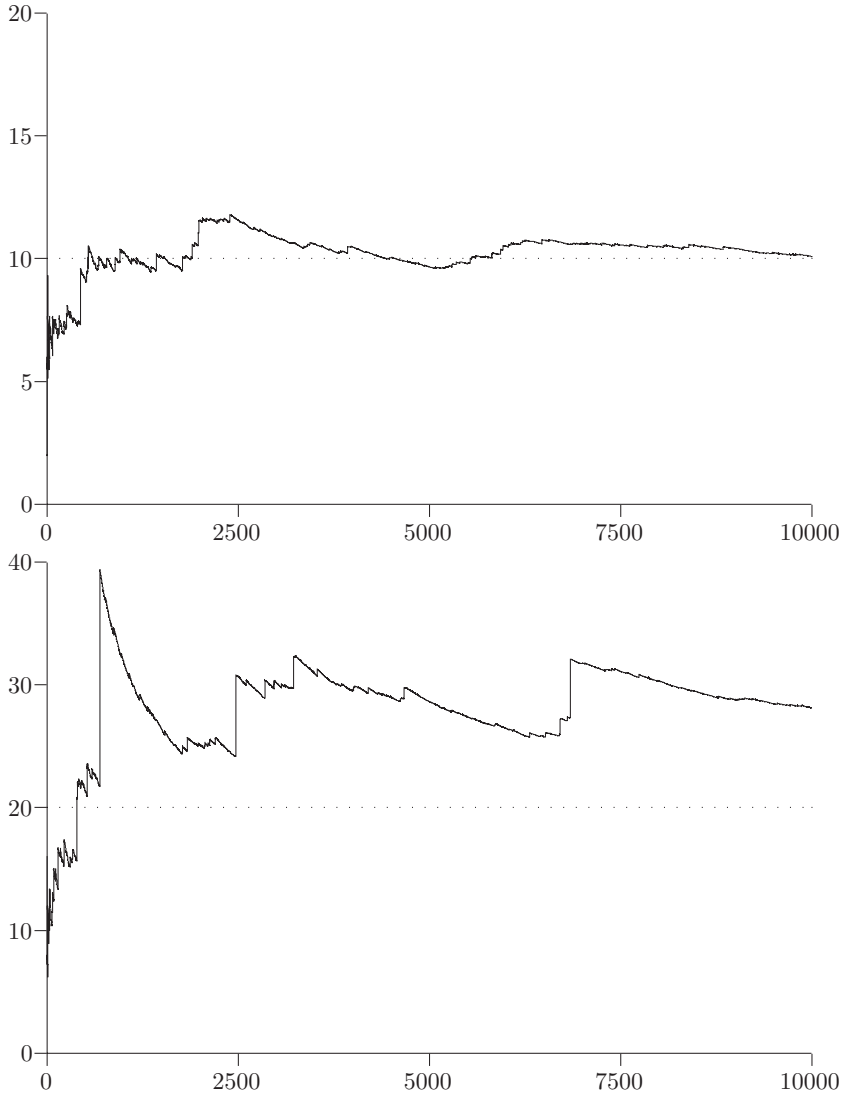


Fig. 2.3. Average payoff in St. Petersburg game.

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.[†]

Roulette is one of the oldest casino games and probably it is of French origin. Roulette is said by some historians to have been invented in 1655 by the French scientist Blaise Pascal. The most common version of roulette uses the numbers $0, 1, \ldots, 36$, 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

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

[†] 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.

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 and 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:

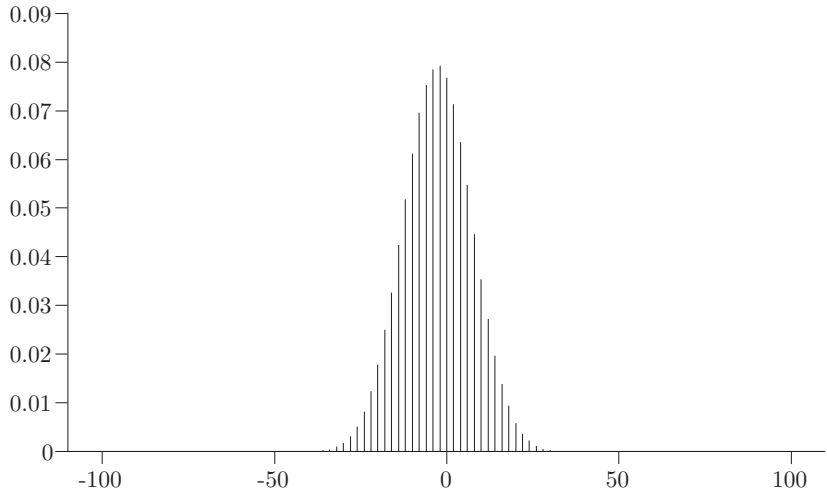


Fig. 2.4. Win/loss calculations for the flat system.

Flat system:

total amount staked	=	100,000,000
total loss	=	2,706,348
loss per unit staked	=	0.0271

Big–Martingale system:

total amount staked	=	384,718,672
total loss	=	10,333,828
loss per unit staked	=	0.0269.

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 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

† This paragraph can be skipped at first reading.

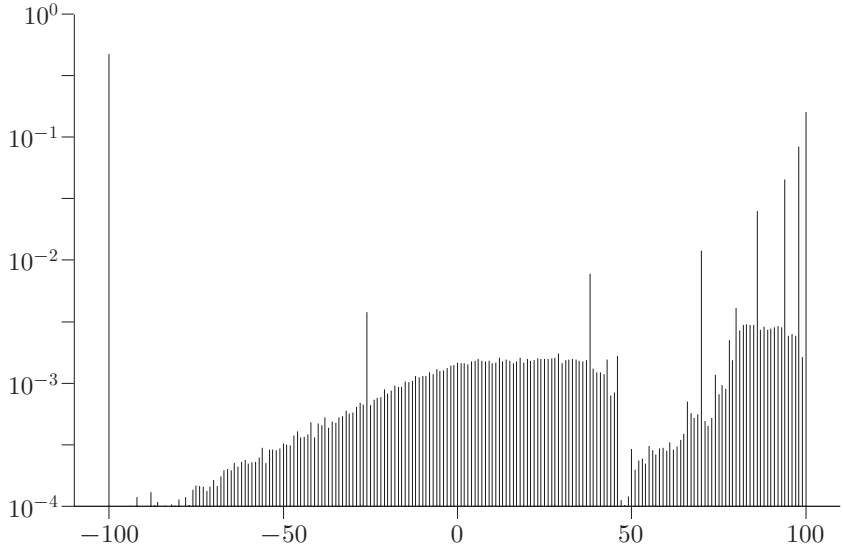


Fig. 2.5. Win/loss calculations for the Big–Martingale system.

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.L. 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. 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 ($0.4 \times 3 + 0.6 \times 0 - 1 > 0$). 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

[†] However, many years before Kelly's publication, W.A. Whitworth already proposed this system in his book *Choice and Chance*, 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 $\exp\left(\frac{1}{n}\sum_{i=1}^{n} \ln(a_i)\right)$.

of f times the amount bet; otherwise, you get nothing back. 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 $\gamma_{\mathit{eff}}(\alpha) = e^{g(\alpha)} - 1$ as $n \to \infty$. The constant $\gamma_{\mathit{eff}}(\alpha)$ gives the effective rate of return for the long run if you bet the same fraction

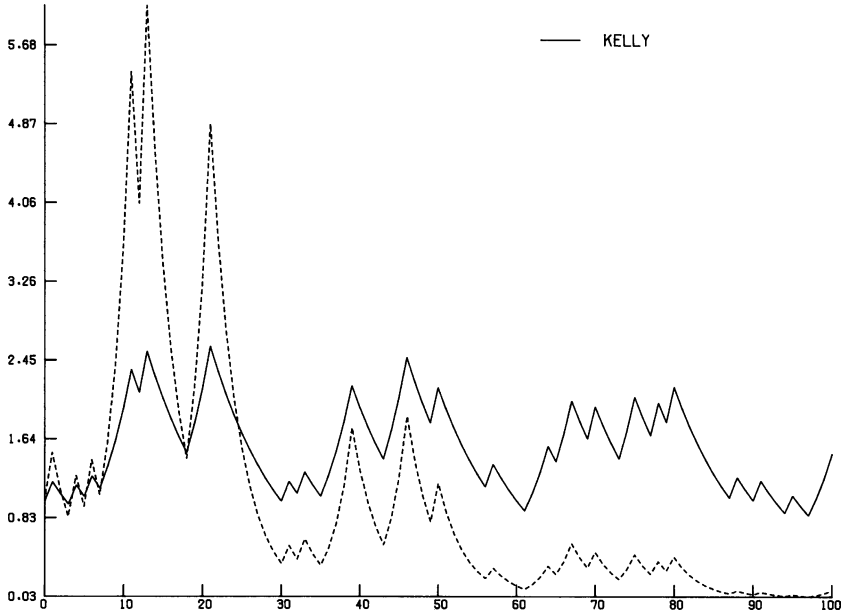


Fig. 2.6. Kelly strategy and an alternative.

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

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

As an illustration, consider the data:

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

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 $\gamma_{\mathit{eff}}(\alpha)$, 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 relation

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

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[†]

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

for any $0 < b < 1 < a$. For example, by betting only half of the Kelly fraction, you give up one-fourth 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.

2.7.3 Derivation of the growth rate

Proof of the relationship

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

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

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

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

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

[†] 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.

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. The method of common random numbers is a simple but very useful technique in probabilistic simulation.

The older random-number generators belong to the class of the so-called multiplicative congruential generators:

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

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 $u_n = \frac{z_n}{m}$. We will not delve into the theory behind this. An understanding of the theory is not necessary in order to use the 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 quality of the multiplicative congruential generator is strongly dependent on the choice of parameters a and m. A much used generator is characterized as $a = 16{,}807$ and $m = 2^{31} - 1$. This generator repeats itself after $m - 1$ values, which is a little over two billion numbers. In the past, this was regarded as plenty. But today, this is not enough for more advanced applications. Nevertheless, the multiplicative congruential generators are still valuable for the simpler applications, despite of the fact that n-dimensional strings of the generated numbers do not pass statistical tests on uniformity in the n-dimensional cube for higher values of n. The newest random-number generators do not use the multiplicative congruential scheme. In fact, they do not involve multiplications or divisions at all. These generators are very fast, have incredibly long periods and provide high-quality pseudo-random numbers. In software tools such as Matlab you will find not only the so-called Christopher

Columbus generator with a cycle length of about 2^{1492} (at ten million random numbers per second, it will take more than 10^{434} years before the sequence of numbers will repeat!), but you will also find the Mersenne twister generator with a cycle length of $2^{19937} - 1$. This generator would probably take longer to cycle than the entire future existence of humanity. It has passed numerous tests for randomness, including tests for uniformity of high-dimensional strings of numbers. The modern generators are needed in Monte Carlo simulations requiring huge masses of random numbers, as is the case in applications in physics and financial engineering.

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, etc., 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 3.7). 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, whereas 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 deck of 52 cards three or four times. This is completely inadequate to achieve a random mix of the cards. Some experienced bridge players are capable of taking advantage of this situation. In professional bridge tournaments and in casinos, computers are being used more and more to ensure a random mix of 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. Roughly speaking, in a riffle shuffle the deck of cards is cut into two more or less equal packets, where the cut is not perfect but involves a random element (the mathematical description of the cut uses the so-called binomial distribution with parameters 52 and 0.5). Once the deck has been cut, the cards from the two packets are interleaved such that the cards from each packet maintain their own relative order (this can be mathematically described by assuming that the cards from the two packets are dropped one at a time such that the next card comes

from the first or second packet with a probability proportional to the number of cards in that packet). It took advanced mathematics to explain the fact that only after seven or more 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 "New Age 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, 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. Let us number the cards in the starting deck, turned face down, in top to bottom order $1 - 2 - \cdots - 26 - 52 - 51 - \cdots - 27$. A yin pile and a yang pile are now made as follows. Riffle shuffle seven times. Then deal the cards one at a time from the top of the deck and place them face up on the table. The yin pile is started as soon as card 1 appears and the yang pile starts as soon as card 27 appears. The cards must be added to the yin pile in the order $1 - 2 - \cdots - 26$ and to the yang pile in the order $27 - 28 - \cdots - 52$. If a card comes up that is not an immediate successor of the top card in either the yin pile or the yang pile, it is placed in a third pile. A single pass through the deck is normally not enough to complete the yin pile or the yang pile. When finished going through the whole deck, take the third pile, turn it face down and repeat the procedure until either the yin pile or the yang pile is completed. Yin wins if the yin pile is completed first. If the deck would have 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 riffle shuffles, it is significantly more likely that the yins will be completed before the yangs. The probability of yin winning can be shown to be 80.7% in this case. This probability is 66.7%, 58.5%, and 54.3%, respectively, after eight, nine, and ten riffle shuffles. Only after fifteen riffle shuffles can we speak of a nearly 50% probability of yin winning. This shows once again the difficulty of getting a fully random mix of the cards by hand.

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

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

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

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

It is instructive to illustrate the above procedure with the birthday problem? What is the probability that in a class of 23 children two or more two children have the same birthday? Assume that the year has 365 days and that all possible birthdays are equally likely. In each run of the simulation model, random

numbers $u_1, \ldots, u_{23}$ are generated and the birthdays $k_i = 1 + \lfloor 365u_i \rfloor$ are computed. The run is said to be successful if $|k_i - k_j| = 0$ for some $i \neq j$. The desired probability is estimated by the ratio of the number of successful runs and the total number of runs. Using 10,000 runs, we obtained the estimate 0.4998 for the probability. The true value of the probability, however, is a little over 50%. The exact value 0.5073 can easily be calculated by an analytical approach, see Section 3.1. The analytical approach becomes very complicated when the problem is slightly changed and one asks for the probability that in a class of 23 children two or more children have a birthday at most one day from each other. On the other hand, the simulation model is hardly affected by the changed problem formulation and needs only a minor adjustment. The test whether a simulation run is successful is now based on $|k_i - k_j| \leq 1$ or $|k_i - k_j| = 364$ for some $i \neq j$, where $|k_i - k_j| = 364$ corresponds to birthdays on January 1 and December 31. Using 10,000 simulation runs, we found the estimate 0.8837, where the exact value is 0.8879.

Another nice illustration of the procedure of drawing a random integer is provided by simulating 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 canceled 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. 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 to assign the mass points x_j 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 the mass points to the buckets 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 $\sum_{j=1}^{l-1} p_j < u \leq \sum_{j=1}^{l} p_j$.

To explain Step 3, note that $b := 1 + \lfloor Bu \rfloor$ corresponds to $\frac{b-1}{B} < u \leq \frac{b}{B}$ and so bucket b contains the point x_l defined by $\sum_{j=1}^{l-1} p_j < u \leq \sum_{j=1}^{l} p_j$. 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

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

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 `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 2; 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 $n - 1$, etc. 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.

Simulation is not a simple gimmick, but you have to build a mathematical model for the simulation. Let us return to the Mississippi problem mentioned in the introductory chapter. What is the probability that any two adjacent letters are different in a random permutation of the eleven letters of the word Mississippi? A simulation model can be constructed by identifying the letter m with the number 1, the letter i with the number 2, the letter s with the number 3, and the letter p with the number 4. A random permutation of the number array $(1, 2, 3, 3, 2, 3, 3, 2, 4, 4, 2)$ can be simulated by using the algorithm above. Note that the initialization of the algorithm now becomes $a[1] = 1$, $a[2] = 2, \ldots, a[10] = 4, a[11] = 2$. To test whether any two adjacent numbers are different in the resulting random permutation $(a[1], a[2], \ldots, a[11])$, you check whether $a[i + 1] - a[i] \neq 0$ for $i = 1, \ldots, 10$. By generating a large number of these random permutations, you obtain an estimate for the desired probability by dividing the number of random permutations in which any two adjacent numbers are different by the total number of random permutations generated. The estimate 0.0582 was obtained by making 100,000 simulation runs. It is never possible to achieve perfect accuracy through simulation. All you can measure is how likely the estimate is to be correct. This issue will be discussed in Section 5.7. Roughly speaking, if you want to achieve one more decimal digit of precision in the estimate of the true probability, you have to increase the number of simulation runs with a factor of about hundred. The probabilistic error bound decreases as the reciprocal square root of the number of simulation runs.

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. This shows how simple it is to simulate a draw from the lotto. The probability of having two consecutive numbers in a draw of the Lotto 6/45 is somewhat more than 50% (the probability is 52.9%). Using simulation, you can easily verify this remarkable fact.

2.9.6 Simulation and probability

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 $x^2 + y^2 \leq 1$. Knowing that the exact value of π is 3.14159 . . ., it is instructive to perform the simulation for several sample sizes. We obtained the estimates 3.24, 3.1536, and 3.1406 using the sample sizes 100, 10,000, and 1,000,000. In general, a large number of simulation runs are required to obtain acceptable precision in the answer. Using the concept of confidence interval, it is possible to estimate the order of magnitude of the statistical error, see Section 5.7. As pointed out before, the probabilistic error bound decreases as the reciprocal square root of the number of simulation runs.

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.

Monte Carlo simulation is often used in simulating physical and mathematical systems in practice, but it is also a very useful tool in the teaching of probability. How to conduct a simulation on your computer? To do this you need to write a computer program. A nice programming environment is provided by the general-purpose software Matlab. Compared to languages such as C or Java, coding in Matlab is much easier, but Matlab programs do not run as fast as those written in a compiled language. Matlab has lots of very nice graphics

routines that you can easily use. Many Matlab tutorials can be found online. Also, the spread-sheet tool Excel and the graphing calculator may be used to perform Monte Carlo simulations of elementary probability problems. In many cases, the simulation methods described in the subsections 2.9.1–2.9.4 suffice. In the problem section below we give a number of geometric probability problems that can be quickly solved by computer simulation but that are otherwise not easily amenable to an analytical solution. In computational probability it is often not clear beforehand whether a probability problem easily allows for an analytical solution. This is also true for simple-looking problems. Also, a small alteration in the formulation of the problem may lead to a much harder problem. Probabilistic simulation is a method that always leads to a numerical answer. However, one should beware of laziness of thinking when such a tool as simulation is available.

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 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 a 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, . . . , 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 fours or three fives 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 twos or three threes 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 $f_1 > 1, 0 \le f_2 < 1$ 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 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.

2.12 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? Use simulation to find these probabilities.

2.13 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.

2.14 Solve Problem 2.13 again when the coefficient A is fixed at the value 1.

2.15 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.

2.16 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 $1 \le k \le m-1$, 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?

2.17 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.

2.18 The popular dice game "drop dead" goes as follows. Each player in turn rolls the five dice and scores when none of the dice thrown show a 2 or a 5. If a 2 or a 5 are not thrown, then the player scores the total of the numbers rolled. If a 2 or 5 is thrown the player scores nothing and puts aside all the dice showing a 2 or 5. These dice are dead and the player continues rolling without them, each time scoring only when no 2s or 5s are rolled and putting aside any dice showing a 2 or a 5. The player's turn ends when all the dice are eliminated. Use simulation to find the expected length and expected total score of a player's turn. Also, simulate the probability that the total score will be more than k points, where $k = 0$, 5, 10, 15, 25, 35, and 50.

2.19 A carnival booth offers the following game of chance. Under each of six inverted cups is a colored ball, in some random order. The six balls are colored red, blue, yellow, orange, green and purple. You wager \$5 to play and you get six tokens. All you have to do is to guess the color of the ball under each of the cups, where you handle one cup at a time. Every time you guess, you risk a token. If your guess is wrong, you lose the token. Each time you guess correctly, the ball is uncovered and you keep your token. If you can guess all six balls before you run out of tokens, you are paid out \$20 for your stake of \$5; otherwise, you lose your stake. Use simulation to find out the house percentage of this game.

2.20 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?

2.21 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 $\frac{1}{37}$?

2.22 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.

2.23 What is the probability that any two adjacent letters are different in a random permutation of the ten letters of the word statistics? 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.

2.24 Suppose that $a(1), a(2), \ldots, a(n)$ is a random permutation of the integers $1, 2, \ldots, n$. Verify experimentally by simulation that the probability of having an up–down permutation has the values 0.1333 and 0.0540 for $n = 5$ and $n = 7$. The permutation $a(1), a(2), \ldots, a(n)$ is said to be an up–down permutation if $a(1) < a(2) > a(3) < a(4) > \cdots$. Also, verify by simulation that the probability of the random permutation satisfying $|a(i) - i| \leq 1$ for all i has the values 0.0667 and 0.0042 for $n = 5$ and $n = 7$.

2.25 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.

2.26 In the best-choice problem from Section 2.3, you now want one of the highest two numbers from 100 slips of paper. Your strategy is to observe the first 34 slips of paper without stopping. Then, observe a stretch of 32 slips and stop if you see a record. If no record appears during that stretch, then you continue until one of the highest two numbers appears. Use simulation to find the probability of getting one of the highest two numbers.

2.27 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 twenty numbers $5, 10, \ldots, 95, 100$ 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. *Remark*: this problem and the next problem are based on the paper "Optimal stopping in the showcase showdown," by P.R. Coe and W. Butterworth, in *The American Statistician* **49** (1995): 271–275.

2.28 Reconsider Problem 2.27 with three candidates A, B and C. Candidate A spins first, candidate B as second and candidate C as last.

(a) Use the optimal stopping rule found in Problem 2.27 to describe the optimal strategy of candidate B.

(b) Use 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?

2.29 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 recounted 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.

2.30 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?

2.31 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.

2.32 A random sequence of 0s and 1s 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 0s or 1s 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 $\log_2(\frac{1}{2}N)$ when N is sufficiently large. *Remark*: More about the longest run in the coin tossing experiment can be found in M.F. Schilling, "The longest run of heads," *The College Mathematics Journal* **21** (1990): 196–207.

2.33 Use simulation to find the probability of getting a run of six different outcomes when rolling a fair die 100 times. Also, find this probability when restricting to the particular run 123456.

2.34 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. 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 $1.18\sqrt{n}$.

2.35 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 $1.65\sqrt{n}$.

2.36 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. Your 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.

2.37 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.

2.38 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.

2.39 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.

2.40 A queue of 50 people is waiting at a box office in order to buy a ticket. The tickets cost five dollars each. For any person, there is a probability of $\frac{1}{2}$ that she/he will pay with a five-dollar note and a probability of $\frac{1}{2}$ that she/he will pay with a ten-dollar 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.

2.41 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 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.

2.42 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).

2.43 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?

2.44 The single-player version of the game of Pig is a dice game in which you want to reach 100 or more points in a minimal number of turns. Each turn, you repeatedly roll a die until either a 1 is rolled or you decide to hold. If a turn ends upon rolling a 1, the sum of the scores in the turn (i.e. the turn total) is lost and nothing is added to your current score. If you decide to hold after having rolled another point than 1, the turn total is added to your current score. Under the hold-at-20 rule you hold when the turn total is 20 or more points with the stipulation that you also hold when the turn total is i or more if your current score lacks i points with $1 \leq i \leq 19$. The rationale behind the hold-at-20 rule: if you put 20 points at stake, your expected loss of $\frac{1}{6} \times 20$ points equals your expected gain of $\frac{5}{6} \times 4$ points. Use computer simulation to find the probability mass function and the expected value of the number of turns needed to reach 100 or more points under the hold-at-20 rule. Also, simulate these performance measures for the five-dice rule in the game of Fast Pig. In this game you have only one roll per turn, but you may roll as many dice as you wish. A turn contributes nothing to your score if one or more of the dice come up with a 1; otherwise, the turn total is added to your score. The five-dice rule prescribes to roll five dice in each turn (the rationale behind this rule: the expected number of points gained in a single turn is maximal when five dice are rolled). Of course, if you still need i points with $1 \leq i \leq 9$, then you roll d_i dice with d_i denoting the smallest integer larger than or equal to $\frac{i}{2}$.

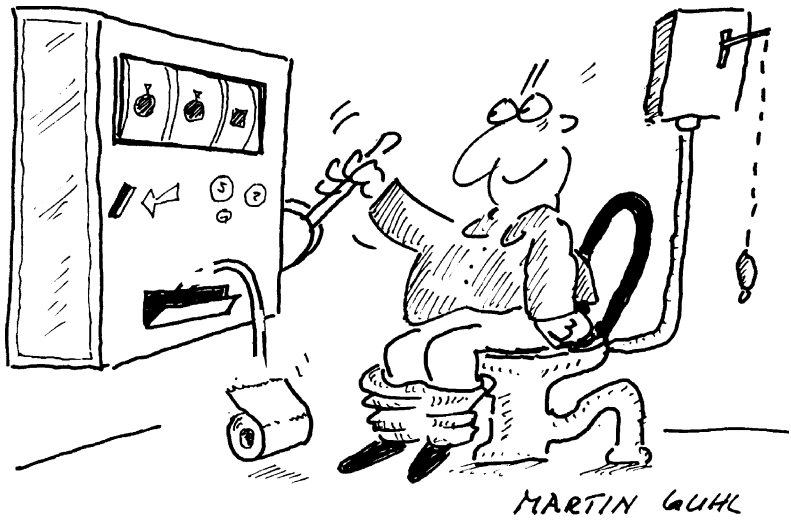
2.45 Consider the following variant of the game of Pig from Problem 2.44. Each turn, you repeatedly roll two dice until either the roll shows a 1 or you hold. In the event of a roll showing a single 1, you lose only the turn total, but in the event of a roll showing a double 1 both the turn total and the current score are lost. You use the following rule: if your current score is i points and the turn total is k, you roll the two dice again if $\frac{1}{36}(i+k) + \frac{10}{36}k < \frac{25}{36} \times 8$ and $i + k < 100$; otherwise, you hold. Use simulation to find the probability mass function and the expected value of the number of turns needed to reach 100 points or more.

2.46 One hundred prisoners are offered a one-time chance to prevent being exiled to a feared prison camp. The names of the 100 prisoners are placed in 100 boxes, one name to a box, and the boxes are lined up on a table in a room. One by one, the prisoners are led into the room. Each prisoner may inspect up to 50 of the boxes to try to find his name. The prisoners must leave behind the room exactly as they found it and they cannot communicate after having left the room. Unless they all find their own names, the whole group will be exiled to the feared prisoner camp. If each prisoner inspects 50 boxes at random, the probability of avoiding the exile of the group is $(1/2)^{100}$ and is thus practically zero. However, the prisoners are allowed to plot a strategy in advance. They decide on a random labeling of the boxes with their own names. Upon entering the room, each prisoner goes to the box labeled with his name. If he finds another prisoner's name in the box, he then looks into the box labeled with that prisoner's name. This process repeats until the prisoner either finds his own name or has inspected 50 boxes. Use simulation to verify that under this strategy the probability of avoiding the exile of the whole group is about 0.312. *Remark*: This problem is taken from Ivars Peterson's column "Puzzling names in boxes," in *Science News*, August 16, 2006. The problem was initiated by the Danish mathematician Peter Bro Miltersen.

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 demonstrate how useful Monte Carlo simulation can be to analyze probability problems. We begin the discussion with the classical birthday problem and the coupon collector's problem. Each of these problems will be solved both by simulation and by an analytical method. Also, the birthday problem is the reason to consider chance coincidences at length. It will be seen that events that looked extremely unlikely were almost to be expected. The casino games of craps and roulette will also be analyzed, using both analytical tools and simulation. Considerable attention is paid to the gambler's ruin problem and optimal stopping in the best-choice problem. Finally, we briefly discuss the bootstrap method which is a powerful simulation method for statistical problems.

3.1 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

[†] 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 10.2

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 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 $1, \ldots, 365$. 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 2.9.2). 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 or more 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 $1, \ldots, n$. 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 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

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 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 $1, \ldots, 49$. 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 $1, \ldots, 49$. 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!

Lottery coincidences abound. An exceptionally striking coincidence happened in the Bulgarian 6/42 Lotto in 2009 after 52 years of lotto history. The numbers 4, 15, 23, 24, 35 and 42 were drawn on September 6 and September 10 in two consecutive drawings. This coincidence drew international news coverage and there were speculations on manipulation. These speculations were also fed by the fact that nobody got all six numbers right in the 6 September drawing, but a record 18 people guessed all six winning numbers in the 10 September drawing. However, some back-of-the-envelope calculations show that coincidences like the one in the Bulgarian Lotto are bound to happen eventually. Imagine observing the Bulgarian 6/42 Lotto for the next 3,000

[†] A humorous story on a similar event in the North Carolina Lottery can be found in L.A. Stefanski, "The North Carolina lottery coincidence," *The American Statistician* **62** (2008): 130–134.

(say) drawings. In each drawing there are $\binom{42}{6} = 5{,}245{,}786$ possible outcomes for the six numbers drawn from the numbers $1, 2, \ldots, 42$. The probability that the same set of six numbers will appear twice in a row somewhere in 3,000 drawings is easily obtained by considering its complementary probability. Letting $N = 5{,}245{,}786$, the probability of the second drawing not repeating the six numbers of the first drawing is $\frac{N-1}{N}$, the probability of the third drawing not repeating the six numbers of the second drawing is $\frac{N-1}{N}$, etc. Thus, by the independence of the drawings, the probability of having no match between the outcomes of two consecutive drawings somewhere in 3,000 drawings is $(\frac{N-1}{N})^{2999} = 0.9994304$. In other words, the probability of the same six numbers twice in a row in 3,000 drawings is 0.0005696. This is still pretty unlikely, but the Bulgarian Lotto is not the only one out there. Let us make the assumption that there are a hundred 6/42 lottos worldwide. It then follows that the probability of seeing a back-to-back repeat in 3,000 drawings somewhere in the world is equal to $1 - (1 - 0.0005696)^{100} = 0.0554$. This 5.5% chance puts the event in the Bulgarian Lotto into context. Surprising, but not unbelievable.

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 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

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

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. 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

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

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. Coincidences are surprising for the person involved, but are not that amazing when you hear they happened to someone else. The New Jersey woman who won the lottery twice in a four-month period must have been perplexed, but the probability of having a double lottery winner somewhere in the world over a four-month period is far from negligible (see Example 4.4 in Chapter 4).

Another example of a noteworthy event concerns the predictions of the outcome of football matches by Paul the Octopus during the World Cup games of 2010. Paul, Germany's oracle octopus, correctly predicted the results of the seven games of the German soccer team and the result of the final between Spain and the Netherlands. This got a lot of news coverage but remarkable predictions as those of Paul are not really surprising viewed in the light of all the other animals that are used to predict sport results or stock prices. We are not hearing about the failed attempts. The excitement surrounding Paul, who became a hero in Spain after his correct prediction of Spain's victory and was offered lifelong protection in Spain against landing on the menu, shows the skewed human perception of chance.

Remarkable events in roulette and craps

On August 18, 1913, black came up twenty-six times in a row on a roulette wheel in the world-famous Casino de Monte Carlo. Can you imagine the chaos, hustle and madness that was going around the table, beginning about the time black had come up fifteen times? People were putting high stakes on red, convinced that the streak of black could not continue. In the end the streak enriched the casino by some millions of francs. How exceptional can we consider a streak of length 26 to be? In 1913, the Monte Carlo casino had been in operation for approximately fifty years. We can roughly estimate that over all of those fifty years, the roulette table had completed between three and five million runs. The probability of the wheel stopping on either red or black twenty-six times in a row in n rounds can be computed to have the value 0.0223 for $n = 3{,}000{,}000$ and the value 0.0368 for $n = 5{,}000{,}000$ (see also Problem 15.19). Thus, it can be said to be exceptional that, in the first fifty years of the existence of the

world's first casino, the roulette wheel stopped twenty-six 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 twenty-five years, a roulette ball will stop on either red or black twenty-six 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.

On May 23, 2009, Patricia DeMauro set a craps record in a casino in Atlantic City by rolling the dice 154 consecutive times in a single turn before she "sevened out" by rolling a seven. She rolled the dice for four hours and 18 minutes, beating the previously reported record set in 1989 by Stanley Fujitake who rolled the dice 119 times at a Las Vegas casino for three hours and six minutes before his turn finally ended. What are the odds? Let us first explain the rules of craps. The player rolling the two dice is called the "shooter". The shooter's turn begins with come-out rolls. These rolls continue until the dice add up to 4, 5, 6, 8, 9, or 10, which establishes the shooter's "point". Once this point is established, the game enters a second stage. The shooter rolls until throwing either the point, which ends the second stage and begins a new set of come-out rolls, or a seven, which ends the shooter's turn. Note that it takes a seven to end the shooter's turn, but the turn cannot end during a come-out roll. This random process can be analyzed by methods described in the Sections 5.9 and 15.3. Using these methods, the probability of not having sevened out after 153 rolls is 1.789×10^{-10}. This is a probability of 1 in 5.59 billion. This certainly is an extremely small probability, but one should take into account the fact that there are many people playing craps at any time. It is estimated that there are about 50 millon shooter's craps turns per year in the U.S., giving about a 1% chance that an event of 154 or more craps rolls in a single turn would occur in a given year. Also, it is interesting to point out that the average number of dice rolls before sevening out is 8.53.

3.2 Coupon collector's problem

In order to introduce a new kind of potato 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 $1, \ldots, 10$ 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 a_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 a_i is equal to $\frac{n-(i-1)}{n}$ and the distribution of Y_i is given by

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

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

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

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 probability 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 probability* 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 probability 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

[†] 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 $P(E_1E_2) = P(E_2 \mid E_1)P(E_1) = \frac{4}{9} \times \frac{5}{10} = \frac{2}{9}$.

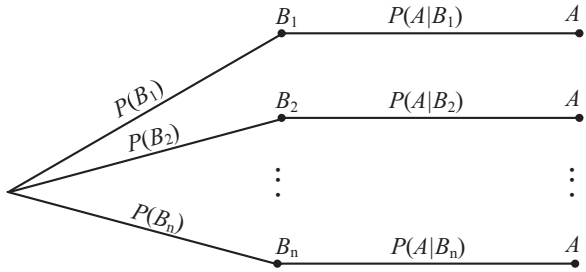


Fig. 3.1. Tree diagram for the law of conditional probabilities.

of 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

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

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 emerged in the early 19th century as a glorious attraction in the casinos of Europe after legalization of the game in 1806 by Napoleon. 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, eighteen are red and eighteen 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 eleven 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 $0, 1, \ldots, 37$. 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 10 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 eleven bets. The player's loss is \$23 if the round goes to eleven bets and the eleventh bet is won; the loss is \$2,023 if the round goes to eleven bets and the eleventh bet is lost. A doubling round goes to eleven bets with a probability of $\left(\frac{19}{37}\right)^{10}$, this being the probability of losing ten bets in a row. This gives

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

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. The same conclusion applies to any other of the mathematical systems that have been devised for roulette (see also Section 2.6).[†] Going back in time, there are players who have broken the bank by detecting biased roulette wheels or by using electronic equipment to predict the path of the ball. The first and most famous biased wheel attack was made in 1873 by the British mechanical engineer Joseph Jagger. He recruited a team of six clerks to record the outcomes of the six roulette wheels at the Monte Carlo casino in Monaco. Jagger detected a biased roulette wheel showing up the nine numbers 7, 8, 9, 17, 18, 19, 22, 28, and 29 far more often than a random wheel would suggest. In a cat-and-mouse game with the casino, Jagger and his team ultimately won two millon francs, which was a fortune for 1873! More recently, gamblers

[†] A nice discussion on roulette systems can be found in the roulette section of the website *The Wizard of Odds* of professor Michael Shackleford.

smuggled into the casino at the Ritz Hotel in London a laser scanner linked to a microcomputer in a mobile phone and using these high-tech devices they won 1.2 million British pounds on the evening of March 16, 2004. The scanner noted where the ball had dropped and measured the declining speed of the wheel. These factors were beamed to the microcomputer which calculated the most likely section of six numbers upon which the ball would finally settle. This information was flashed on to the mobile phone just before the wheel made its third spin, by which time all bets must be placed.

3.5 Gambler's ruin problem

In casino games, the gambler often has the intention to play the game repeatedly until he either increases his bankroll to a predetermined level or goes broke. Imagine a gambler who starts with an initial bankroll of a dollars and then on each successive gamble either wins one dollar or loses one dollar with probabilities p and $q = 1 - p$ respectively. The gambler stops playing after having raised his bankroll to $a + b$ dollars or running out of money, whichever happens first. Here a and b are positive integers. What is the gambler's probability of reaching his desired goal before going broke? This is commonly known as the *gambler's ruin problem*. The progress of the gambler's bankroll forms a random walk in which there is a probability p of moving one unit in the positive direction and a probability q of moving one unit in the other direction. The random walk eventually reaches one of the absorbing states 0 or $a + b$ and then stops. Letting $P(a, b)$ denote the probability of the gambler ending up with $a + b$ dollars, then this probability is given by the gambler's ruin formula

$$P(a, b) = \frac{1 - (q/p)^a}{1 - (q/p)^{a+b}},$$

where $P(a, b)$ should be read as $\frac{a}{a+b}$ if $q = p$. This formula has a long history and goes back to Blaise Pascal (1623–1662) and Christiaan Huygens (1629–1695). A derivation of the gambler's ruin formula can be found in Example 8.4 in Chapter 8.

An interesting application of the gambler's ruin formula concerns the Zarin case. In 1980 David Zarin, a compulsive gambler, was given virtually unlimited credit at a casino in Atlantic City, and ran up a debt of three million dollars. According to New Jersey law, his debt to the casino was not legally enforceable, and the casino settled out of court for a much smaller amount. But then the U.S. government tax-agency IRS claimed taxes on the remainder of the debt,

on the grounds that cancelation of the debt made it taxable income. Since Zarin had never actually received any cash but was only given chips for gambling at the craps table, the court finally ruled that Zarin had no tax obligations. The court never asked what Zarin's credit line was actually worth. This question was addressed in the paper by M. Orkin and R. Kakigi, "What is the worth of free casino credit?," *The American Mathematical Monthly* **102** (1995): 3–8. Using the gambler's ruin formula for the game of craps, it can be argued that the "worth" of a sufficiently large free line of credit is approximately \$195,000. Let us assume that Zarin played the "pass line" bet at craps. The payoff odds of such a bet are one to one and the probability of winning this bet is $p = 244/495$, see Section 3.3. In a casino game with a positive house edge, it is optimal for the gambler to play boldly if he wants to reach his target with maximal probability, where bold play means that the gambler bets as much as possible on each bet or just enough to reach his goal, whichever is less. This mathematically deep result can be intuitively explained by noting that in bold play the gambler exposes his bankroll the least time to the house advantage of the casino. In Zarin's case the house limit for a pass line bet at craps was \$15,000. Let us also assume that Zarin made \$15,000 pass line bets. To apply the gambler's ruin formula, we take \$15,000 as unit. Imagine that Zarin's goal was to win an additional b units on top of the starting capital of $a = 3{,}000{,}000/15{,}000 = 200$ units. The gambler incurs no debt if he goes broke and otherwise leaves the casino with b units in his pocket. Hence the expected utility of the free credit line of a units can be defined as

$$u(a, b) = 0 \times (1 - P(a, b)) + b \times P(a, b),$$

where $P(a, b) = (1 - (q/p)^a)/(1 - (q/p)^{a+b})$, by the gambler's ruin formula. It is reasonable to take $\max_b u(a,\ b)$ as the "worth" of the credit line of a units. For sufficiently large a, the function $u(a, b)$ is maximal for

$$b \approx \frac{1}{\ln(q/p)},$$

which is independent of a. Furthermore, the maximal expected utility is approximately equal to

$$\frac{e^{-1}}{\ln(q/p)},$$

where $e = 2.71828\ldots$ is the base of the natural logarithm.

A sketch of the derivation of these two results goes as follows. Writing $P(a, b) = (1 - (q/p)^a)/(1 - (q/p)^{a+b})$ as $[(q/p)^{-a} - 1]/[(q/p)^{-a} - (q/p)^b]$ and noting that $(q/p)^{-a}$ tends to 0 as a tends to infinity (use the fact

that $q/p > 1$), it follows that $u(a, b) = bP(a, b)$ tends to $b(q/p)^{-b}$ as a tends to infinity. Next it is a matter of calculus to verify that $b(q/p)^{-b}$ is maximal for $b \approx 1/\ln(q/p)$ and that $P(a,\ 1/\ln(q/p)) \approx e^{-1}$ for large a. It is a remarkable result that the probability of raising the bankroll with $b \approx 1/\ln(q/p)$ units before running out of money is always about 37% when the starting capital is sufficiently large.

Applying the above formulas to the Zarin case with $a = 200$, $p = 244/495$ and $q = 251/495$, we find a maximizing value of $b = 35.355$ units and a maximal expected utility of 13 units. Each unit represents \$15,000. This implies that the "worth" of the credit line of three million dollars is approximately given by \$195,000.

3.6 Optimal stopping

In real life there is an abundance of decision-making problems in a dynamic but random environment. The best-choice problem is one of the most famous examples. This problem, which is also known as the secretary problem, the Sultan's dowry problem and the googol game was already touched upon in Section 2.3.1. The standard best-choice problem can be formulated as follows. A known number of n items is to be presented one by one in random order. The observer is able at any time to rank the items that have been presented so far from the best to the worst without ties. The decision to accept or reject each item must be based on the relative ranks of the items presented so far. Once an item has been accepted, the process stops. The observer has only one chance to accept or reject each item; he cannot go back and choose one he has previously rejected. If the last item is presented, it must be accepted. What is the observer's optimal strategy in order to maximize the probability of getting the item with the highest rank?

A reasonable strategy is to reject the first s items for some integer $s \geq 0$ in order to obtain a baseline to work with, and then choose the next item which is highest in the relative ranks of the observed items. Let $p(s, n)$ denote the probability of getting the item with the highest rank when using this strategy. The probability $p(s, n)$ can be easily obtained by simulation. The simulation is based on a random permutation of the integers $1, 2, \ldots, n$, imagining that integer 1 corresponds to the item with the highest rank, integer 2 corresponds to the item with the second highest rank, and so on. However, it is also easy to find an explicit expression for $p(s, n)$ by analytical tools. To do so, let A_k be the event that the kth item in the sequence has the highest rank of all n items and let B_k be the event that the item having the highest rank among the first

$k-1$ items appears in the first s items. Some reflections show that $P(A_k) = \frac{1}{n}$ and $P(B_k \mid A_k) = \frac{s}{k-1}$ for $k > s$. Using the basic formula,

$$p(s,n) = \sum_{k=s+1}^{n} P(A_k B_k) = \sum_{k=s+1}^{n} P(B_k \mid A_k) P(A_k),$$

we obtain the explicit expression

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

Let s_n^* denote the value of s for which $p(s,n)$ is maximal and let p_n^* denote the maximal probability of getting the item with the highest rank. Then the following two results can be proved:

(a) s_n^* is the largest value of s for which

$$\frac{1}{s} + \cdots + \frac{1}{n-1} - 1 \geq 0.$$

(b) For n sufficiently large,

$$s_n^* \approx \frac{n}{e},$$

where $e = 2.71828\ldots$ is the base of the natural logarithm. Moreover, p_n^* tends to $\frac{1}{e} = 0.3679$ as n gets very large.[†]

The asymptotic values for s_n^* and p_n^* already give excellent approximations for moderate values of n. The switching level s_n^* has the values 7 and 37 for $n = 20$ and $n = 100$, where the corresponding values for p_n^* are 0.3842 and 0.3710.

What does the optimal strategy become when your goal is to find one of the items of the highest two ranks? Then, the optimal strategy is characterized by two switching levels s_1 and s_2. Under this strategy you observe s_1 items without stopping. Then, observe the items $s_1 + 1, \ldots, s_2$ and stop if you see an item with the highest rank observed so far. If no record appears during that stretch of items, then continue to the next stage where you stop if you see an item with one of the highest two ranks observed so far. Letting $s_{1,n}^*$ and $s_{2,n}^*$ denote the optimal values of s_1 and s_2 when the number of items is n, the following result

[†] An outline of the proof is as follows. For fixed n, let $\Delta_s = p(s,n) - p(s-1,n)$. It is a matter of simple algebra to verify that $\Delta_s = (1/n)[\sum_{j=s}^{n-1} 1/j - 1]$ and that Δ_s is decreasing in s. This proves the first result. To prove the second result, we use the approximation $\sum_{j=1}^{k-1} 1/j \approx \ln(k) + \gamma$ for sufficiently large k, where $\gamma = 0.57722\ldots$ is Euler's constant. Setting $\Delta_s = 0$, we obtain $\ln(n) - \ln(s_n^*) \approx 1$ for n large, implying that $\ln(n/s_n^*) \approx 1$ and so $n/s_n^* \approx e$. Substitution of $s = n/e$ into $p(s,\, n)$ gives that p_n^* tends to $1/e$ as n gets very large.

can be proved. For n sufficiently large, $s^*_{1,n} \approx 0.3470n$ and $s^*_{2,n} \approx 0.6667n$. The maximal probability of getting one of the items with the highest two ranks tends to 0.5736 as n gets very large. These asymptotic results provide excellent approximations. For $n = 20$, the optimal values of s_1 and s_2 are 7 and 13 with a probability of 0.6046 of getting one of the best two. The optimal strategy for $n = 100$ items has the parameter values $s_1 = 34$ and $s_2 = 66$ with a probability of 0.5796 of getting one of the best two. In the same way, the optimal strategy is characterized by three switching levels s_1, s_2 and s_3 when the goal is to find one of the best three with maximal probability. The optimal values $s^*_{1,n}$, $s^*_{2,n}$ and $s^*_{3,n}$ satisfy $s^*_{1,n} \approx 0.3367n$, $s^*_{2,n} \approx 0.5868n$ and $s^*_{3,n} \approx 0.7746n$ for n sufficiently large, where the probability of getting one of the best three tends to 0.7082 as n gets very large. As an illustration, consider the values $n = 20$ and $n = 100$. For $n = 20$, the optimal values of s_1, s_2 and s_3 are 6, 12 and 15 with a probability of 0.7475 of getting one of the best two. The optimal strategy for $n = 100$ items has the parameter values $s_1 = 33$, $s_2 = 58$ and $s_3 = 77$ with a probability of 0.7160 of getting one of the best three.

More about optimal stopping problems can be found in the fascinating paper "Knowing when to stop," by T.P. Hill, in *American Scientist* **97** (2007): 126–133. Among others this paper discusses optimal stopping problems in which only partial information is available about the number of items to be presented to the observer. Also, the paper discusses a surprising result for the following problem.

How to pick the largest of two hidden numbers?

Two different numbers are written on separate slips of paper. You choose one of the two slips and the number on it is shown to you. Then you must judge whether the observed number is larger than the hidden number on the other slip. The surprising claim is that there is a method guaranteeing that the probability of picking the largest number is more than one-half, no matter what the numbers are. This method is based on randomization. It will be explained for the case that the two different numbers come from the set of the nonnegative integers. The idea is that you arm yourselves with a probability mass function $\{p_k\}$ assigning a positive probability to each nonnegative integer, say the so-called Poisson distribution $p_k = e^{-\lambda}\lambda^k/k!$ for some $\lambda > 0$. You simulate a random observation r from the probability mass function you have chosen. Next use the following rule. If the observed number on the slip of paper is larger than r, guess that the observed number is the largest; otherwise, guess that the hidden number is the largest. How can such a simple-minded rule guarantee a win probability of more than 0.5? This can be argued as follows. Suppose

the hidden numbers are a and b with $a < b$. If r is smaller than a, then the observed number will always be larger than r and sticking to this number gives a win probability of 0.5. If r is larger than or equal to b, then the observed number is smaller than or equal to r and switching to the hidden number gives a win probability of 0.5. But if $a \leq r < b$, then you win with certainty, because you switch to the hidden number b when the observed number is a and you stick to b when the observed number is b. Thus the overall probability of guessing the largest number is more than 0.5. For example, if the two hidden numbers are 58 and 72 and you have armed yourselves with the Poisson probabilities $p_k = e^{-\lambda}\lambda^k/k!$ with $\lambda = 50$, then your overall win probability is $0.5\sum_{k=0}^{57} p_k + 0.5\sum_{k=72}^{\infty} p_k + \sum_{k=58}^{71} p_k = 0.5714$.

3.7 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

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

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 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

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

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.

Bootstrap method

In the statistical analysis of the 1970 draft lottery, we used a powerful, generally applicable form of statistical methodology, namely the *bootstrap method.* This new method, developed in 1977 by the 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 itself 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.

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.

3.8 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 In a class of 30 children, each child writes down a randomly chosen number from $1, 2, \ldots, 250$. What is the probability that two or more children have chosen the same number?

3.9 You bet your friend that, of the next fifteen 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 twenty-five 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 3.8–3.11 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 thirty times a week on average over a period of fifty 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 fifty years.

3.14 In the Massachusetts Numbers Game, a four-digit number is drawn from the numbers 0000, 0001, . . . , 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 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 fifteen and thirty. 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 fifteen-year span, that span consisting of the period from fifteen to thirty years ago. Listeners whose birthday falls 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 thirty-day run of the game?

3.17 Consider the following application of the coupon collector's problem with unequal probabilities. You repeatedly roll two fair dice until you have obtained all possible totals from 2, 3, . . . , 12. Use simulation to find the expected value of the number of rolls you need. Also, simulate the probability that you will roll a 2, 3, 4, 5, 6, 8, 9, 10, 11, and 12 before rolling a seven. What is the house edge of this craps side bet with a payout of \$175 for each dollar bet?

3.18 In a particular game, you begin by tossing a die. If the toss results in k points, then you go on to toss k dice together. If the sum of the points resulting from the toss of the k dice is greater than (less than) twelve, you win (lose) one dollar, and if the sum of those points is equal to twelve, 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.

3.19 In the old English game of Hazard, a player must first determine which of the five numbers from 5, . . . , 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?

3.20 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?

3.21 Use simulation to find out how many people are required to have an at least even chance that there are two or more days with matching birthdays.

3.22 Two players A and B play a game in which they choose each a number from the numbers $10, 11, \ldots, 99$, independently of each other. Player A wins the game if the first digit of the product of the two numbers is 1, 2, or 3; otherwise, player B wins. Suppose player A uses a randomized strategy, where each of the numbers 11, 17, 28, 44, and 67 is chosen with equal probability. Experiment with a few values for the number chosen by player B and verify that the probability of player A winning is 60% or more whatever number player B chooses. *Remark*: this multiplication game is related to Benford's law that will be discussed in Section 5.10.2.

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 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.

3.25 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?

3.26 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 of $\frac{1}{2}$. What is the probability that the drunkard will ever visit the point which is one unit distance south from his starting point? Explain why the answer $q = \frac{1}{2}(\sqrt{5} - 1)$ can be obtained from the equation $q = 0.5 + 0.5q^3$. Use the result for the drunkard's walk to give the probability that the number of heads will ever exceed twice the number of tails if a fair coin is tossed over and over.

3.27 You have \$800 but you desperately need \$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?

3.28 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?

3.29 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.

3.30 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, ..., "thirteen" 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 fifteen 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 twelve 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.



4.1 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

$$X = \text{the total number of successes in } n \text{ independent trials of the Bernoulli experiment.}$$

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 1 and $n-k$ positions can be chosen for a 0 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 two 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 probability has the numerical values 8.85×10^{-3}, 2.46×10^{-4}, and 2.56×10^{-6} for $k = 2$, 3, and 4, respectively.

4.2 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

Table 4.1. *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

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.

Many practical phenomena can be described according to the Poisson distribution. Evidence of this lies in the following important result (see Section 9.6 for a proof):

> **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 is the number of trials and p is the probability of success.**

In mathematical terms, the binomial probability $\binom{n}{k}p^k(1-p)^{n-k}$ tends to the Poisson probability $e^{-\lambda}\lambda^k/k!$ for any $k \geq 0$ if n tends to infinity and p tends to zero while $np = \lambda$. To give you an idea of how quickly the binomial distribution describing the number of successes in a series of independent Bernoulli trials approaches the Poisson distribution, we refer to Table 4.1. In this table 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. The value λ of the product is usually known in practical applications. The Poisson model is uniquely determined by this quantity λ. For large λ (say $\lambda \geq 25$), it would be exceptional for a Poisson distributed random variable to take on a value that differs more than $3\sqrt{\lambda}$ from its expected value λ. This fact is extremely useful for statistical applications of the Poisson model.

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 14 cavalry corpses over the 20 years 1875 to 1894 (see also Problem 12.14 in Chapter 12). 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, etc.

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 4.3 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, twelve 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?).

How exceptional is the occurrence of twelve fires in the past year if the yearly number of fires is a Poisson distributed random variable X with $E(X) = 8$? To answer this question, you calculate the probability of twelve or more fires and not the probability of exactly twelve fires. The desired probability is given by

$$P(X \geq 12) = 0.112.$$

In view of the order of magnitude of this probability, the twelve fires in the past year might be just explained as a chance variation. Such an explanation would be highly implausible if 50% more fires would have occurred in a big city facing 50 fires per year on average. Such an event has a probability of the order of 5.8×10^{-4}.

Example 4.4[†] 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, each of whom submit more than one entry for each weekly draw. For example, if every week 50 million people randomly submit 5 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

[†] 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.

submitting 5 six-number sequences is

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

It is reasonable to model the number of times that a given player will win a jackpot in the next 200 drawings of a Lotto 6/42 by a Poisson distribution with expected value

$$\lambda_0 = 200 \times \frac{5}{\binom{42}{6}} = 1.906 \times 10^{-4}.$$

For the next 200 drawings, this means that

$$\begin{aligned} &P(\text{any given player wins the jackpot two or more times}) \\ &= 1 - e^{-\lambda_0} - e^{-\lambda_0}\lambda_0 = 1.817 \times 10^{-8}. \end{aligned}$$

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

$$\lambda = 50{,}000{,}000 \times (1.817 \times 10^{-8}) = 0.9084.$$

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 $1 - e^{-\lambda} = 0.5968$. 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 $1, \ldots, n$ 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 7.11 of Chapter 7. A numerical comparison of the exact distribution with the Poisson distribution reveals that $n = 15$ is sufficiently large in order for the Poisson probabilities to match the exact probabilities in at least 8 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 3.30. 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 thirteen 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 at least one 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. Simulation shows that the Poisson distribution

with parameter $\lambda = 4$ provides a very accurate approximation to the exact probability distribution of the number of matches.

Several real-life applications of the hat-check problem are given in Problem 4.22. A nice real-life application is the following. In 2008, a famous self-proclaimed psychic medium participated in a one-million dollar paranormal challenge to test for psychic ability. A child was chosen by the medium as one whom he could connect with telephathically. The medium was shown the ten target toys that he would have to pick up during the test and then he was placed in an isolated room. A toy was chosen at random and given to the child. The medium's task was to connect to the child and state which toy he was playing with. This was repeated ten times and at the end of the trials the medium's answers were checked against the toys the child actually played with. If the medium had six or more correct guesses, then he would win one million dollars. The medium correctly guessed one toy choice out of ten, thus failing the test and losing the challenge. By chance alone, a correct guess of one out of ten random choices was to be expected while the probability of winning the challenge was just 6.0×10^{-4}.

A lottery problem

What is the probability that, in 30 lottery drawings of six distinct numbers from the numbers $1, \ldots, 45$, not all 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 $1, \ldots, 45$ induces 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 all 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 (see Problem 7.48). 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 $\lambda(m) = 365 \times p$. This probability gives the probability of having represented all of the 365 birthdays in the group of m people. 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 a group size of 2287 people, as can be calculated by using the inclusion-exclusion formula from Chapter 7 (see Problem 7.49). More advanced methods to attack 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, the arrival of urgent calls to an emergency center, etc. 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

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

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

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

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

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

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

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 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:

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

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 4.5 Out in front of Central Station, multiple-passenger taxicabs wait until either they have 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:

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

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 takes 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

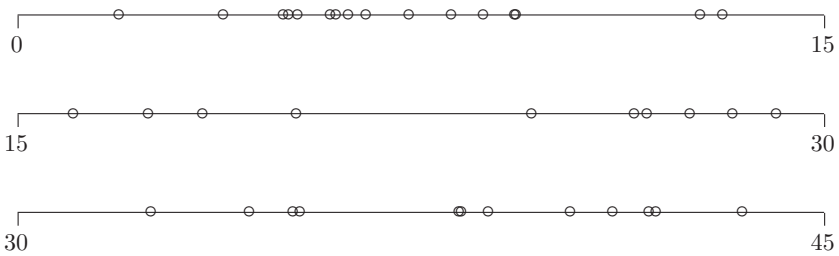


Fig. 4.1. Arrival times of a Poisson process.

(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) 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 twenty 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 4.6 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 $N(t)$ be the number of accidents occurring in the time interval $(0, t)$, where a day is taken as time unit. Denote by the epoch $t = u - 1$ the beginning of day u ($1 \leq u \leq 7$). The probability we are looking for is $P(N(u) - N(u-1) = 1 \text{ for } u = 1, \ldots, 7 \mid N(7) = 7)$ with the convention $N(0) = 0$. By the property **B** of the Poisson process, the random variables $N(1), N(2) - N(1), \ldots, N(7) - N(6)$ are independent and so this probability can be written as

$$\frac{P(N(1) = 1) \times P(N(2) - N(1) = 1) \times \cdots \times P(N(7) - N(6) = 1)}{P(N(7) = 7)}.$$

By property **C**, the random variables $N(1), N(2) - N(1), \ldots, N(7) - N(6)$ are Poisson distributed with expected value $\lambda/7$ and $N(7)$ is Poisson distributed

with expected value λ. Thus, the desired probability is given by

$$\frac{e^{-\lambda/7}(\lambda/7) \times e^{-\lambda/7}(\lambda/7) \times \cdots \times e^{-\lambda/7}(\lambda/7)}{e^{-\lambda}\lambda^7/7!} = \frac{7!}{7^7} = 0.0162.$$

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. How often do lotto numbers come out evenly spaced? A tendency towards bunching is exactly how the Poisson process behaves. 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 inter-arrival 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 inter-arrival time.

1. Generate a random number u between 0 and 1.
2. Take $x = -\frac{1}{\alpha}\ln(u)$ as the inter-arrival 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 4.7 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 10-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 10-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 $1 - \sum_{j=0}^{3} e^{-0.402}(0.402)^j/j! = 0.00079$.

A remarkable use of the splitting property of the Poisson process is made in the following example dealing with a variant of the coupon collector's problem. This example is for the mathematically adept reader.

Example 4.8 In the casino game of craps two dice are repeatedly rolled. What is the probability that a total of 2, 3, 4, 5, 6, 8, 9, 10, 11, and 12 will be rolled before rolling a seven?

Solution. A beautiful trick is used to solve this problem. Imagine that rolls of the two dice occur at epochs generated by a Poisson process with rate $\alpha = 1$. Each roll results in a total of j with probability p_j, where $p_2 = p_{12} = \frac{1}{36}$, $p_3 = p_{11} = \frac{2}{36}$, $p_4 = p_{10} = \frac{3}{36}$, $p_5 = p_9 = \frac{4}{36}$, $p_6 = p_8 = \frac{5}{36}$, and $p_7 = \frac{6}{36}$. For any fixed j, the rolls resulting in a total of j occur according to a Poisson process with rate $\alpha p_j = p_j$. Secondly, and very importantly, these Poisson processes are independent of each other. The analysis now proceeds as follows. Fix t and suppose that a roll with a total of 7 occurs for the first time at time t. For any $j \neq 7$, the probability that at least one roll with a total of j will occur in $(0, t)$ is $1 - e^{-p_j t}$. Thus, by the independence of the separate Poisson processes, the product

$$(1 - e^{-p_2 t}) \cdots (1 - e^{-p_6 t})(1 - e^{-p_8 t}) \cdots (1 - e^{-p_{12} t})$$

represents the conditional probability of a total of 2, 3, 4, 5, 6, 8, 9, 10, 11, and 12 showing up in $(0, t)$ given that the first appearance of a total of 7 is at time t. The probability that a total of 7 will occur for the first time in the very small interval $(t, t + \Delta t)$ is equal to $p_7 e^{-p_7 t}\,\Delta t$ except for a negligible term. Averaging the above product over the density function $p_7 e^{-p_7 t}$, it follows that the desired probability is given by

$$\int_0^\infty (1 - e^{-p_2 t}) \cdots (1 - e^{-p_6 t})(1 - e^{-p_8 t}) \cdots (1 - e^{-p_{12} t})\, p_7 e^{-p_7 t}\, dt.$$

A justification of the unconditioning step is provided by the law of conditional probability that will be discussed in Section 13.2. Finally, numerical integration gives the value 0.005258 for the probability that a total of 2, 3, 4, 5, 6, 8, 9, 10, 11, and 12 will be rolled before rolling a seven. In some casinos the payout for this craps side bet is \$175 for each dollar bet. This corresponds to a house edge of 7.99%. Another craps side bet is the following. You win if any total is rolled twice before a seven. A win pays 2 for 1. It is left to the reader to verify

that your win probability is given by

$$1 - \int_0^\infty (e^{-p_2 t} + p_2 t e^{-p_2 t}) \cdots (e^{-p_6 t} + p_6 t e^{-p_6 t}) \\ \times (e^{-p_8 t} + p_8 t e^{-p_2 t}) \cdots (e^{-p_{12} t} + p_{12} t e^{-p_{12} t})\, p_7 e^{-p_7 t}\, dt = 0.47107.$$

This corresponds to a house edge of 5.79%.

4.3 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 4.9 In the game "Lucky 10," twenty 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 4.2 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 in Table 4.2. 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 twenty numbers drawn from the numbers $1, \ldots, 80$ are identified as $R = 20$ red balls in an urn and that the remaining sixty, 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 4.2, you will get an expected payoff of:

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

[†] The modern lottery with prize money attached has its origins in what is now the Netherlands and Flanders: 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 4.2. *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}

Table 4.3. *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%

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 4.3 are calculated in the same way.

In Table 4.3, the expected payoff per dollar staked on 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.

Example 4.10[†] 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

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

Table 4.4. *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

payoffs and win probabilities are given in Table 4.4. 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 probability is. It can best be attempted as follows: if you enter twelve 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:

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

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, 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 $1, \ldots, 45$ (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 4.5. No statistical tests are necessary in

Table 4.5. *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

order to recognize that these people did not choose their numbers randomly. Table 4.5 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 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

[†] 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.

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 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, remind 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

[†] 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.

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 twelve dice, and at least three sixes in one throw of eighteen 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 configurations. 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 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

† 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.

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 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?

4.9 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?

4.10 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?

4.11 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.

4.12 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?

4.13 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?

4.14 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.

4.15 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.

4.16 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.

4.17 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.

4.18 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.

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

4.20 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.

4.21 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.

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

(a) In a particular branch of a company, the fifteen 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?

4.23 A total of 126 goals were scored by the 32 soccer teams in the 48 games played in the group matches of the 1998 World Cup soccer. It happened that 26 times zero goals were scored by a team, 34 times one goal, 24 times two goals, eight times three goals, one time four goals, two times five goals, and one time six goals. Can you make plausible that the number of goals scored per team per game can be quite well described by a Poisson distribution with an expected value of 1.3125?

4.24 In the first five months of the year 2000, trams 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?

4.25 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 fifteen or more of these 75 cars are parked illegally?

4.26 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).

4.27 Consider the following side bet in craps. You win if the shooter rolls a 2, 3, 4, 5, and 6 before rolling a seven. A win pays 35 for 1. Use the trick from Example 4.8 to find the win probability and the house edge.

4.28 In Lottoland, there is a weekly lottery in which six regular numbers plus one bonus number are drawn at random from the numbers $1, \ldots, 45$ without replacement. 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 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. The central limit theorem will also be applied to go more deeply in the analysis of betting systems and to introduce the Brownian

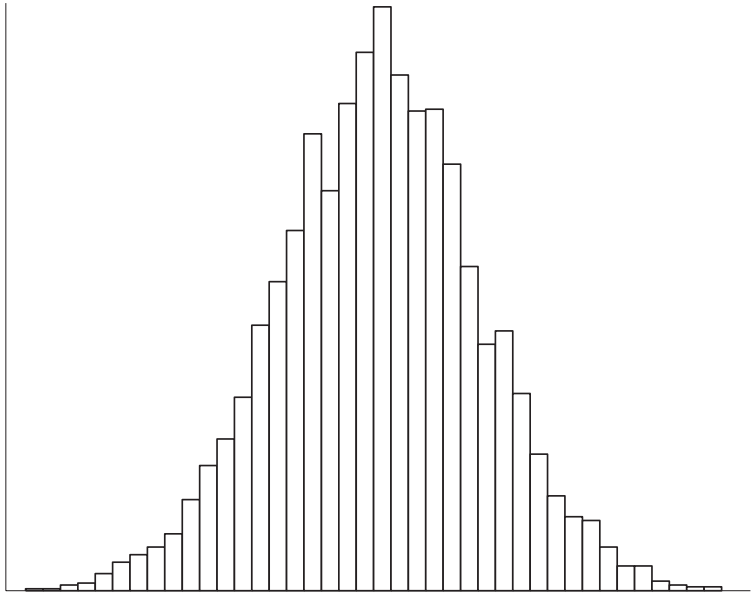


Fig. 5.1. Histogram of heights.

motion process. This random process is widely used in physics and in the modeling of financial markets. To conclude this chapter, we discuss several statistical topics, including the difference between Bayesian statistics and classical statistics.

5.1 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

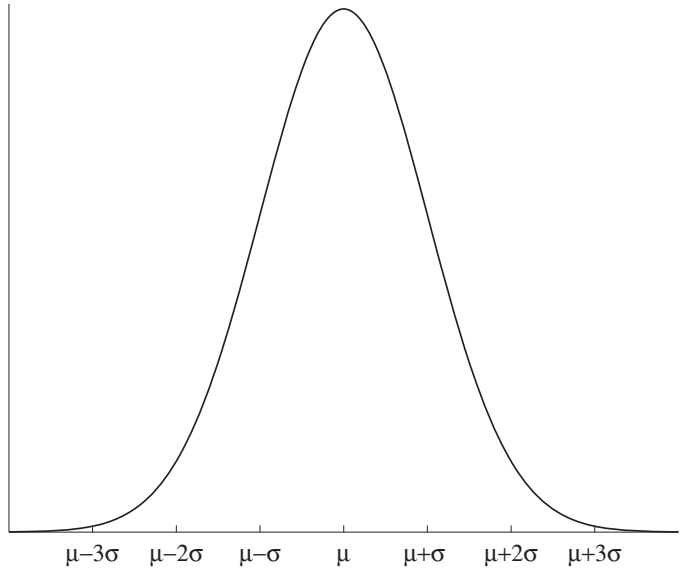


Fig. 5.2. The normal curve.

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

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

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 $\sqrt{2}, \pi = 3.14159\ldots$ and $e = 2.71828\ldots$ 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, etc., 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

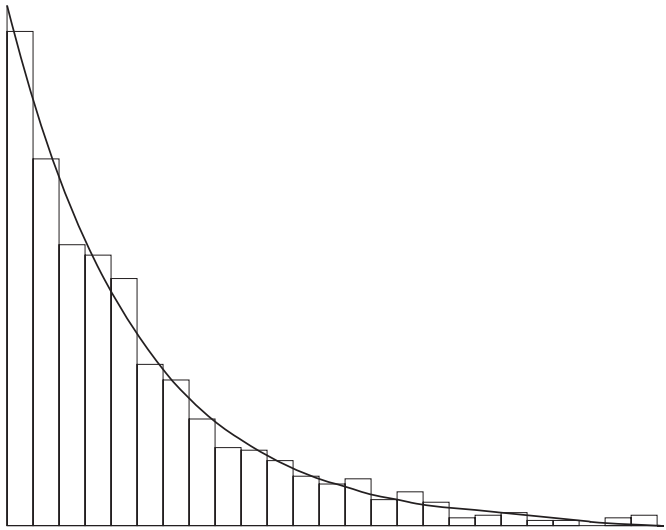


Fig. 5.3. Histogram of decay times.

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 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 each of the intervals (a, b), $[a, b)$, $(a, b]$, and $[a, b]$ gets assigned the same probability. 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 provided that $f(x)$ is continuous at the point v. 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 is a measure for 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 $\Phi(z)$ 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

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

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, $P(X \leq \mu - z\sigma) = P(X \geq \mu + z\sigma)$. 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 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

[†] This example is borrowed from Sam Savage in his article, "The flaw of averages," October 8, 2000 in the *San Jose Mercury News*.

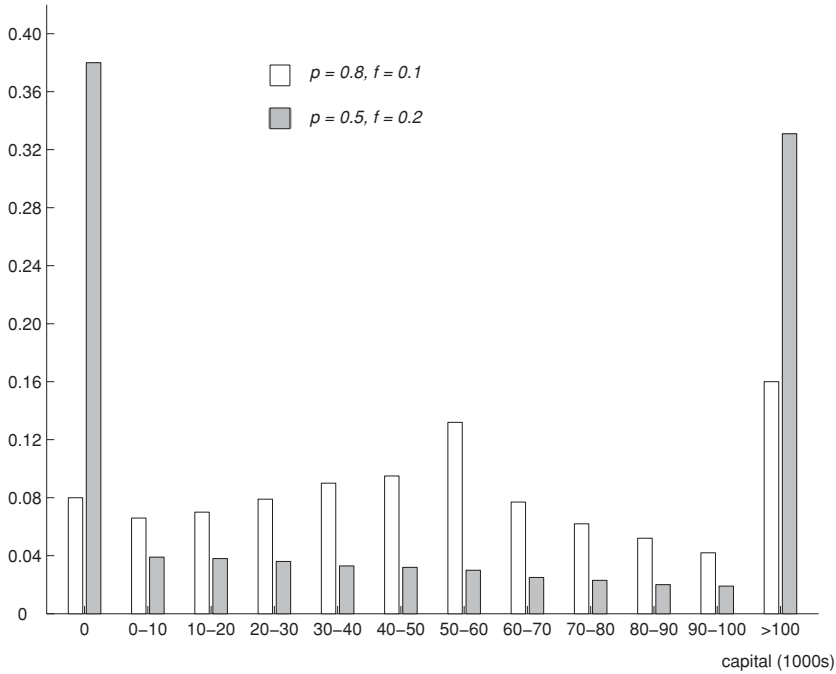


Fig. 5.4. Distribution of invested capital after 15 years.

$(1+r)^{20}A - \sum_{k=0}^{19}(1+r)^k x = 0$ 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 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,

$$\sigma^2(X) = E\left[(X-\mu)^2\right].$$

Another common notation for the variance of X is $\mathrm{var}(X)$.[†] Why not use $E(|X-\mu|)$ as the measuring gauge for the spread? The answer is simply that it

[†] 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 $E[(X-\mu)^2] = \sum_{i=1}^{6}(i-3.5)^2 P(X=i) = 2.917$.

is much easier to work with $E\left[(X-\mu)^2\right]$ than with $E(|X-\mu|)$. The variance $\sigma^2(X) = E\left[(X-\mu)^2\right]$ 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

$$P\big(X > \mu + a\big) \leq \frac{\sigma^2(X)}{\sigma^2(X) + a^2} \quad \text{and} \quad P\big(X < \mu - a\big) \leq \frac{\sigma^2(X)}{\sigma^2(X) + a^2}$$

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

$$\sigma(X) = \sqrt{E\left[(X-\mu)^2\right]}.$$

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

$$\operatorname{cov}(X, Y) = E\Big[\,(X - E(X))\,(Y - E(Y))\,\Big].$$

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.

Property 2. *For every two random variables X and Y,*

$$\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y) + 2\,\text{cov}(X, Y).$$

5.2.2 Independent random variables

Random variables X and Y are said to be *uncorrelated* if $\text{cov}(X, Y) = 0$. 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 $\text{cov}(X, Y) = 0$ if X and Y are independent (see Chapter 11). Thus, Property 2 implies that

$$\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y) \qquad \text{for independent } X \text{ and } Y.$$

[†] 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.

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 that after waiting for more than twenty, 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 busses, 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

† This section is highly specialized and may be skipped over.

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

$$\frac{1}{2}(1 + c_T^2)E(T)$$

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 arrival rate λ (the Poisson process was discussed in Section 4.2.4 and is a process in which arrivals occur completely randomly in time). 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 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 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 $n \geq 1$,*

$$\sigma\,(X_1 + X_2 + \cdots + X_n) = \sigma\sqrt{n}.$$

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.

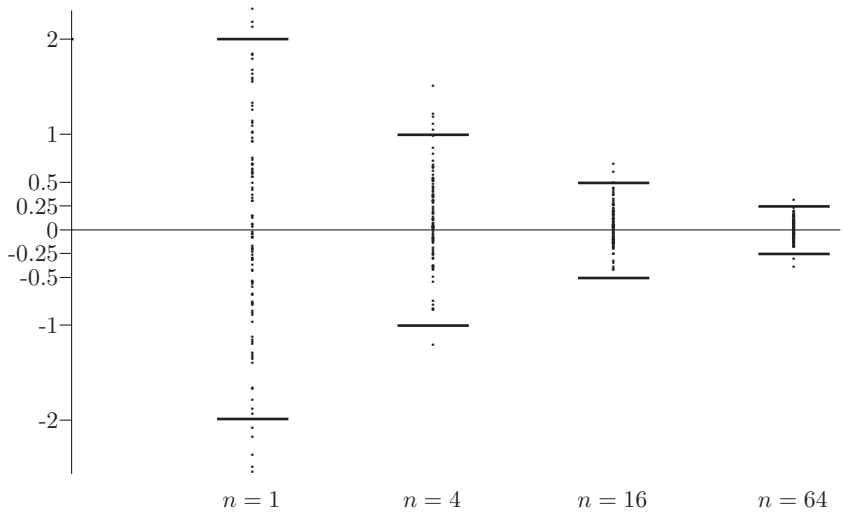


Fig. 5.5. Simulation outcomes for the sample mean.

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 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 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 $a < b$,*

$$\lim_{n\to\infty} P\left(a \le \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le b\right) = \Phi(b) - \Phi(a),$$

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 reformulate 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 the 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 that, 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, etc., 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 5,227 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 $c > 0$,

$$P\left(\left|(X_1 + X_2 + \cdots + X_n) - n\mu\right| > c\sigma\sqrt{n}\right) \approx 2\{1 - \Phi(c)\}$$

[†] 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.

when n is sufficiently large. In particular,

$$P\left(\left|(X_1+\cdots+X_n)-n\mu\right| > c\sigma\sqrt{n}\right) \approx \begin{cases} 0.3173 & \text{for } c=1 \\ 0.0455 & \text{for } c=2 \\ 2.70\times10^{-3} & \text{for } c=3 \\ 6.33\times10^{-5} & \text{for } c=4 \\ 5.73\times10^{-7} & \text{for } c=5. \end{cases}$$

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 5,227 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 5,227 or more heads appearing in 10,000 tosses of a fair coin (and not the probability of exactly 5,227 heads). The number of times the coin lands heads can be written as the sum $X_1+X_2+\cdots+X_{10,000}$, where

$$X_i = \begin{cases} 1 & \text{if the } i\text{th toss turns heads} \\ 0 & \text{otherwise.} \end{cases}$$

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

$$\sigma = \sqrt{\left(0-\frac{1}{2}\right)^2\times\frac{1}{2}+\left(1-\frac{1}{2}\right)^2\times\frac{1}{2}} = \frac{1}{2}.$$

Hence the expected value and the standard deviation of the number of heads in 10,000 tosses are given by 5,000 and $\frac{1}{2}\sqrt{10{,}000}=50$. Tossing 5,227 heads lies

$$\frac{5{,}227-5{,}000}{50} = 4.54$$

standard deviations above the expected value of 5,000 heads. The chance of this or more extreme outcome happening is approximately 1 in 3.5 million.[†] The claim of your friend is fakery! Note that the judgment is based on the probability of getting 5,227 or more heads in 10,000 tosses and not on the probability of getting exactly 5,227 heads (any individual outcome has a very small probability).

[†] Use the fact that $P(X>\mu+c\sigma)=P(\frac{X-\mu}{\sigma}>c)=1-\Phi(c)$ for an $N(\mu,\sigma^2)$ distributed random variable X.

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 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 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

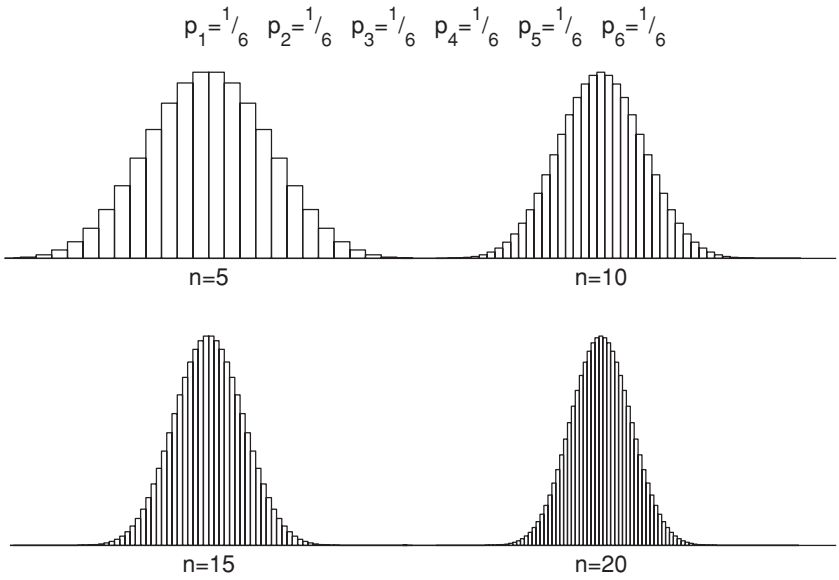


Fig. 5.6. Probability histogram for the unbiased die.

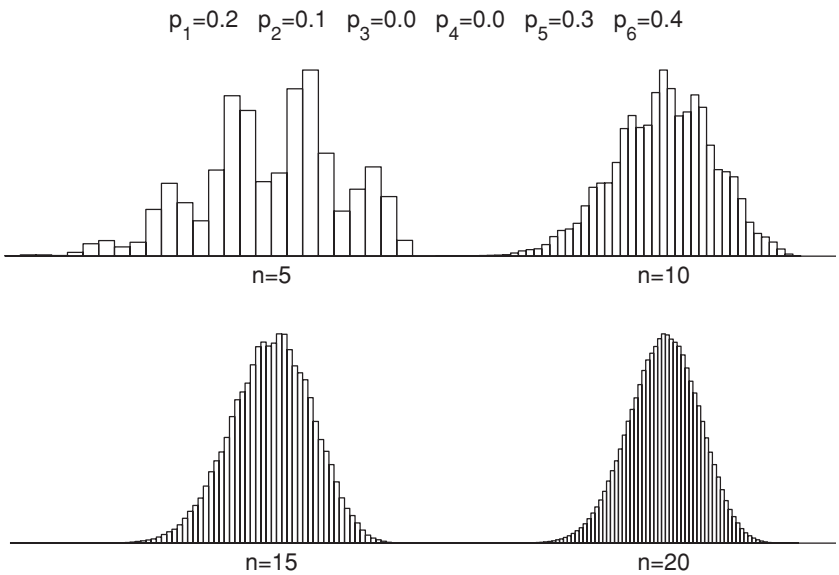


Fig. 5.7. Probability histogram for a biased die.

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 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 applications and go more deeply into the concept of z-value. This concept is a quick and useful tool to judge whether an observation from an approximately normal distribution is exceptional.

5.6.1 The z-value

The key to finding the solution to the illustrative example in Section 5.4.1 was 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 5.1 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 $p = 0.5$. This binomial model can be approximated by the model of the normal distribution. A binomial random variable with parameters $n = 199$ and $p = 0.5$ has expected value $199 \times 0.5 = 99.5$ and standard deviation $0.5\sqrt{199} = 7.053$. Thus, for the z-value, we find

$$z = \frac{57 - 99.5}{7.053} = -6.03.$$

In words, 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 $\Phi(-6.03)$ and 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.

5.6.2 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 Section 4.2). The binomial distribution with parameters n and p tends to a Poisson distribution with parameter λ if n tends to infinity and p tends to zero while $np = \lambda$. A binomially distributed random variable X with parameters n and p can be written as the sum of n independent Bernoulli variables $I_1, \ldots, I_n$ with success probability p, where $I_k = 1$ if the kth trial is a success and $I_k = 0$ otherwise. The expected value and the standard deviation of the binomial random variable X are given by $E(X) = np$ and $\sigma(X) = \sqrt{np(1-p)}$. Hence, by the central limit theorem, the distribution of X can be approximated by the normal distribution with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ if n is large enough (a rule of thumb is $np(1-p) \geq 25$).

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 an approximately 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?

The Poisson model is a practically useful model to make quick statistical assessments. As you have seen, the Poisson model may be an appropriate model for the frequently occurring situation of many experiments each having a small success probability, where one knows only the expected value of the number of successes. What's very useful is that the standard deviation of the Poisson distribution is the square root of the expected value of the distribution and nearly all of

the mass of the distribution is within three standard deviations from the expected value.

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 $\left[\overline{X}(n) - \mu\right] / \left(\sigma/\sqrt{n}\right)$ is an approximately standard normal random variable, it follows that

$$P\left(-z_{1-\frac{1}{2}\alpha} \leq \frac{\overline{X}(n) - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{1}{2}\alpha}\right) \approx 1 - \alpha$$

or, stated differently,

$$P\left(\overline{X}(n) - z_{1-\frac{1}{2}\alpha}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}(n) + z_{1-\frac{1}{2}\alpha}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$

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

$$S^2(n) = \frac{1}{n}\sum_{k=1}^{n}\left[X_k - \overline{X}(n)\right]^2.$$

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. 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

[†] 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).

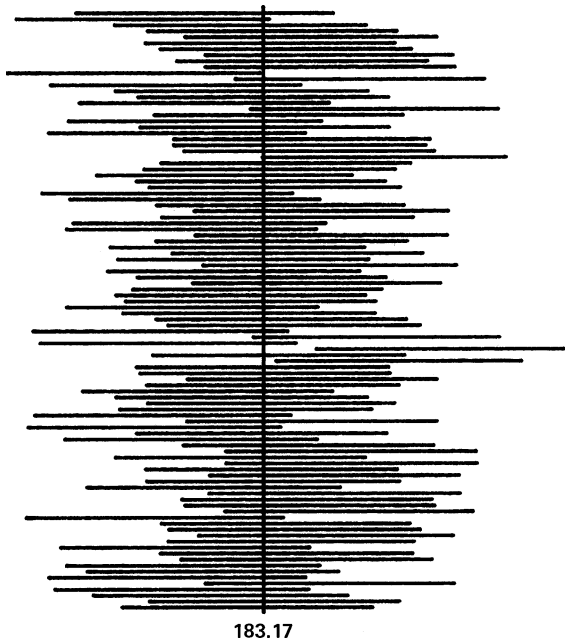


Fig. 5.8. One hundred 95% confidence intervals.

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%. In other words, 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 one hundred 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

[†] This is the frequentist approach for constructing a confidence interval. It requires that many independent replications of the experiment can be done. The frequentist confidence interval is a natural concept in the statistical analysis of the output of simulation trials of a stochastic system. In physics and medicine among others, random experiments are often not easy to repeat or cannot be repeated under the same conditions. In such situations one might use another approach resulting in a so-called Bayesian confidence interval for the unknown quantity, see Section 13.4.

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 one hundred approximate 95% confidence intervals for μ by simulating the sales one hundred 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).

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

$$\overline{X}(n) \pm z_{1-\frac{1}{2}\alpha} \frac{\sqrt{\overline{X}(n)\left[1-\overline{X}(n)\right]}}{\sqrt{n}}.$$

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 0.809 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 $1.96\sqrt{0.809\,(1-0.809)}/\sqrt{2{,}500} = 0.015$. 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 5.2 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

$$\overline{X}(n) \pm 1.96 \frac{\sqrt{\overline{X}(n)\left[1 - \overline{X}(n)\right]}}{\sqrt{n}} = 0.4886 \pm 0.0013.$$

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 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

$$E(W_n) = n\,(1 - 2p) \text{ and } \sigma(W_n) = 2\,[p\,(1 - p)]^{\frac{1}{2}}\,\sqrt{n}.$$

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:

$$\begin{array}{ll}
\text{loss probability} = 0.3553 & \text{if } n = 100 \\
\text{loss probability} = 0.1876 & \text{if } n = 1{,}000 \\
\text{loss probability} = 0.0033 & \text{if } n = 10{,}000 \\
\text{loss probability} = 6.1\times 10^{-18} & \text{if } n = 100{,}000 \\
\text{loss probability} = 3.0\times 10^{-161} & \text{if } n = 1{,}000{,}000.
\end{array}$$

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 $(36/k) - 1$ 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 10.13 in Section 10.5, 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 10.13, 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 gets nothing paid. 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

$$\begin{aligned} \mu &= p \ln(1 - \alpha + \alpha f) + (1 - p) \ln(1 - \alpha) \\ \sigma^2 &= p[\ln(1 - \alpha + \alpha f) - \mu]^2 + (1 - p)[\ln(1 - \alpha) - \mu]^2. \end{aligned}$$

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

$$e^{n\mu+\ln(V_0)+n\sigma^2/2} \approx a V_0,$$

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

$$\begin{aligned} P(V_n \geq a V_0) &= P\big(\ln(V_n) \geq \ln(a) + \ln(V_0)\big) \\ &= P\left(\frac{\ln(V_n) - n\mu - \ln(V_0)}{\sigma\sqrt{n}} \geq \frac{\ln(a) - n\mu}{\sigma\sqrt{n}}\right). \end{aligned}$$

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 of 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 of no 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

$$\mu_1 = 0.5\ln(1 + 0.8\alpha) + 0.5\ln(1 - 0.6\alpha)$$
$$\sigma_1 = [0.5(\ln(1 + 0.8\alpha) - \mu_1)^2 + 0.5(\ln(1 - 0.6\alpha) - \mu_1)^2]^{1/2}.$$

Noting that $P(V_n > x) = P\,(\ln(V_n/V_0) > \ln(x/V_0))$, we find

$$P(V_n > x) = P\left(\frac{\ln(V_n/V_0) - n\mu_1}{\sigma_1\sqrt{n}} > \frac{\ln(x/V_0) - n\mu_1}{\sigma_1\sqrt{n}}\right)$$
$$\approx 1 - \Phi\left(\frac{\ln(x/V_0) - n\mu_1}{\sigma_1\sqrt{n}}\right),$$

where $V_0 = 10{,}000$. 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 no more than \$1.95 after 52 weeks is approximately equal to 0.500. This probability is practically equal to 0 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.

An approximation formula

In the above example one might also pose the following question. What is the probability that during the next 52 weeks your bankroll exceeds b times your initial bankroll for a given value of $b > 1$? Let us address this question for a slight generalization of the Kelly model formulated in Problem 2.10. You are offered a sequence of independent betting opportunities. Each bet pays out 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 $f_1 > 1, 0 \leq f_2 < 1$. Also, there is an interest rate r, that is, each dollar you do not stake at a betting opportunity is worth $1 + r$ dollars at the next betting opportunity. It is assumed that $pf_1 + (1 - p)f_2 > 1 + r$. Suppose you use a betting strategy in which you bet each time the same fraction α of your current bankroll. Let the random variable V_n denote your bankroll after the nth bet, where V_0 is your initial bankroll. What is the probability

$$P(V_n > bV_0 \text{ for some } k,\ 1 \leq k \leq N)$$

for given values of b and N? This question can be converted into a question about a hitting probability in a random walk. Equivalently, we can ask what is $P(\ln(V_n/V_0) > \ln(b) \text{ for some } k,\ 1 \leq k \leq N)$? By the same arguments as in Section 2.7.3, we find

$$\ln\left(\frac{V_n}{V_0}\right) = \sum_{i=1}^{n} \ln[(1-\alpha)(1+r) + \alpha R_i],$$

where $R_1, R_2, \ldots$ are independent random variables with $P(R_i = f_1) = p$ and $P(R_i = f_2) = 1 - p$. The independence of the R_i implies that the random variables $\ln[(1-\alpha)(1+r) + \alpha R_i]$ are also independent and so the partial sums $\sum_{i=1}^{n} \ln[(1-\alpha)(1+r) + \alpha R_i]$ form a random walk process. The expected value and standard deviation of the $\ln[(1-\alpha)(1+r) + \alpha R_i]$ are

$$\mu_\alpha = ph_1 + (1-p)h_2,$$
$$\sigma_\alpha = \sqrt{ph_1^2 + (1-p)h_2^2} - \mu_\alpha^2,$$

where $h_j = \ln[(1-\alpha)(1+r) + \alpha f_j]$ for $j = 1, 2$. It is a matter of simple algebra to verify that the asymptotic growth factor $\mu(\alpha)$ is maximal for

$$\alpha^* = \min\left(\frac{(1+r)(pf_1 + (1-p)f_2) - (1+r)^2}{(1+r)(f_1+f_2) - f_1 f_2 - (1+r)^2},\ 1\right).$$

Using the fact that the process $\{\ln(V_n/V_0),\ n \geq 1\}$ is a random walk, the following approximation can be obtained (see also Section 5.9.2):

$$\begin{aligned} &P(V_n > bV_0 \text{ for some } k,\ 1 \leq k \leq N) \\ &\approx \Phi\left(\frac{\mu_\alpha}{\sigma_\alpha}\sqrt{N} - \frac{\ln(b)}{\sigma_\alpha\sqrt{N}}\right) + e^{2\ln(b)\mu_\alpha/\sigma_\alpha^2}\,\Phi\left(-\frac{\mu_\alpha}{\sigma_\alpha}\sqrt{N} - \frac{\ln(b)}{\sigma_\alpha\sqrt{N}}\right), \end{aligned}$$

where $\Phi(x)$ is the standard normal distribution function. This useful approximation is taken from Edward O. Thorp, "The Kelly criterion in blackjack, sports betting, and the stock market," revised version 1998, www.bjmath.com. As illustration, consider the numerical example with $f_1 = 1.8$, $f_2 = 0.4$, $p = 0.5$, $r = 0$, and $N = 52$. For $b = 2$, 3, 4 the approximate probabilities are 0.675, 0.466, 0.333 and 0.246 when $\alpha = \alpha^*$ and 0.428, 0.143, 0.048 and 0.017 when $\alpha = 0.5\alpha^*$. The simulated values based on 100,000 runs are 0.639 ($\pm$0.003), 0.445 ($\pm$0.003), 0.317 ($\pm$0.003) and 0.231 ($\pm$0.003) when $\alpha = \alpha^*$ and 0.393 ($\pm$0.003), 0.134 ($\pm$0.002), 0.045 ($\pm$0.001) and 0.014 ($\pm$0.001) when $\alpha = 0.5\alpha^*$, where the numbers between brackets indicate the 95% confidence intervals.

5.9 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 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,

† This section contains more advanced material.

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.

5.9.1 Construction of a Brownian motion process

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

$$E(X_i) = (2p - 1)\delta \quad \text{and} \quad \text{Var}(X_i) = 4p(1 - p)\delta^2$$

(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. A Brownian motion process with $\mu = 0$ and $\sigma^2 = 1$ is called the *standard Brownian motion process*. By scaling the normal distribution, it follows that the process $\{\frac{X(t)-\mu t}{\sigma}\}$ is standard Brownian motion if $\{X(t)\}$ is a Brownian motion process with drift parameter μ and variance parameter σ^2.

The Brownian motion process is often referred to as the Wiener process after the American mathematician Norbert Wiener who laid the mathematical

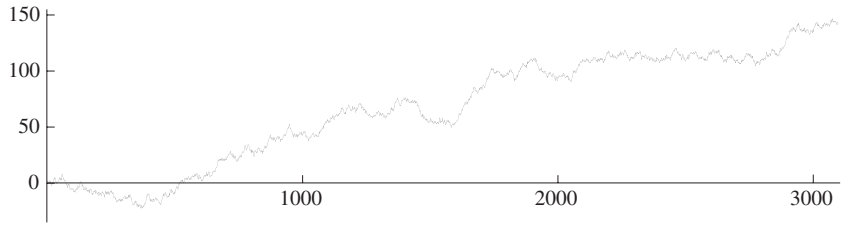


Fig. 5.9. A realization of Brownian motion.

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 the 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 like mountains and clouds as the end result of random walks.

5.9.2 Hitting probabilities in Brownian motion

Hitting probabilities play an important role in diverse problems in gambling and the theory of financial derivatives. For a Brownian motion process with drift parameter μ and variance parameter σ^2, the following famous formula holds:

$$P(\text{process hits } c \text{ before } -d) = \frac{1 - e^{-2d\mu/\sigma^2}}{1 - e^{-2(d+c)\mu/\sigma^2}}$$

for any $c, d > 0$, where the probability should be read as $\frac{d}{d+c}$ if $\mu = 0$. In Problem 5.32 the reader is asked to derive this formula from the gambler's ruin formula from Section 3.5. By rewriting this formula and letting $c \to \infty$, it is easily verified that

$$P(\text{process evers hits } -d) = e^{-2d\mu/\sigma^2} \text{ for any } d > 0$$

when the Brownian motion process has a *positive* drift parameter μ. Similarly, for a Brownian motion process with a *negative* drift parameter μ,

$$P(\text{process evers hits } c) = e^{2c\mu/\sigma^2} \quad \text{for any } c > 0.$$

Many analytical results are available for the Brownian motion process. Approximating a random walk by a Brownian motion process and using these analytical results, useful approximations can be obtained for random walks. In

particular, hitting probabilities in random walks with jumps of non-unit length can be approximated by using the gambler's ruin formula for the Brownian motion process. The Brownian motion approximation is in terms of the expected value and the standard deviation of the jumps of the random walk. To explain how the approximation works, consider the random walk that was analyzed in Section 3.5. This random walk goes in each step either one unit to the right or one unit to the left with probabilities $p = \frac{244}{495}$ and $q = \frac{251}{495}$ respectively. It follows from the gambler's ruin formula in Section 3.5 with $a = 200$ and $b = 35$ that the probability that this random walk reaches the point 235 before it falls down to the point 0 is 0.37077 when the random walk starts at the point 200. The random walk can be approximated by a Brownian motion process with drift parameter μ and variance parameter σ^2, where

$$\mu = 1 \times \frac{244}{495} - 1 \times \frac{251}{495} = -0.014141,$$
$$\sigma^2 = 1^2 \times \frac{244}{495} + (-1)^2 \times \frac{251}{495} - \mu^2 = 0.99980.$$

Plugging $c = 35$ and $d = 200$ into the gambler's ruin formula for the Brownian motion process, we find the approximate value 0.37072 for the hitting probability in the random walk. This approximation is very close to the exact value 0.37077.

Another useful analytical result for Brownian motion is the following result that will be stated without proof. For a Brownian motion process $\{X(t)\}$ with drift parameter μ and variance parameter σ^2,

$$P(X(t) \geq a \text{ for some } t,\ 0 \leq t \leq T)$$
$$= \Phi\left(\frac{\mu}{\sigma}\sqrt{T} - \frac{a}{\sigma\sqrt{T}}\right) + e^{2a\mu/\sigma^2}\,\Phi\left(-\frac{\mu}{\sigma}\sqrt{T} - \frac{a}{\sigma\sqrt{T}}\right)$$

for any $a > 0$, where $\Phi(x)$ is the standard normal distribution function.[†] Similarly, by writing $X(t) \leq -a$ as $-X(t) \geq a$, we have for any $a > 0$,

$$P(X(t) \leq -a \text{ for some } t,\ 0 \leq t \leq T)$$
$$= \Phi\left(\frac{-\mu}{\sigma}\sqrt{T} - \frac{a}{\sigma\sqrt{T}}\right) + e^{-2a\mu/\sigma^2}\,\Phi\left(\frac{\mu}{\sigma}\sqrt{T} - \frac{a}{\sigma\sqrt{T}}\right).$$

Let us illustrate the last formula with the casino game of blackjack. Imagine that you are going to play a blackjack variant using a basic strategy for which

[†] This result can also be used to give the probability distribution of the first time $T(a,\ b)$ at which standard Brownian motion hits the line $a + bt$, where $a > 0$. It holds that $P(T(a,\ b) \leq T) = P(X(t) \geq a \text{ for some } t,\ 0 \leq t \leq T)$ for the Brownian motion process $\{X(t)\}$ with drift parameter $\mu = -b$ and variance parameter $\sigma^2 = 1$. In particular, the probability that standard Brownian notion will ever hit the line $a + bt$ is $\min(1,\ e^{-2ab})$.

the net payoff of a one-dollar bet has an expected value of $-\$0.0039$ and a standard deviation of \$1.148. Your initial bankroll is \$400 and you bet \$10 each time. What is the probability that you will go broke within 800 bets? Using the fact that the net payoff of a ten-dollar bet has an expected value of $\mu = -0.039$ and a standard deviation of $\sigma = 11.48$, it follows that your probability of going broke within 800 bets is approximately equal to

$$\Phi\left(\frac{0.039}{11.48}\sqrt{800} - \frac{400}{11.48\sqrt{800}}\right)$$
$$+ e^{2\times 400\times 0.039/(11.48)^2}\,\Phi\left(-\frac{0.039}{11.48}\sqrt{800} - \frac{400}{11.48\sqrt{800}}\right) = 0.2494.$$

You cannot do much about the house percentage of 0.39%, but you can control your probability of ruin through the bet size. If you bet five dollars each time, then your probability of going broke within 800 bets is approximately 0.0174 when your initial bankroll is \$400 (verify!).

5.9.3 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

$$\delta = \sigma\sqrt{\Delta} \quad \text{and} \quad p = \frac{1}{2}\left\{1 + \frac{\gamma}{\sigma}\sqrt{\Delta}\right\}$$

for given values of γ 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 $\gamma - \frac{1}{2}\sigma^2$ 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 $t > 0$,

$$S_t^\Delta = (1 + X_1) \times \cdots \times (1 + X_{\lfloor t/\Delta \rfloor})S_0.$$

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 $p = \frac{1}{2}(1 + \frac{\gamma}{\sigma}\sqrt{\Delta})$. 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 become negligible as $\Delta \to 0$. Thus, using the expansion of $\ln(1 + X_i)$, we find that $\ln\left(S_t^\Delta/S_0\right)$ 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 c has an $N(c + \nu, \tau^2)$ 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 $\gamma - \frac{1}{2}\sigma^2$ and variance parameter σ^2.†

The following useful result can be given for the stock price process $\{S_t\}$. For any a, b with $0 < b < 1 < a$,

$$P(\text{the stock price increases to } aS_0 \text{ before falling down to } bS_0) = \frac{1 - b^{2\mu/\sigma^2}}{1 - (b/a)^{2\mu/\sigma^2}},$$

where $\mu = \gamma - \frac{1}{2}\sigma^2$. Using the fact that $\{\ln(S_t/S_0)\}$ is a Brownian motion process with drift parameter $\mu = \gamma - \frac{1}{2}\sigma^2$ and variance parameter σ^2, the

† The process $\{S_t\}$ is a so-called geometric Brownian motion: a random process $\{Y(t)\}$ is said to be a geometric Brownian motion if $Y(t) = y_0 e^{X(t)}$ with $\{X(t)\}$ is a Brownian motion 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.

result follows directly from the gambler's ruin formula for Brownian motion by taking $c = \ln(a)$ and $d = -\ln(b)$ in this formula. To illustrate the above formula for the hitting probability, consider the wealth process in the Kelly model from the last paragraph of Section 5.8.4. In this paragraph it was shown that the process $\{\ln(V_n/V_0),\ n \geq 1\}$ is a random walk, where V_n is the size of the bankroll after the nth bet when each bet is a fixed fraction α of your current bankroll. Approximating this random walk by a Brownian motion process with drift parameter μ_α and variance parameter σ_α^2, it follows from the above formula that

$$P(\text{the process } \{V_n\} \text{ reaches } aV_0 \text{ without falling down first to } bV_0) \approx \frac{1 - b^{2\mu_\alpha/\sigma_\alpha^2}}{1 - (b/a)^{2\mu_\alpha/\sigma_\alpha^2}}$$

for any constants a, b with $0 < b < 1 < a$, where V_0 is the initial bankroll. Let us take the same numerical data $f_1 = 1.8$, $f_2 = 0.4$, $r = 0$ and $p = 0.5$ as in the last paragraph of Section 5.8.4. Then the Kelly strategy prescribes the betting fraction $\alpha^* = \frac{5}{24}$. Taking $a = 2$ and $b = 0.5$, then the probability that your bankroll reaches aV_0 without falling down first to bV_0 has the approximate values 0.6661, 0.8866, and 0.9796 for the Kelly strategy with $\alpha = \alpha^*$ and the fractional Kelly strategies with $\alpha = 0.5\alpha^*$ and $\alpha = 0.3\alpha^*$. A simulation study with one million runs gives the simulated values 0.6781 ($\pm$0.0009), 0.8960 ($\pm$0.0006), and 0.9822 ($\pm$0.0003) showing that the approximation is excellent. It is interesting to note that the simplified approximation in Section 2.7.2 gives nearly the same answers as the Brownian motion approximation which involves the expected value μ_α and the standard deviation σ_α. The simplified approximation does not require μ_α and σ_α but uses the ratio $c = \alpha/\alpha^*$ and even seems to improve slightly the Brownian motion approximation.

5.9.4 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 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 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 $E(S_\Delta/S_0) = E\left(e^W\right)$. Since W is $N(\eta\Delta, \sigma^2\Delta)$ distributed, a basic result for the normal distribution tells us that $E\left(e^W\right) = e^{\eta\Delta + \frac{1}{2}\sigma^2\Delta}$ (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 $N\left((r - \frac{1}{2}\sigma^2)T, \sigma^2 T\right)$ distribution, it is matter of integral calculus 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

$$d_1 = \frac{\ln(S_0/K) + (r + \frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}} \quad \text{and} \quad d_2 = d_1 - \sigma\sqrt{T},$$

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

$$d_1 = 0.0498673470,\ d_2 = 0.0211410061,\ d_3 = 0.0032776263,$$
$$d_4 = 0.0000380036,\ d_5 = 0.0000488906,\ d_6 = 0.0000053830.$$

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 $\Phi(x) = 1 - \Phi(-x)$.

5.10 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. The first list is a fake. To put it differently, the first list was made less random to make it feel more random. Most people tend to avoid noting long sequences of consecutive heads or tails. Truly random sequences do not share this human tendency!

5.10.1 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 twenty attempts? We learned

in 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

$$u_k(i) = \text{the probability of getting a run of } r \text{ heads during } n \text{ tosses when the current state of the tossing process is } (i, k).$$

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 twenty 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 the Sections 14.1 and 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

Table 5.1. *Probability of a run of length r in n tosses.*

n/r	3	4	5	6	7	8	9	10
10	0.826	0.465	0.217	0.094	0.039	0.016	0.006	0.002
25	0.993	0.848	0.550	0.300	0.151	0.073	0.035	0.017
50	1.000	0.981	0.821	0.544	0.309	0.162	0.082	0.041
75	1.000	0.998	0.929	0.703	0.438	0.242	0.126	0.064
100	1.000	1.000	0.972	0.807	0.542	0.315	0.169	0.087
150	1.000	1.000	0.996	0.918	0.697	0.440	0.247	0.131
200	1.000	1.000	0.999	0.965	0.799	0.542	0.318	0.172

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

$$v_k(i) = \text{the probability of getting a run of } r \text{ heads or } r \text{ tails during } n \text{ tosses when the current state of the tossing process is } (i, k).$$

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 $1 \le i \le r-1$ and $v_j(r) = 1$ for $0 \le j \le n-1$. 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. In Table 5.1 we give the probability of getting a run of r heads or r tails in n tosses of a fair coin for several values of r and n. The numbers in the table may help correct common mistaken notions about the lengths of runs in coin tossing. More about success runs in the Sections 14.1 and 15.3.

5.10.2 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,

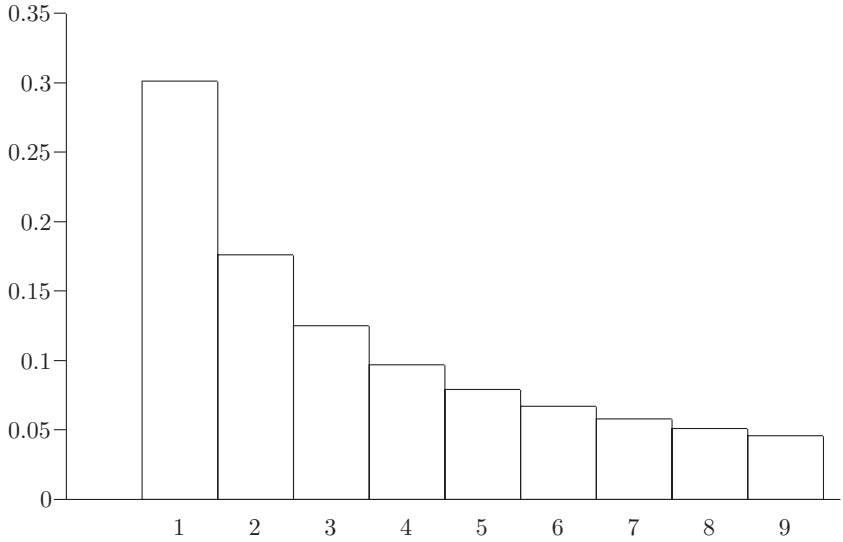


Fig. 5.10. Probability distribution of the first significant digit.

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

$$P(\text{first significant digit} = d) = \log_{10}(1 + \frac{1}{d}) \qquad \text{for } d = 1, 2, \ldots, 9.$$

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 a nearly perfect approximation if all these different data sets were combined. Many familiar mathematical sequences, including the Fibonacci numbers, the powers of 2, and the factorial sequence $(n!)$, all follow Benford's law exactly.

Of course, not every data set 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. Also, common sequences such as the natural numbers and the primes do not follow Benford's law.

The ubiquity of Benford's law leads to the question of what properties real-life data sets must satisfy in order to follow Benford's law. A simple fact is the following. If a collection of numbers satisfies Benford's law, then it can be mathematically proven that 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 real-life datasets satisfy Benford's law. An explanation for this phenomenon was recently given by the American mathematician Theodore P. 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 T.P. 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. Tax returns that deviate from this law, over the course of many years, appear to be fraudulent in many cases. People just don't think Benford-like and therefore Benford's law is a useful diagnostic tool for identifying number-pattern anomalies worthy of closer investigation. Benford's law is applicable to surprisingly many large data sets on real-life activities prone to deception.

Benford's law has also applications in the natural sciences and in physics. It can be used to detect signals in contrast to background noise. An earthquake could be detected from just the first digit distribution of the ground displacement counts on a seismometer.

Multiplication game

Benford's law also underlies the amazing multiplication game that was analyzed in K. Morrison,"The multiplication game," *Mathematics Magazine* **83** (2010): 100–110. Imagine that a casino offers you the following game. The dealer hits a button and from a slot in the table comes a slip of paper with a four-digit positive integer on it that you cannot see. Next you use a keypad to choose a number of your own – any positive integer you like with as many digits as you wish. You are shown the slip of paper on which your integer is printed along with the product of the two integers. You win if the first digit of the product is 4 through 9 and you lose if it is 1, 2, or 3. The casino offers to give you \$2.40 back for each dollar bet if you win. This looks tempting, but it sounds almost too good to be true. Just to be sure, you calculate all the products of two four-digit numbers between 1,000 and 9,999 and you find that a proportion of 0.430 of the products are winners for you. So you decide to play the game. However, when playing the game, you will soon find out that the game is unfavorable to you. Whatever strategy you follow, the casino will win in the long-run over 60% of the time. How could this happen? The answer is that the casino uses randomization to choose its four-digit number. This goes as follows. Each time the casino generates a random number u from the uniform distribution on (0, 1) and then takes the largest integer that is less than or equal to $10^4 \times 10^u$. In this way the casino can guarantee for itself a win rate of at least $\log_{10}(4) - 10^{-3} = 0.60106$, whatever strategy the player uses. A truly amazing result! A proof of it can be found in the Morrison paper.[†]

5.11 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

[†] The number $x = 10^u$ can be seen as a random sample from the Benford density $f(x)$ which is defined as $f(x) = \frac{1}{\ln(10)} \frac{1}{x}$ for $1 < x < 10$ and $f(x) = 0$ otherwise. If the random variable X has the Benford density $f(x)$, then $P(X \leq x) = \log_{10}(x)$ for $1 \leq x \leq 10$. The following noteworthy result also holds. If the positive random variable Y is independent of X, then data sampled from the probability distribution of the product XY satisfy Benford's law.

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 $P(X_1 = k) = 1/s$ 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.

5.12 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.

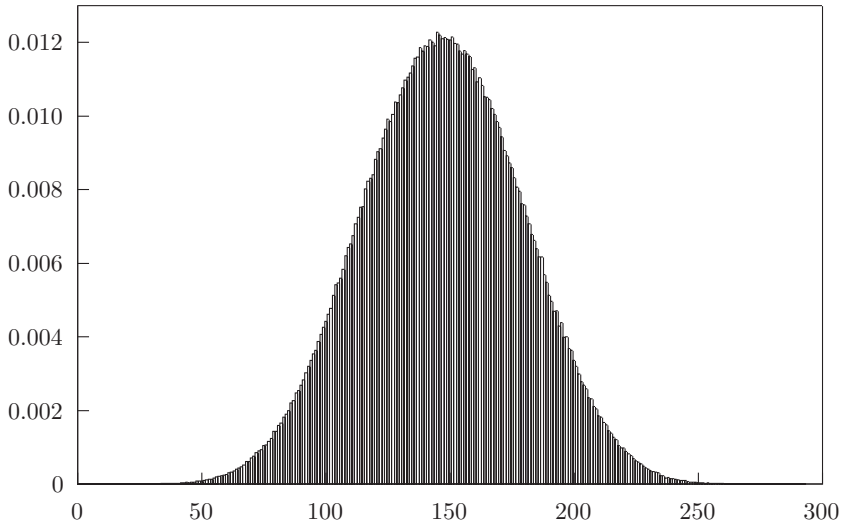


Fig. 5.11. Probability histogram for $r = 6$ and $s = 49$.

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 of whether to reject the null hypothesis at the 5% significance level instead of at, for example, the 0.1% level is also subjective. For some purposes, probability is best thought of as subjective. A subjective probability then describes your personal belief about how likely the null hypothesis is true. In the Bayesian approach probabilities are always conditional; they are assigned based on information.

The fundamental difference between the classical and Bayesian approach can be nicely illustrated with the following example. Imagine 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. The classical approach in statistics is to calculate the probability of correctly answering 26 or more of the 50 questions by luck alone. This excess probability is equal to 0.0049 and is called the p-value in classical statistics. On the basis of the small p-value one might conclude that the student is bluffing and in fact did prepare for the exam. However, one might have information about earlier achievements of the student in question. The Bayesian approach does use this information and then calculates the probability that the student did not prepare for the exam given that he/she answered 26 of the 50 questions correctly. This is the probability we are actually looking for. The Bayesian approach 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 may 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 20% the student was totally unprepared. In other words, $P(H) = 0.2$ and $P(\overline{H}) = 0.8$. Let us make the additional assumption that any of the 50 questions is answered correctly with probability $\frac{1}{3}$ if the student is totally unprepared and with probability 0.7 if the student is well-prepared. Where the classical

[†] The issue of subjectivity in statistical analysis is nicely discussed in J.O. Berger and D.A. Berry, "Statistical analysis and the illusion of objectivity," *American Scientist* **76** (1988): 159–165.

approach uses unobserved values that are more extreme than the observed value, the Bayesian approach uses only the observed value. The tool to revise the prior probability of the hypothesis in the light of the observed data is Bayes' rule. This rule is discussed in Chapter 8. Using Bayes' rule, we then conclude that with a probability of 0.1806 the student did not study and could only complete the exam by guessing, given the fact that he/she gave a correct answer to exactly 26 of the 50 questions. The Bayesian evidence against the hypothesis is not very strong. In general, the Bayesian approach is more careful in making a conclusion than the classical approach in which the p-value often exaggerates the strength of the evidence against the hypothesis.

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 analysis. Bayesian statistics is discussed in more detail in the Chapters 8 and 13. The spirit of Reverend Bayes (1702–1761) is still very much alive!

5.13 Problems

5.1 You draw twelve random numbers from (0,1) and average these twelve 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 $0.453\ (\pm 0.010)$. 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 millimeters and a standard deviation of 121.4 millimeters. Over many

years, what is the proportion of years that the annual rainfall in Amsterdam is below 550 millimeters?

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 group?

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 $\left(\sigma_B^2 - \sigma_A\sigma_B\rho_{AB}\right) / \left(\sigma_A^2 + \sigma_B^2 - 2\sigma_A\sigma_B\rho_{AB}\right)$. *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

$$\sigma^2\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma^2(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \text{cov}\left(X_i, X_j\right).$$

For the case that $X_1, \ldots, X_n$ all have the same variance σ^2 and $\text{cov}(X_i, X_j)$ 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 twenty 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 twenty 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 twenty-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 twenty 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 twenty 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 twenty years. What are the expected value and the standard deviation of the investor's capital after twenty 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 County 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 1.15277.

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.

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 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?

5.28 Use the gambler's ruin formula from Section 3.5 to derive the gambler's ruin formula for the Brownian motion process. *Remark*: use the fact that $\lim_{\Delta \to 0}(1 + a\Delta)^{1/\Delta} = e^a$ for any constant a.

5.29 A fair coin is tossed infinitely often. The initial bankroll is 1,000 dollars. Each time the coin comes up heads the current bankroll is increased by 50% but is decreased by 50% otherwise. Use the gambler's ruin formula for Brownian motion to approximate the probability that the size of the bankroll will ever become larger than the initial size of the bankroll. *Hint*: the process describing the logarithm of your bankroll is a random walk with a negative drift.

5.30 A drunkard is wandering back and forth on a line. At each step he moves two units distance to the right with a probability of $\frac{1}{2}$, or one unit to the left with a probability of $\frac{1}{2}$. The drunkard starts at the origin. Use Brownian motion to

approximate the probability that the drunkard will enter the negative part of the line during his first ten steps. What is the Brownian motion approximation of the probability that the drunkard will ever enter the negative part of the line?

5.31 Suppose that you are going to play a blackjack variant using a basic strategy for which the net payoff of a one-dollar bet has an expected value of −\$0.039 and a standard deviation of \$1.148. Your initial bankroll is \$400 and you bet ten dollars each time. What is an approximation for the probability of doubling your bankroll before going broke?

5.32 You are offered a sequence of favorable betting opportunities. Each bet results in one of four possible outcomes with probabilities $p_1 = 0.10$, $p_2 = 0.25$, $p_3 = 0.15$, $p_4 = 0.50$ and payoffs $f_1 = 4$, $f_2 = 2$, $f_3 = 1$, $f_4 = 0$ for each dollar staked. Calculate the Kelly betting fraction. What is the Brownian motion approximation for the probability of doubling your bankroll without having it halved first when you use half Kelly? What is an approximation for the probability of doubling your bankroll within 400 bets?

5.33 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 $\alpha^2\sigma^2$, 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 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 joke 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 joke 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 joke 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 $\frac{1}{3}$

[†] 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.

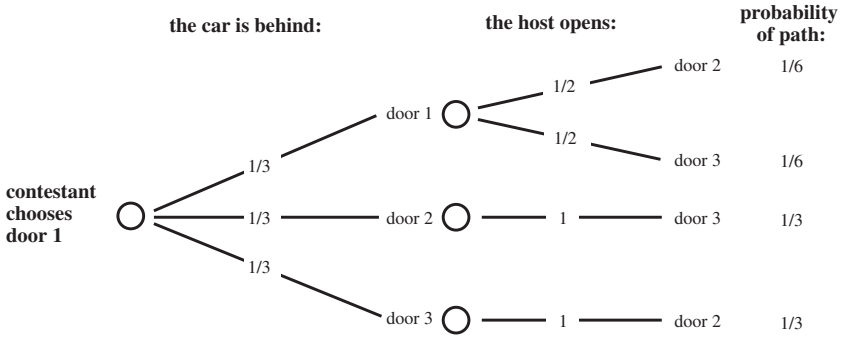


Fig. 6.1. Chance tree for the Monty Hall dilemma.

probability that the automobile will be hidden behind his chosen door, and a $\frac{2}{3}$ 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!

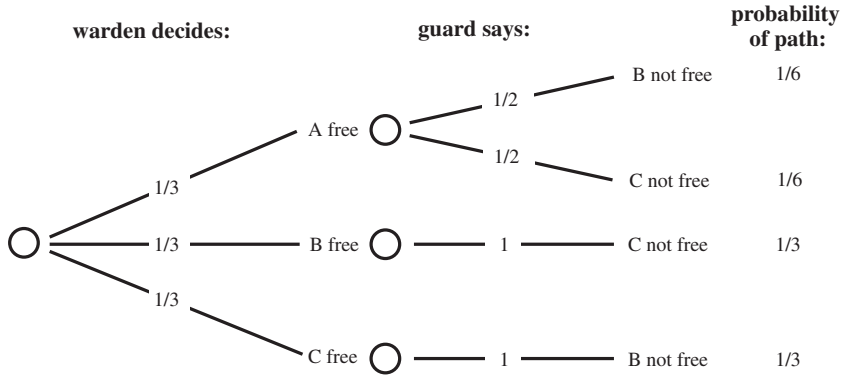


Fig. 6.2. Chance tree for the prisoner's problem.

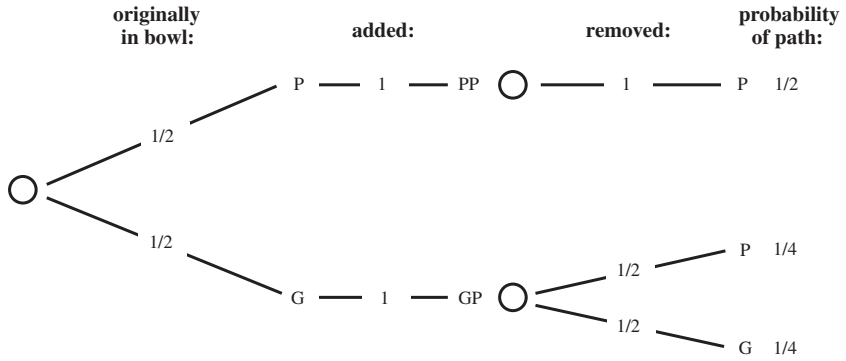


Fig. 6.3. Chance tree for the sushi delight problem.

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

$$P(\text{path 1} \mid \text{path 1 or path 2}) = \frac{P(\text{path 1})}{P(\text{path 1}) + P(\text{path 2})} = \frac{1/2}{1/2 + 1/4} = \frac{2}{3}.$$

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 6.8.

6.2 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 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 policlinic to test a subgroup of persons among whom it is known that 1 out of 2 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,

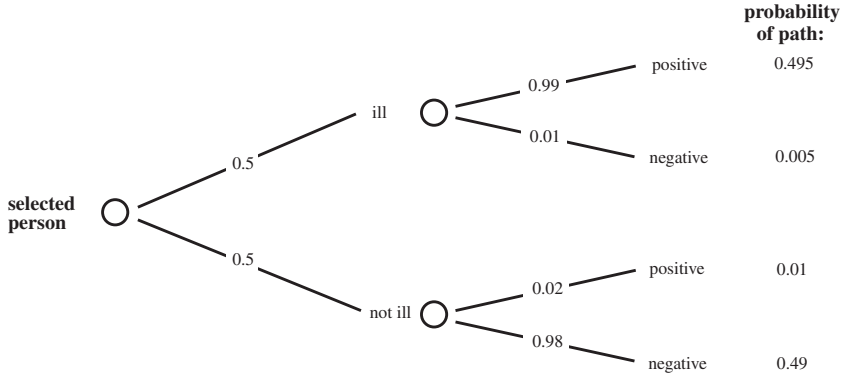


Fig. 6.4. Chance tree for the subgroup.

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 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.

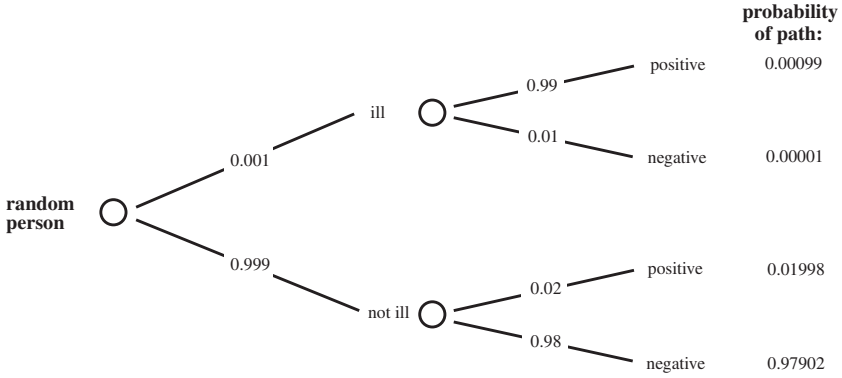


Fig. 6.5. Chance tree for the entire population.

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

[†] 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.

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

$$\begin{aligned}
&P(\text{disease}) = 0.001, \quad P(\text{no disease}) = 0.999,\\
&P(\text{positive} \mid \text{disease}) = 0.99, \quad P(\text{negative} \mid \text{disease}) = 0.01,\\
&P(\text{positive} \mid \text{no disease}) = 0.02, \quad P(\text{negative} \mid \text{no disease}) = 0.98.
\end{aligned}$$

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

$$\begin{aligned}
P(\text{positive}) &= P(\text{positive and disease}) + P(\text{positive and no disease})\\
&= P(\text{positive} \mid \text{disease})P(\text{disease})\\
&\quad + P(\text{positive} \mid \text{no disease})P(\text{no disease}).
\end{aligned}$$

Consequently, the desired probability $P(\text{disease} \mid \text{positive})$ satisfies the formula

$$\begin{aligned}
&P(\text{disease} \mid \text{positive})\\
&= \frac{P(\text{positive} \mid \text{disease})P(\text{disease})}{P(\text{positive} \mid \text{disease})P(\text{disease}) + P(\text{positive} \mid \text{no disease})P(\text{no disease})}.
\end{aligned}$$

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 can better calculate the posterior probability according to Bayes' rule by using first principles as done in the above example. Another insightful representation of the formula of Bayes will be discussed in Section 8.3.

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 10, nine 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?

[†] In his book *Calculated Risks* (Simon & Schuster, 2002) Gerd Gigerenzer advocates that doctors and lawyers be educated in more understandable representations of risk.

6.3 You are one of fifty thousand 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 fifty thousand 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 joke 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 joke 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 joke 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 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 twenty drivers on the roads in the region in question is driving under the influence of alcohol. 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. The witness identified the cab as White. 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 one out of one hundred 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 one 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 one 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?

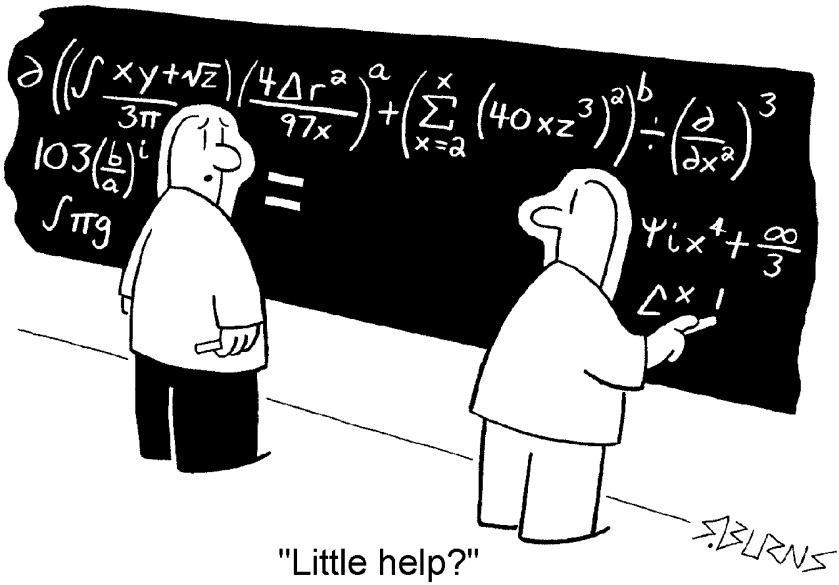
6.16 A sum of money is placed in each of two envelopes. The amounts differ from one another, but you do not know what the values of the two amounts are. You do know that the values lie between two boundaries m and M with $0 < m < M$. You choose an envelope randomly. After inspecting its contents, you 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.

1. Choose an envelope and look to see how much cash is inside.
2. Pick a random number between m and M.
3. 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



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 so-called sample space, where the sample space represents the set of all possible outcomes the experiment. 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 is the fundamental link between theory and the real world. Its proof has to be postponed until Chapter 14.

The purpose of this chapter is to discuss the axioms of probability theory and to derive from the axioms a number of 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. In a few of the examples counting methods and binomial coefficients will be used. A short but self-contained treatment of these basic tools from combinatorics can be found in the Appendix at the end of this book.

7.1 Probabilistic foundations

A probability model is a mathematical representation of a real-word situation or a random experiment. It consists of a complete description of all possible outcomes of the experiment and an assignment of probabilities 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. Let us give a few examples.

- The experiment is to toss a coin once. The sample space is the set $\{H, T\}$, 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.
- The experiment is to roll a die once. 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 gets assigned a probability of $\frac{1}{6}$ if the die is fair.
- The experiment is to choose a letter at random from the word statistics. The sample space is the set $\{s, t, a, i, c\}$. The probabilities $\frac{3}{10}$, $\frac{3}{10}$, $\frac{1}{10}$, $\frac{2}{10}$, and $\frac{1}{10}$ are assigned to the five outcomes s, t, a, i, and c.
- The experiment is to repeatedly roll a fair die until the first six shows up. The sample space is the set $\{1, 2, \ldots\}$ of the positive integers. Outcome k indicates that the first six shows up on the kth roll. The probabilities $\frac{1}{6}, \frac{5}{6} \times \frac{1}{6}, (\frac{5}{6})^2 \times \frac{1}{6}, \ldots$ can be assigned to the outcomes $1, 2, 3, \ldots$.
- The experiment is to measure the time until the first emission of a particle from a radioactive source. The sample space is the set $(0, \infty)$ of the positive real numbers, where the outcome t indicates that it takes a time t until the first emission of a particle. Taking an appropriate unit of time, the probability $\int_a^b e^{-t}\, dt$ can be assigned to each time interval (a, b) on the basis of physical properties of radioactive material, where $e = 2.71828\ldots$ is the base of the natural logarithm.

Various choices for the sample space are sometimes possible. In the experiment of tossing a coin twice, a possible choice for the sample space is the set $\{HH, HT, TH, TT\}$. Another possible choice is the set $\{0, 1, 2\}$, where the outcome indicates the number of heads obtained. The assignment of probabilities to the elements of the sample space differs for the two choices.

In the first three examples above the sample space is a *finite* set. In the fourth example the sample space is a so-called *countably infinite* set, while in the fifth example the sample space is a so-called *uncountable* set. Let us briefly

explain these basic concepts from set theory. 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 countably infinite 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 countably infinite. 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, whereas sets that are either finite or countably infinite are called countable.

7.1.1 Axioms of probability theory

A probability model consists of a sample space together with the assignment of probability, where probability is a function that assigns numbers between 0 and 1 to subsets of the sample space. The axioms of probability are mathematical rules that the probability function must satisfy. In the informal Section 2.2.2, we discussed already these rules for the case of a finite sample space. The axioms of probability are essentially the same for a chance experiment with a countable or 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. The probability measure on the sample space is denoted by P. It

assigns to each subset A a probability $P(A)$ and 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 $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ *for every collection of pairwise disjoint subsets* $A_1, A_2, \ldots$.

The *union* $\bigcup_{i=1}^{\infty} A_i$ of the subsets $A_1, A_2, \ldots$ is defined as 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. 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* (or disjoint) if the corresponding sets $A_1, A_2, \ldots$ are pairwise disjoint.

The first two axioms simply express a probability as a number between 0 and 1. The crucial third axiom 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. Starting with these three axioms and a few definitions, a powerful and beautiful theory of probability can be developed.

The standard notation for the sample space is the symbol Ω. An outcome of the sample space is denoted by ω. A sample space together with a collection of events and an assignment of probabilities to the events is called a *probability space*. For a finite or countably infinite sample space Ω, it is sufficient to assign a probability $p(\omega)$ to each element $\omega \in \Omega$ such that $p(\omega) \geq 0$ 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

$$P(A) = \sum_{\omega \in A} p(\omega).$$

In other words, $P(A)$ is the sum of the individual probabilities of the outcomes ω that belong to the set A. It is left to the reader to verify that P satisfies the Axioms 7.1 to 7.3.

A probability model is constructed with a specific situation or experiment in mind. The assignment of probabilities is part of the translation process from a concrete context into a mathematical model. Probabilities may be assigned to events any way you like, as long the above axioms are satisfied. To make

your choice of the probabilities useful, the assignment should result in a "good" model for the real-world situation.

Equally likely outcomes

In many experiments with finitely many outcomes $\omega_1, \ldots, \omega_N$ it is natural to assume that all these outcomes are equally likely to occur. In such a case, $p(\omega_i) = \frac{1}{N}$ for $i = 1, \ldots, N$ and each event A gets assigned the probability

$$P(A) = \frac{N(A)}{N},$$

where $N(A)$ is the number of outcomes in the set A. This model is sometimes called the classical probability model.

Example 7.1 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 chosen as $\{(i, j, k)$: $i, j, k = 1, \ldots, 6\}$, 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 7.2 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. The players decided upon the following strategy. A player who sees that the other two players wear a hat with the same color guesses the opposite color for his/her own hat; otherwise, the player says nothing. What is the probability of winning the game under this strategy?

Solution. This chance experiment can be seen as tossing a fair coin three times. As sample space, we take the set consisting of the eight elements RRR, RRB,

RBR, BRR, BBB, BBR, BRB and RBB, where R stands for a red hat and B for a blue hat. Each element of the sample space is equally probable and gets assigned a probability of $\frac{1}{8}$. The strategy is winning if one the six outcomes RRB, RBR, BRR, BBR, BRB or RBB occurs (verify!). Thus the probability of winning the game under the chosen strategy is $\frac{3}{4}$.

Uncountable sample space

The following two examples illustrate the choice of a probability measure for an uncountable sample space. These two examples deal with so-called geometric probability problems. In the analysis of the geometric probability problems we use the continuous analog of the assignment of probabilities in the probability model having a finite number of equally likely outcomes.

Example 7.3 You randomly throw a dart at a circular dartboard with radius R. It is assumed that the dart is infinitely sharp and lands on a completely random point on the dartboard. How do you calculate the probability of the dart hitting the bull's-eye having radius b?

Solution. The sample space of this experiment consists of the set of pairs of real numbers (x, y) with $x^2 + y^2 \leq R^2$, where (x, y) indicates the point at which the dart hits the dartboard. This sample space is uncountable. We first make the following observation. The probability that the dart lands exactly on a *prespecified* point is zero. It makes only sense to speak of the probability of the dart 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.The assumption of the dart hitting the dartboard on a completely 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. Hence the probability of the dart hitting the bull's-eye is $\pi b^2/(\pi R^2) = b^2/R^2$.

Example 7.4 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

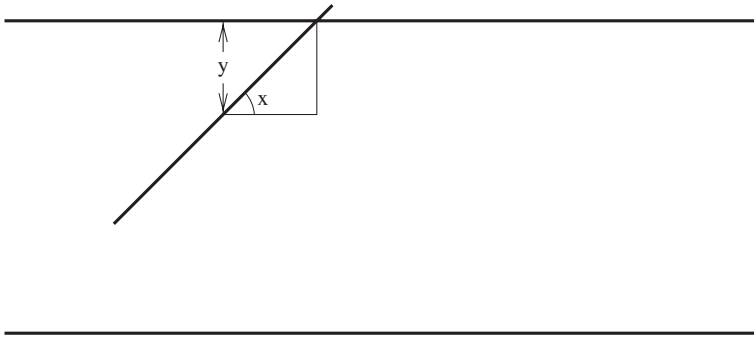


Fig. 7.1. The landing of Buffon's needle.

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 $0 \leq y \leq \frac{1}{2}D$. Dropping the needle at random on the floor can be seen to be equivalent to choosing a random point in the rectangle R. 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

$$P(\text{needle intersects one of the lines}) = \frac{2L}{\pi D}.$$

In the following problems you should first specify a sample space before calculating the probabilities.

Problem 7.1 Three fair dice are rolled. What is the probability that the sum of the three numbers shown is an odd number? What is the probability that the product of the three numbers shown is an odd number?

Problem 7.2 In a township, there are two plumbers. On a particular day three 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?

Problem 7.3 Four black socks and five white socks lie mixed up in a drawer. You grab two socks at random from the drawer. What is the probability of having grabbed one black sock and one white sock?

Problem 7.4 Two letters have fallen out of the word Cincinnati at random places. What is the probability that these two letters are the same?

Problem 7.5 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 7.6 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 7.7 You have four mathematics books, three physics books and two chemistry books. The books are put in random order on a bookshelf. What is the probability of having the books ordered per subject on the bookshelf?

Problem 7.8 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. What is the probability that one of the three friends will have to pay for all three tickets?

Problem 7.9 You choose a letter at random from the word Mississippi eleven times without replacement. What is the probability that you can form the word Mississippi with the eleven chosen letters? *Hint*: it may be helpful to number the eleven letters as $1, 2, \ldots, 11$.

Problem 7.10 You choose a number at random from the numbers $1, 2, \ldots, 100$ ten times. What is the probability of choosing ten distinct numbers? What is the probability that the first number chosen is larger than each of the other nine numbers chosen?

Problem 7.11 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 7.12 Two people have agreed to meet at the train station. Independently of one another, each person is to appear at a completely random moment between between 12 p.m and 1 p.m. What is the probability that the two persons will meet within 10 minutes of one another?

Problem 7.13 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 $q > 0$.

Problem 7.14 A point is chosen at random within the unit square. What is the probability that the product of the two coordinates of the point is larger than 0.5?

Problem 7.15 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 a given value d with $0 < d < h$? 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?

Problem 7.16 Consider the following variant of Buffon's needle problem from Example 7.4. A rectangular card with side lengths a and b is dropped at random on the floor. It is assumed that the length $\sqrt{a^2 + b^2}$ of the diagonal of the card is smaller than the distance D between the parallel lines on the floor. Show that the probability of the card intersecting one of the lines is given by $\frac{2(a+b)}{\pi D}$.

Problem 7.17 Choose randomly a point within a circle with radius r and construct the (unique) chord with the chosen point as its midpoint. What is the probability that the chord is longer than a side of an equilateral triangle inscribed in the circle?

Problem 7.18 A stick is broken in two places. The breaking points are chosen at random on the stick, independently of each other. What is the probability that a triangle can be formed with the three pieces of the broken stick? *Hint*: the sum of the lengths of any two pieces must be greater than the length of the third piece.

Problem 7.19 You choose a number v at random from $(0, 1)$ and next a number w at random from $(0, 1 - v)$. What is the probability that a triangle can be formed with the side lengths v, w and $1 - v - w$? *Hint*: represent w as $y(1 - v)$,

where y is a random number from (0, 1). Next answer the following question. What is the probability that a triangle can be formed with three pieces of a broken stick if the stick is first broken at random into two pieces and next the longer piece is randomly broken into two.

7.1.2 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 $n \geq 1$. For a nondecreasing sequence $\{E_n\}$, let us use the notation $E = \lim_{n\to\infty} E_n$ for the union $E = \bigcup_{i=1}^{\infty} E_i$. Then, the continuity property states that

$$\lim_{n\to\infty} P(E_n) = P(\lim_{n\to\infty} E_n).$$

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,

$$\begin{aligned} P\big(\lim_{n\to\infty} E_n\big) &= P\Big(\bigcup_{i=1}^{\infty} E_i\Big) = P\Big(\bigcup_{i=1}^{\infty} F_i\Big) = \sum_{i=1}^{\infty} P(F_i) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P(F_i) = \lim_{n\to\infty} P\Big(\bigcup_{i=1}^{n} F_i\Big) \\ &= \lim_{n\to\infty} P\Big(\bigcup_{i=1}^{n} E_i\Big) = \lim_{n\to\infty} P(E_n), \end{aligned}$$

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 this case the set $\lim_{n\to\infty} E_n$ denotes the intersection $\bigcap_{i=1}^{\infty} E_i$ of the sets $E_1, E_2, \ldots$. The *intersection* of the sets $E_1, E_2, \ldots$ is defined as the set of all outcomes that belong to each of the sets E_i.

Problem 7.20 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) $P\big(\bigcup_{k=1}^{\infty} A_k\big) \leq \sum_{k=1}^{\infty} P(A_k)$ for any sequence of subsets $A_1, A_2, \ldots$ (this result is known as *Boole's inequality*).

Problem 7.21 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 $C = \{\omega : \omega \in A_k \text{ for infinitely many } k\}$. 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 7.5 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 7.6 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. Take as sample space for this chance experiment 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. For notational convenience, let us represent 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 ∞. Then we can also express the sample space Ω as

$$\Omega = \{1, 2, \ldots\} \cup \{\infty\}.$$

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

$$p(n) = \left(\frac{5}{6}\right)^{n-1} \frac{1}{6}$$

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

$$\begin{aligned} P(A) &= \sum_{n=0}^{\infty} p(2n+1) = \sum_{n=0}^{\infty} \left(\frac{5}{6}\right)^{2n} \frac{1}{6} \\ &= \frac{1}{6} \sum_{n=0}^{\infty} \left(\frac{25}{36}\right)^{n} = \frac{1}{6} \left(\frac{1}{1 - \frac{25}{36}}\right) = 0.5436. \end{aligned}$$

Problem 7.22 In a tennis tournament between three players A, B, and C, each player plays the others once. The strengths of the players 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 7.23 Two people take turns selecting a ball at random from a bowl containing 3 white balls and 7 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 7.24 In repeatedly rolling two dice, what is the probability of getting a total of 6 before a total of 7 and what is the probability of getting a total of 8 before a total of 6 and before a total of 7? What is the probability of getting a total of 6 and a total of 8 in any order before two 7s?

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

[†] This section can be skipped without loss of continuity.

how we deal with such experiments by way of an illustration of a compound 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. This sample space is uncountable. The nontrivial proof of this remarkable fact will not be given. In order to be able to define a probability measure on the uncountable 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 elements ω have the finite beginning $T, T, \ldots, T, H$, and thus the event 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 Hs). 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 term almost sure convergence is also used for this type of convergence. The terminology of an "almost sure" event means that realizations not in this event are theoretically possible but will not happen in reality. 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 14.4).

Problem 7.25 In the coin-tossing experiment of repeatedly tossing a fair coin, a run of length r is said to have occurred if either heads or tails have just been tossed r times in a row. Prove that a run of length r is certain to occur somewhere if a fair coin is tossed indefinitely often.[†] *Hint*: analyze the coin-tossing experiment as an infinite sequence of physically independent r-experiments. In each r-experiment the coin is tossed r times. Define B_n as the event that no run of length r occurs in the r-experiments $1, \ldots, n$ and evaluate $P(\bigcap_{n=1}^{\infty} B_n)$.

Problem 7.26 You toss a fair coin until you obtain 10 heads in a row and then you stop. What is the probability of seeing at least 10 consecutive tails in the sequence prior to stopping?

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 denoted by AB. Another notation for the intersection is $A \cap B$. 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.

[†] Actually, we have the stronger result stating that a run of length r will occur infinitely often with probability 1.

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,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{n} P(A_i).$$

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 now 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 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

$$\begin{aligned} & s - \binom{s}{2} + \binom{s}{3} - \cdots + (-1)^{s-1} \binom{s}{s} \\ &= 1 - \left[\binom{s}{0} - \binom{s}{1} + \binom{s}{2} - \binom{s}{3} + \cdots + (-1)^{s} \binom{s}{s}\right] \\ &= 1 - (-1+1)^{s} = 1, \end{aligned}$$

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.

In the examples below the *fundamental counting principle* will be used. This principle can be formulated as follows. Suppose a task involves a sequence of r choices. Let n_1 be the number of possible ways the first choice can be made, and let n_2 be the number of possible ways the second choice can be made after the first choice has been done, and let n_3 be the number of possible ways the third choice can be made after the first two choices have been done, etc. Then the total number of different ways the task can be performed is $n_1 \times n_2 \times \cdots \times n_r$.

The complement rule will be illustrated with the following problem posed by the gambler Chevalier de Méré to the famous French mathematician Blaise Pascal in 1654. This problem marks the beginning of probability theory.

Example 7.7 How many rolls of a fair die are required to have at least a 50% chance of rolling at least one six? How many rolls of two fair dice are required to have at least a 50% chance of rolling at least one double six?

Solution. Let us first consider the experiment of rolling a single die r times, where r is fixed in advance. The sample space of this experiment consists of

all elements $(i_1, i_2, \ldots, i_r)$, where i_k denotes the outcome of the kth roll of the die. The sample space has $6 \times 6 \times \cdots \times 6 = 6^r$ elements. All elements are equally likely. Let A be the event that at least one six is obtained in the r rolls of the die. To compute $P(A)$, it is easiest to compute the probability of the complementary event A^c that no six is obtained in the r rolls of the die. The set A^c consists of $5 \times 5 \times \cdots \times 5 = 5^r$ elements. Hence $P(A^c) = 5^r/6^r$ and so

$$P(A) = 1 - \frac{5^r}{6^r}.$$

This probability has the value 0.4213 for $r = 3$ and the value 0.5177 for $r = 4$. Hence four rolls of the die are required.

The following sample space is taken for the experiment of rolling two dice r times. Imagining that one die is blue and the other is red, the sample space consists of all elements $((i_1, j_1), (i_2, j_2), \ldots, (i_r, j_r))$, where i_k and j_k denote the outcomes of the blue and the red die on the kth roll of the two dice. The sample space consists of $36 \times 36 \times \cdots \times 36 = 36^r$ elements. All elements are equally likely. Let A be the event that at least one double six is obtained in the r rolls of the two dice. The complementary event A^c of rolling no double six in r rolls of the two dice can occur in $35 \times 35 \times \cdots \times 35 = 35^r$ ways. Hence $P(A^c) = 35^r/36^r$ and so

$$P(A) = 1 - \frac{35^r}{36^r}.$$

This probability has the value 0.4914 for $r = 24$ and the value 0.5055 for $r = 25$. Hence 25 rolls of the two dice are required.

Rule 7.3 is often called the *addition rule* for probabilities and is illustrated with the following example.

Example 7.8 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

$$P(A \cup B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}.$$

A more elaborate illustration of Rule 7.3 will be given in the following example.

Example 7.9 The eight soccer teams which have reached the quarterfinals of the Champions League are formed by two teams from each of the countries England, Germany, Italy and Spain. The four matches to be played in the quarterfinals are determined by drawing lots.
(**a**) What is the probability that the two teams from the same country play against other in each of the four matches?
(**b**) What is the probability that there is a match between the two teams from England or between the two teams from Germany?

Solution. Number the eight teams as $1, \ldots, 8$ and take as sample space all possible permutations of the integers $1, \ldots, 8$. Any permutation $t_1, \ldots, t_8$ corresponds to the four matches (t_1, t_2), (t_3, t_4), (t_5, t_6) and (t_7, t_8). The total number of elements in the sample space is $8 \times 7 \times \cdots \times 1 = 8!$. Each element is equally likely and gets assigned the same probability of $1/8!$.
(**a**) Let A be the event that the two teams from the same country play against each other in each of the four matches. The event A corresponds to $8 \times 1 \times 6 \times 1 \times 4 \times 1 \times 2 \times 1 = 384$ elements in the sample space and so

$$P(A) = \frac{384}{8!} = 0.0095.$$

(**b**) Let E be the event that there is a match between the two teams from England and let G be the event that there is a match between the two teams from Germany. The desired probability $P(E \cup G)$ satisfies $P(E \cup G) = P(E) + P(G) - P(EG)$. The events E and G each correspond to $4 \times (2 \times 1 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1) = 5760$ elements in the sample space and the event EG corresponds to $4 \times 3 \times (2 \times 1 \times 2 \times 1 \times 4 \times 3 \times 2 \times 1) = 1152$ elements in

the sample space and so

$$P(E) = P(G) = \frac{5760}{8!} \quad \text{and} \quad P(EG) = \frac{1152}{8!}.$$

This gives

$$P(E \cup G) = \frac{5760}{8!} + \frac{5760}{8!} - \frac{1152}{8!} = 0.2571.$$

The following example deals with a variant of the birthday problem. The classical birthday problem asks for the probability that in a class of n children two or more children have the same birthday. Assuming that the year has 365 days and that all possible birthdays are equally likely, the complement rule gives that this probability equals $1 - (365 \times 364 \times \cdots \times (365 - n + 1))/365^n$. The smallest value of n for which this probability is 50% or more is $n = 23$. The probability has the value 0.5073 for $n = 23$. The event that *at least* two children have the same birthday can occur in several ways. In the following example we calculate the probability that *exactly* two children have the same birthday.

Example 7.10 What is the probability that in a class of 23 children exactly two children have birthdays on the same day?

Solution. Take as the sample space all ordered 23-tuples of numbers selected from the integers $1, 2, \ldots, 365$, where the first number of the tuple is the birthday of the first child, the second number is the birthday of the second child, etc. The total number of elements of the sample space is $365 \times 365 \times \cdots \times 365 = 365^{23}$. Each element of the sample space is equally likely and thus gets assigned a probability of $1/365^{23}$. The trick is now to consider all possible combinations of two children from the group of 23 children. The total number of such combinations is given by the binomial coefficient $\binom{23}{2} = 253$. Number the 253 possible combinations of two children as $1, \ldots, 253$ and define A_i as the event that the two children from combination i are the only two children having the same birthday. The events $A_1, \ldots, A_{253}$ are mutually exclusive and so the desired probability is given by $P(A_1 \cup A_2 \cup \cdots \cup A_{253})$. For any fixed i, the event A_i corresponds to $365 \times 1 \times 364 \times 363 \times \cdots \times (365 - 21)$ elements in the sample space and so

$$P(A_i) = \frac{365 \times 1 \times 364 \times 363 \times \cdots \times 344}{365^{23}} = 0.0014365.$$

By Rule 7.1, $P(A_1 \cup A_2 \cup \cdots \cup A_{253}) = P(A_1) + P(A_2) + \cdots + P(A_{253})$ and so the probability that in a class of 23 children exactly two children have the same birthday is equal to $253 \times 0.0014365 = 0.3634$. It is left to the reader to verify that the probability of having exactly two days on which exactly two

children have the same birthday has the value 0.1109. Adding up the probabilities 0.3634 and 0.1109, we come close to value 0.5073 of the probability that at least two children have the same birthday.

Problem 7.27 What is the probability of getting at least one ace in a poker hand of five cards dealt from an ordinary deck of 52 cards? What is the probability of getting five cards of the same suit in a poker hand of five cards?

Problem 7.28 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 7.29 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 7.30 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 numbers $1, \ldots, 6$ 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?

Problem 7.31 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 7.32 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 all of the three destinations will be chosen.

Problem 7.33 You roll a fair die six times in a row. What is the probability that all of the six face values will appear? What is the probability that one or more sixes will appear?

Problem 7.34 In the Lotto 6/42, six distinct numbers are picked at random from the numbers $1, \ldots, 42$. What is the probability that number 10 will be picked?

Problem 7.35 Two black socks, two brown socks and one white sock lie mixed up in a drawer. You grab two socks without looking. What is the probability that you have grabbed two black socks or two brown socks?

Problem 7.36 John and Paul play the following game. They each roll one fair die. John wins the game if his score is larger than Paul's score or if the product of the two scores is an odd number. Is this a fair game?

Problem 7.37 Consider a random permutation of the numbers $1, 2, \ldots, n$ and answer the following questions.
(**a**) What is the probability that a particular number r belongs to a cycle of a given length k?
(**b**) What is the probability that the cycle covering the particular number r contains only numbers larger than or equal to r? *Hint*: let A_k be the event that the cycle covering the number r has length k and contains only numbers larger than or equal to r. Use the identity $\sum_{j=s}^{m} \binom{j}{s} = \binom{m+1}{s+1}$.
(**c**) What is the probability that two particular numbers r and s belong to a same cycle?

A cycle of the permutation is a subset whose elements trade places with one another in the permutation; for example, in the permutation $(2, 4, 3, 1)$ of $(1, 2, 3, 4)$, the set $\{1, 4, 2\}$ is a cycle of length 3 with $1 \to 4 \to 2 \to 1$.

Problem 7.38 Consider Problem 2.46 again. Give an explicit expression for the probability that the group will not be exiled. *Hint*: consider the contents of the boxes as a random permutation of the integers $1, 2, \ldots, 2n$ with $n = 50$ and verify that the probability of having no cycle of length $n + 1$ or more is equal to $1 - \sum_{k=n+1}^{2n} 1/k$.

Problem 7.39 Two-up is a traditional Australian gambling game. A spinner repeatedly throws two pennies up in the air. "Heads" appear if both pennies land with the head side facing up, "tails" appear if both pennies land with the tails side facing up, and "odds" appear if one penny lands with the head side up and the other lands with the tails side up. Consider the following game variant. The spinner wins if three "heads" appear before any appearance of "tails" and also before any intervening sequence of five consecutive "odds". A win pays 8.5 for 1. What is the win probability of the spinner? Is the game favorable for the bettor?

Problem 7.40 A random experiment has the possible outcomes $O_1, O_2, \ldots$ with respective probabilities $p_1, p_2, \ldots$. Independent repetitions of the experiment are performed. Define an appropriate probability space to verify that the probability of the outcome O_1 occurring before the outcome O_2 is given by $p_1/(p_1 + p_2)$. Also, verify the more general result that the probability of the rth appearance of outcome O_1 occurring before the sth appearance of outcome O_2 is given by

$$\sum_{k=0}^{s-1} \binom{r+k-1}{r-1} \left(\frac{p_1}{p_1+p_2}\right)^r \left(\frac{p_1}{p_1+p_2}\right)^k \quad \text{for any } r, s \geq 1.$$

Hint: define $A_{n,k}$ as the event that the rth appearance of outcome O_1 occurs at the nth trial and that outcome O_2 appears exactly k times in the first $n-1$ trials. Use the identity $\sum_{l=0}^{\infty}(l+1)\cdots(l+m)\,x^l = m!/(1-x)^{m+1}$ for any $m = 1, 2, \ldots$ and $|x| < 1$.

Problem 7.41 From an ordinary deck of 52 cards, cards are randomly drawn one by one and without replacement. What is the probability that an ace will appear before any of the cards 2 through 10?

Problem 7.42 Twenty-five people each choose two distinct numbers at random from the numbers $1, 2, \ldots, 25$, independently of each other. What is the probability that at least two people will choose the same two numbers?

Inclusion-exclusion rule

Many probability problems can be reduced to finding the probability of the occurrence of at least one event from a sequence of appropriately defined events. How to do this is explained in Rule 7.4. This rule extends Rule 7.3 and is known as the *inclusion-exclusion rule*. It 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 7.11 Letters to n different persons are randomly put into n preaddressed 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

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \binom{n}{1}\frac{(n-1)!}{n!} - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} - \cdots + (-1)^{n-1}\binom{n}{n}\frac{1}{n!}.$$

Here we use the fact that the binomial coefficient $\binom{n}{k}$ gives the total number of ways one can choose k different events from the events $A_1, \ldots, A_n$. 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 $e = 2.71828\ldots$ (see the Appendix). Thus, for large n, the probability that at least one person will receive the correct letter is approximately equal to $1 - e^{-1} = 0.6321$, 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 so-called Poisson probability $e^{-1}/j!$ as the number of envelopes becomes large. The Poisson approximation is very accurate for n not too small and was already obtained by other arguments in Section 4.2.3. Moreover, several real-life applications of the envelope problem are discussed in this section.

Many probability problems of a combinatorial nature can be solved by using the inclusion-exclusion rule. We give two more examples.

Example 7.9 (continued) What is the probability that in none of the four matches two teams from the same country play against each other?

Solution. Let A_i denote the event that two teams from the same country play against each other in the ith match that results from drawing lots. Using the complement rule, the desired probability is given by $1 - P(A_1 \cup A_2 \cup A_3 \cup A_4)$. In the solution of part (b) we already found that

$$P(A_i) = \frac{5760}{8!} \quad \text{for all } i \quad \text{and} \quad P(A_iA_j) = \frac{1152}{8!} \quad \text{for all } i < j.$$

For any $i < j < k$ the event $A_iA_jA_k$ corresponds to $4 \times 3 \times 2 \times (2 \times 1 \times 2 \times 1 \times 2 \times 1 \times 2 \times 1) = 384$ elements in the sample space and so

$$P(A_iA_jA_k) = \frac{384}{8!}.$$

Since the event A_4 is implied by the event $A_1A_2A_3$, it follows that $P(A_1A_2A_3A_4)$ is equal to $P(A_1A_2A_3) = \frac{384}{8!}$. Applying the inclusion-exclusion

rule, we find

$$P(A_1 \cup A_2 \cup A_3 \cup A_4) = \binom{4}{1}\frac{5760}{8!} - \binom{4}{2}\frac{1152}{8!} + \binom{4}{3}\frac{384}{8!} - \binom{4}{4}\frac{384}{8!}$$
$$= 0.4286$$

and so the desired probability is $1 - 0.4286 = 0.5714$.

Example 7.12 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 7.43 Three people each choose five distinct numbers at random from the numbers $1, 2, \ldots, 25$, independently of each other. What is the probability that the three choices of five numbers have no number in common?

Problem 7.44 Consider Problem 7.43 again. Calculate the probability that the three choices of five numbers have exactly j numbers in common for $j = 1, 2$. *Hint*: apply the trick used in Example 7.11 to find the probability that exactly j persons receive a correct letter.

Problem 7.45 What is the probability that in a thirteen-card bridge hand at least one suit will be missing? What is the probability that in a five-card poker hand at least one suit will be missing?

Problem 7.46 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 thirteen ranks four times each. A match occurs if the rank the dealer is calling corresponds to the card he turns over. A "supermatch" occurs if all four cards

of some same rank are matched (e.g., all four aces). What is the probability that a "supermatch" will occur?

Problem 7.47 What is the probability that a hand of 13 cards from an ordinary deck of 52 cards contains four of a kind?

Problem 7.48 What is the probability that not all of the numbers $1, 2, \ldots, 45$ will come up in 30 drawings of the Lotto 6/45? In each drawing six distinct numbers are chosen at random from $1, 2, \ldots, 45$.

Problem 7.49 Consider the coupon collector's problem from Section 3.2 again. Use the inclusion-exclusion formula to find the probability of collecting all of the n flippos in no more than r purchases. As an application, compute the answer to the following question. How many spins of the wheel in American roulette are required to have at least an even chance that the ball has landed in all of the 38 sections?

Problem 7.50 Three people each write down the numbers $1, \ldots, 10$ in a random order. What is the probability that the three people all have at least one number in the same position?

Problem 7.51 Each of ten persons chooses one name at random from the names of the other nine persons. What is the probability that at least two person choose each other's name?

Problem 7.52 Three are n people in a room and each of them is armed and angry. At a chime of a clock, everyone spins around and shoots at random one other person. What is the probability that there are no survivors after the first round?

8

Conditional probability and Bayes

The concept of conditional probability lies at the heart of probability theory. It is an intuitive concept. To illustrate this, 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 52 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 $\frac{4}{52} \times \frac{3}{51}$. Letting A_1 be the event that the first card is an ace and A_2 the event that the second card is an ace, one intuitively applies the fundamental formula $P(A_1A_2) = P(A_1)P(A_2 \mid A_1)$, where $P(A_2 \mid A_1)$ is the notation for the conditional probability that the second card will be an ace given that the first card was an ace.

The purpose of this chapter is to present the basics of conditional probability. You will learn about the multiplication rule for probabilities and the law of conditional probabilities. These results are extremely useful in problem solving. Much attention will be given to Bayes' rule for revising conditional probabilities in light of new information. This rule is inextricably bound up with conditional probabilities. The odds form of Bayes' rule is particularly useful and will be illustrated with several examples. Following on from Bayes' rule, we explain Bayesian inference for discrete models and give several statistical applications.

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 $P(A \mid B)$. The relative frequency interpretation also tells us how a conditional probability must be estimated in a simulation study.

Example 8.1 Someone has rolled two dice. You know that one of the dice turned up a face value of six. What is the probability that the other die turned up a six as well?

Solution. Let A be the event that two sixes show up and B the event that at least one six shows up. The sample space of this experiment is the set $\{(i, j) \mid i, j = 1, \ldots, 6\}$, where i and j denote the outcomes of the two dice. A probability of $\frac{1}{36}$ is assigned to each element of the sample space. The event A is given by the set $A = \{(6, 6)\}$ and the event B by the set

$B = \{(1, 6), \ldots, (5, 6), (6, 6), (6, 5), \ldots, (6, 1)\}$. The probability of event B occurring is $\frac{11}{36}$ and the probability of both event A and event B occurring is $\frac{1}{36}$. Given that you know that event B has occurred, the probability that event A has also occurred is $P(A \mid B)$. 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 $\frac{1}{6}$. An intuitive explanation of the difference between the two probabilities is the fact that in the second scenario you have the additional information which one of the two dice (red and blue) shows a six.

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). Let A be the event that John misses and let B be 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 $P(B) = \frac{3}{60} + \frac{4}{60} + \frac{2}{60} = \frac{9}{60}$. We are now ready to determine the conditional probability $P(A \mid B)$. Applying the formula $P(A \mid B) = P(AB)/P(B)$, we can conclude that

$$P(\text{John misses} \mid \text{exactly two darts hit the target}) = \frac{2}{9}.$$

Problem 8.1 Three fair dice are rolled. What is the conditional probability that the sum of the three faces is 10 given that the three dice are showing different faces?

Problem 8.2 You toss a nickel (5 cents), a dime (10 cents) and a quarter (25 cents), independently of each other. What is the conditional probability that the quarter shows up heads given that the coins showing up heads represent an amount of at least 15 cents?

Problem 8.3 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 8.4 The experiment is to toss a fair coin once and to roll a fair die once. Let A be the event that the coin lands "heads" and B the event that the die lands "six". After the experiment, you are informed that at least one of the two events has occurred. What is the probability that both events have occurred and what is the probability that event A has occurred, given the information?

Problem 8.5 You simultaneously grab two balls at random from an urn containing two red balls, one blue ball and one green ball. What is the probability that you have grabbed two non-red balls given that you have grabbed at least one non-red ball? What is the probability that you have grabbed two non-red balls given that you have grabbed the green ball? Can you give an intuitive explanation of why the second probability is larger than the first one?

Problem 8.6 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?

Problem 8.7 Suppose a bridge player's hand of thirteen cards contains an ace. What is the probability that the player has at least one more ace? What is the answer to this question if you know that the player had the ace of hearts? Can you explain why the second probability is larger than the first one?

8.1.1 Assigning probabilities by conditioning

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. The result $P(AB) = P(A \mid B)P(B)$ is called the *multiplication rule* for probabilities. In many cases, this rule 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 $P(1^{\text{st}}$ marble is red$) = \frac{7}{10}$ and $P(2^{\text{nd}}$ marble is white $\mid 1^{\text{st}}$ 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 for 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. More generally, by a repeated application of the basic formula $P(A) = P(A \mid B)P(B)$, we have the following useful rule.

Rule 8.1 *For any sequence of events* $A_1, A_2, \ldots, A_n$,

$$\begin{aligned} &P(A_1 A_2 \cdots A_n) \\ &= P(A_1) \times P(A_2 \mid A_1) \times P(A_3 \mid A_1 A_2) \times \cdots \times P(A_n \mid A_1 A_2 \cdots A_{n-1}). \end{aligned}$$

Many probability problems with a solution by counting arguments can be more easily tackled by using conditional probabilities. To illustrate this, consider question (a) from Example 7.9. For $i = 1, \ldots, 4$, let A_i be the event that two teams from the same country play against each other in the ith match. The four matches are sequentially determined by lot. The probability asked in question (a) is then given by $P(A_1 A_2 A_3 A_4)$. Obviously, $P(A_1) = \frac{1}{7}$,

$P(A_2 \mid A_1) = \frac{1}{5}$, $P(A_3 \mid A_1,\ A_2) = \frac{1}{3}$ and $P(A_4 \mid A_1,\ A_2,\ A_3) = 1$ and so the desired probability $P(A_1A_2A_3A_4)$ equals $\frac{1}{7} \times \frac{1}{5} \times \frac{1}{3} \times 1 = 0.0095$. As another illustration, what is the probability that it takes 10 or more cards before the first ace appears if cards are randomly drawn one by one from an ordinary deck of 52 cards? To answer this question by counting arguments, we take as sample space the set of all possible permutations of the integers $1, 2, \ldots, 52$. Each of the 52! elements is equally likely. Let A be the event that it takes 10 or more cards before the first ace appears. The set A contains $48 \times 47 \times \cdots \times 40 \times 41!$ elements of the sample space and so $P(A) = 48 \times 47 \times \cdots \times 40 \times 41!/52! = 0.4559$. To answer the question with conditional probabilities, define A_i as the event that the ith card drawn is not an ace. Applying Rule 8.1, we find $P(A_1A_2 \cdots A_9) = \frac{48}{52} \times \frac{47}{51} \times \cdots \times \frac{40}{44} = 0.4559$.

The following important remark is made. In using conditional probabilities, you usually perform the probability calculations without explicitly specifying a sample space; an assignment of probabilities to properly chosen events suffices. Solving a probability problem by counting arguments always requires the specification of a sample space.

Example 8.3 A group of fifteen tourists is stranded in a city with four hotels of the same class. Each of the hotels has enough room available to accommodate the fifteen 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

$$\sum_{i_A=1}^{12} \sum_{i_B=1}^{13-i_A} \sum_{i_C=1}^{14-i_A-i_B} \frac{1}{16} \times \frac{1}{16 - i_A} \times \frac{1}{16 - i_A - i_B} = 0.2856.$$

The next example deals with a famous problem known as the *gambler's ruin problem*.

Example 8.4 John and Pete enter a coin-tossing game and they play until one of them has won all of the money. John starts with a dollars and Pete with b dollars. They flip a fair coin. If the coin lands heads, John gets one dollar from Pete; otherwise, John loses one dollar to Pete. What is the probability that John will win all of the money? Next consider the situation that ten people including John and Pete enter the coin-tossing game. Each player has the same starting capital. The first two players, selected by lot, will play until one wins all the other's money. The winner having doubled his money will next play against another player until one has won all the other's money, and so on. Suppose that John is in the first game and Pete plays the last game against the survivor of the first nine players. Who has the best chance to be the ultimate winner – John or Pete?

Solution. Let $P(a, b)$ denote the probability of John winning all of the money when John starts with a dollars and Pete with b dollars. It will be proved that

$$P(a, b) = \frac{a}{a+b}.$$

Let E be the event that John wins all of the money. A recurrence equation for $P(E) = P(a, b)$ is obtained by conditioning on the outcome of the first flip. Let H be the event that the first flip lands heads and let $\overline{H}$ be the event that the first flip lands tails. The event E is the union of the disjoint events EH and $E\overline{H}$ and so $P(E) = P(EH) + P(E\overline{H})$. Applying the formula $P(AB) = P(A \mid B)P(B)$, it next follows that

$$P(E) = P(E \mid H)P(H) + P(E \mid \overline{H})P(\overline{H}).$$

If the first flip lands heads, you get the changed situation that John has $a+1$ dollars and Pete has $b-1$ dollars. This gives $P(E \mid H) = P(a+1, b-1)$. Similarly, $P(E \mid \overline{H}) = P(a-1, b+1)$. The coin is fair and so $P(H) = P(\overline{H}) = \frac{1}{2}$. Thus we obtain the recurrence equation

$$P(a, b) = \frac{1}{2}P(a+1, b-1) + \frac{1}{2}P(a-1, b+1).$$

Copying this equation with the starting capitals a and b replaced by i and $a+b-i$ gives a system of equations for $P_i = P(i, a+b-i)$. These equations are

$$P_i = \frac{1}{2}P_{i+1} + \frac{1}{2}P_{i-1} \qquad \text{for } i = 1, \ldots, a+b-1$$

with the boundary conditions $P_0 = 0$ and $P_{a+b} = 1$. The solution can be found through the standard theory for linear difference equations. However, one could also make the educated guess that the win probability P_i must be proportional to the capital i. Substituting a solution of the form $P_i = ci$ for some constant c into the equations and using the boundary conditions, it follows that $P_i = \frac{i}{a+b}$ for $i = 0, 1, \ldots, a+b$. This gives the desired result $P(a,b) = \frac{a}{a+b}$.[†]

The answer to the second question is that each of the ten players has the same probability of $\frac{1}{10}$ of being the ultimate winner. This answer might surprise you more than the answer to the first question. Let us first argue that the probability of John being the ultimate winner is $\frac{1}{10}$. Define A_i as the event that John enters the ith round and wins this round. Assuming that each player has a starting capital of a dollars, we then have $P(A_1) = \frac{a}{a+a} = \frac{1}{2}$, $P(A_2 \mid A_1) = \frac{2a}{2a+a} = \frac{2}{3}$, $P(A_3 \mid A_1,\ A_2) = \frac{3a}{3a+a} = \frac{3}{4}$, and so on. This gives

$$P(A_1 A_2 \cdots A_9) = \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \times \cdots \times \frac{9}{10} = \frac{1}{10},$$

verifying that John will be the ultimate winner with probability $\frac{1}{10}$. By similar arguments, the player who begins with the second round will be the ultimate winner with probability $\frac{1}{3} \times \frac{3}{4} \times \cdots \times \frac{9}{10} = \frac{1}{10}$. Following this reasoning through, we find the same win probability of $\frac{1}{10}$ for each player.

Problem 8.8 Solve the Problems 7.27, 7.34, 7.35 and 7.41 by using conditional probabilities.

Problem 8.9 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 can join the line any time, what position in the line maximizes your probability of getting the free ticket?

Problem 8.10 You travel from Amsterdam to Sydney 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 Sydney with you? If your luggage does not reach Sydney with you, what is the probability that it was lost at the airport of Dubai?

[†] For the case of an unfair coin landing heads with probability p and landing tails with probability $q = 1 - p$, it can be shown along the same lines that the probability of John winning all of the money is given by $P(a,b) = [1 - (q/p)^a]/[1 - (q/p)^{a+b}]$.

Problem 8.11 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 in a random order, one by one. What is the probability of all seven individuals taking their seats without having to squeeze past an already seated individual?

Problem 8.12 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 $i = 1, \ldots, 4$?

Problem 8.13 A box contains N identical balls, where one ball has a winning number written on it. In a prespecified order N persons each choose a ball at random from the box. Show that each person has the same probability of getting the winning ball.

Problem 8.14 What is the probability that in the next m draws of the 5/39 Lotto there will be three consecutive draws with the feature that the same five numbers appear in two or more of these three draws? In each draw five distinct numbers are drawn at random from $1, 2, \ldots, 39$. *Hint*: define A_i as the event that each of the three draws $i, i + 1$ and $i + 2$ has a different set of five numbers.

Problem 8.15 You repeatedly roll a fair die. What is the probability that the outcome of a previous roll repeats itself before the fifth roll of the die?

Problem 8.16 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 to 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?

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 8.5 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 8.5. In addition to the events A and B from Example 8.5, 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 a detail as in Example 8.5, 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 8.17 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?

8.2 Law of conditional probability

The law of conditional probability is a natural partitioning rule. Let us consider the following experiment. A fair die is rolled to yield a number between 1 and 6, and then a fair coin is tossed that many times. Many people would reason as follows to find the probability that heads will not appear. If the number k is rolled, then the probability of no heads is $(\frac{1}{2})^k$. Hence the desired probability is

$$\frac{1}{2} \times \frac{1}{6} + \frac{1}{4} \times \frac{1}{6} + \frac{1}{8} \times \frac{1}{6} + \frac{1}{16} \times \frac{1}{6} + \frac{1}{32} \times \frac{1}{6} + \frac{1}{64} \times \frac{1}{6} = 0.1641.$$

This intuitive reasoning is justified by the following rule.

Rule 8.2 *Let A be an event that can only occur if 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 probability*. The proof of this law is simple and instructive. In terms of sets, 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 probability. 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.

It is often the case that the unconditional probability $P(A)$ of an event A is found most easily by by conditioning on an appropriate sequence of *mutually exclusive* events $B_1, \ldots, B_n$. A nice illustration of the law of conditional probability is provided by the craps example in Section 3.3 of Chapter 3. Another nice illustrative example is the following one.

Example 8.6 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 their 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. Using the standard trick of calculating the complementary probability that each person has a different birthday, we find

$$P(A \mid B_i) = \begin{cases} 1 - \frac{23 \times 22 \times \cdots \times (23-i+1)}{23^i}, & 2 \le i \le 23 \\ 1, & i \ge 24. \end{cases}$$

Obviously, $P(A \mid B_0) = P(A \mid B_1) = 0$. The probability $P(B_i)$ is given by

$$P(B_i) = \binom{180}{i} \left(\frac{23}{365}\right)^i \left(1 - \frac{23}{365}\right)^{180-i}, \qquad 0 \le i \le 180.$$

Putting the pieces together, we obtain from $P(A) = \sum_{i=2}^{180} P(A \mid B_i) P(B_i)$ that

$$P(A) = 1 - P(B_0) - P(B_1) - \sum_{i=2}^{23} \frac{23 \times 22 \times \cdots \times (23 - i + 1)}{23^i} P(B_i).$$

This yields the value 0.8841 for the probability $P(A)$.

The following example uses conditional probabilities to solve a real-life problem.

Example 8.7 Two candidates A and B remain in the final of a television game show. At this point, each candidate must spin a wheel of fortune. The numbers $1, 2, \ldots, 100$ are listed on the wheel and when the wheel has stopped spinning, a pointer randomly stops on one of the numbers. Each player 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. A player whose total exceeds 100 gets a final score of zero. The winner is the player who gets the highest score. Should both players have the same final score, then the player to spin the wheel first is the winner. Player A has to spin first. What is the optimal strategy for player A?

Solution. The optimal strategy of the second player B is obvious. This player stops after the first spin only if the score is larger than the final score of player A. To find the optimal strategy of player A, let $S(a)$ be the probability

of player A winning if A stops after the first spin giving a score of a points and $C(a)$ the probability of player A winning if A continues after the first spin giving a score of a points. Let us take a general set-up and assume that the numbers $1, 2, \ldots, R$ are listed on the wheel. To find $S(a)$, we condition on the outcome of the first spin of player B. Given that player B has scored b points in the first spin with $b \leq a$, then player A is the final winner if player B scores in the second spin either at most $a - b$ points or more than $R - b$ points. By the law of conditional probability, we find

$$S(a) = \frac{1}{R} \sum_{b=1}^{a} \left(\frac{a-b}{R} + \frac{b}{R} \right) = \frac{a^2}{R^2} \quad \text{for } 1 \leq a \leq R.$$

The expression for $C(a)$ follows directly from $S(a)$. By conditioning on the outcome of the second spin of player A, we have

$$C(a) = \frac{1}{R} \sum_{k=1}^{R-a} S(a+k) = \frac{1}{R} \sum_{k=1}^{R-a} \frac{(a+k)^2}{R^2} \quad \text{for } 1 \leq a \leq R.$$

Taking $R = 100$, numerical calculations give that $C(a) > S(a)$ for $1 \leq a \leq 53$ and $S(a) > C(a)$ for $53 < a \leq 100$. Hence, player A stops after the first spin if this spin gives a score larger than 53; otherwise, A continues. The probability of player A winning is $\frac{1}{100} \sum_{a=1}^{53} C(a) + \frac{1}{100} \sum_{a=54}^{100} S(a) = 0.4596$.

Problem 8.18 You have two identical boxes in front of you. One of the boxes contains 10 balls numbered 1 to 10 and the other box contains 25 balls numbered 1 to 25. You choose at random one of the boxes and pick a ball at random from the chosen box. What is the probability of picking the ball with number 7 written on it?

Problem 8.19 A bag contains three coins. One coin is two-headed and the other two are normal. A coin is chosen at random from the bag and is tossed twice? What is the probability of getting two heads? If two heads appear, what is the inverse probability that the two-headed coin was chosen?

Problem 8.20 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 of unlocking the box with the sports car?

Problem 8.21 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 fifty-fifty 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 8.22 You play a series of independent games against an opponent. You win any game with probability p and lose it with probability $q = 1 - p$. The ultimate winner is the first player to achieve a lead of two games. What is the probability that you will be the ultimate winner? *Hint*: condition on the outcomes of the first two games and derive an equation for the unknown probability.

Problem 8.23 You repeatedly roll two fair dice. What is the probability of two consecutive totals of 7 appearing before a roll with double sixes?

Problem 8.24 Let's return to the casino game Red Dog from Problem 3.25. Use the law of conditional probability to calculate the probability of the player winning. *Hint*: argue first that the probability of a spread of i points is given by $(12 - i) \times 4 \times 4 \times 2 \times 50!/52!$.

Problem 8.25 Consider again Problem 2.19. Use conditional probabilities to obtain the probability of guessing all six balls before running out of tokens. *Hint*: define $p(i, t)$ as your success probability once you have reached cup i with t tokens left. Use a conditioning argument to obtain a recursion equation for the $p(i, t)$ and calculate $p(1, 6)$.

Problem 8.26 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 probability to find a recursion equation for the 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 8.27 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 probability to obtain a recursion equation for the a_k. Calculate a_k for $k = 5$, 10, 25, and 50.

Problem 8.28 Consider Problem 2.39 again. What is the probability that the kth dwarf can sleep in his own bed for $k = 1, \ldots, 7$? *Hint*: use a general set-up

and define $p(k, n)$ as the probability that the kth dwarf will not sleep in his own bed for the situation of n dwarfs with the first dwarf being drunk. Use a conditioning argument to obtain a recursion equation for the $p(k, n)$.

Problem 8.29 (difficult) Consider the three-player variant of Example 8.7. What is the optimal strategy for the first player A and what is the overall win probability of player A? Next use the probability distribution function of the final score of player A to find the overall win probabilities of the players B and C.

8.3 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 already 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 general rule is that odds a to b correspond to a probability of $\frac{a}{a+b}$. The odds form of Bayes' rule reads as follows:

Rule 8.3 *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.

[†] 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.

In applying Bayes' rule in odds form, the main step is identifying the hypothesis H and the evidence E. We illustrate this with several examples.

Example 8.8 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?

Solution. Obviously, the hypothesis H is that the wreck is in the area in question and the evidence E is that the wreck has not been detected in that area. By Bayes' rule in odds form,

$$\frac{P(H \mid E)}{P(\overline{H} \mid E)} = \frac{0.4}{0.6} \times \frac{0.1}{1} = \frac{1}{15}.$$

The updated probability that the wreck is in the area is given by $\frac{1}{1+15} = \frac{1}{16}$.

Example 8.9 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?

Solution. We can without restriction assume that the color on the visible side of the chosen card is red. Let H be the hypothesis that both sides of the chosen card are red and let E be the evidence that the visible side of the chosen card is red. Then, $P(H) = \frac{1}{3}$, $P(\overline{H}) = \frac{2}{3}$, $P(E \mid H) = 1$ and $P(E \mid \overline{H}) = \frac{1}{2} \times 0 + \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Hence

$$\frac{P(H \mid E)}{P(\overline{H} \mid E)} = \frac{1/3}{2/3} \times \frac{1}{1/4} = 2.$$

This gives $P(H \mid E) = \frac{2}{3}$. Hence the bet is unfavorable to you.

Example 8.10 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 $P(H \mid E)/[1 - P(H \mid E)] = 10$, 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 8.11 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 as well 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

[†] 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.

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 probability. In this example, Bayes' formula is given by

$$\frac{P(G \mid EM)}{P\left(\overline{G} \mid EM\right)} = \frac{P(G \mid M)}{P\left(\overline{G} \mid M\right)} \frac{P(E \mid GM)}{P\left(E \mid \overline{G}M\right)},$$

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.

The next example involves a subtle application of Bayes' rule in odds form.

Example 8.12 A diamond merchant has lost a case containing 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, and each is considered a suspect in the diamond theft. 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 perpetrators 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 generally quite 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 has matching DNA, 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 reasoning of the defence is on the whole correct and can be made more precise by Bayes' rule. Let us perform the analysis under the assumption that each of the 150 thousand inhabitants of the city is equally likely to be the finder of the diamond and that the finder keeps the diamond with probability p_0 when the finder is a person in the database and with probability p_1 otherwise. It will appear that only the value of the ratio $r = p_0/p_1$ is required and thus not the values of p_0 and p_1. The following formula will be derived:

$$P(\text{suspect is the thief}) = \frac{r}{r + (n - n_0)\lambda},$$

where $n = 150{,}000$, $n_0 = 5{,}120$ and $\lambda = 10^{-6}$. This posterior probability has the value 0.8735 for $r = 1$ and the value 0.9325 for $r = 2$. This result confirms that the counsel for defense is correct in stating that it is not beyond reasonable doubt that the suspect is guilty so that the suspect must be released if there is no additional proof. The derivation of the above formula proceeds as follows. Define H as the hypothesis that the thief is a person from the database. Let E_1 be the evidence that the thief has the observed DNA profile and let E_2 be the evidence that only one person in the database has the observed DNA profile. The subtlety of the Bayesian analysis lies in the specifications of H, E_1 and E_2. With this notation the probability that the suspect is the thief is given by $P(H \mid E_1, E_2)$ (verify!). The prior odds of H is

$$\frac{P(H)}{P(\overline{H})} = \frac{(n_0/n)p_0}{[(n - n_0)/n]p_1} = r\frac{n_0}{n - n_0}.$$

Using the fact that $P(E_1 \mid H) = P(E_1 \mid \overline{H}) = \lambda$, it follows from Bayes' rule in odds form that $P(H \mid E_1)/P(\overline{H} \mid E_1) = P(H)/P(\overline{H})$. In other words, the prior odds of hypothesis H does not change based on the evidence E_1. Next we use evidence E_2. The relations

$$P(E_2 \mid H,\ E_1) = (1 - \lambda)^{n_0 - 1} \quad \text{and} \quad P(E_2 \mid \overline{H},\ E_1) = n_0\lambda(1 - \lambda)^{n_0 - 1}$$

hold. Using the above result $P(H \mid E_1)/P(\overline{H} \mid E_1) = P(H)/P(\overline{H})$, we obtain from another application of Bayes' rule in odds form that

$$\frac{P(H \mid E_1,\ E_2)}{P(\overline{H} \mid E_1,\ E_2)} = \frac{P(H \mid E_1)}{P(\overline{H} \mid E_1)} \times \frac{P(E_2 \mid H,\ E_1)}{P(E_2 \mid \overline{H},\ E_1)} = r\frac{n_0}{n - n_0} \times \frac{1}{n_0\lambda}.$$

This gives $P(H \mid E_1,\ E_2) = r/(r + (n - n_0)\lambda)$, as was to be shown.

Problem 8.30 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 8.31 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 8.32 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.

Problem 8.33 You choose a family at random from all two-child families having at least one boy. What is the probability that the chosen family has two boys? What is the answer to this question if the chosen family has at least one boy born on a Sunday ? What is the probability that the chosen family has two boys if the family has at least one boy born on one of the first k days of the week ($k = 1, \ldots, 7$)? Assume that each child is a boy or a girl with equal probability. *Remark*: The wording of the boy–girl problem is of utmost importance. In some formulations of the problem, the boy–girl problem becomes ambiguous and leaves room for different interpretations.

Problem 8.34 Your friend shakes thoroughly two dice in a dice-box. He then looks into the dice-box. Your friend is a honest person and always tells you if he sees a six in which case he bets with even odds that both dice show an even number. Is this a fair bet?

Problem 8.35 Suppose a person has confessed a crime under a given amount of police pressure. Besides this confession there is no other proof that the person

is guilty. Use Bayes' rule in odds form to verify that the confession only adds to evidence of guilt if the confession is more likely to come from the guilty than from the innocent. Do you think that in real life a hardened person who is guilty is more likely to confess than an unstable person who is innocent?

Problem 8.36 Use Bayes' rule in odds form to analyze the Monty Hall dilemma from Section 6.1 and the test paradox from Section 6.2. Also, solve the Problems 6.7, 6.11 and 6.13 by using Bayes' rule in odds form.

8.4 Bayesian statistics – discrete case

Bayesian statistics takes a fundamentally different viewpoint from that of classical statistics. Unlike the frequentist approach, the Bayesian approach to statistical inference treats population parameters not as fixed, unknown constants but as random variables – subject to change as additional data arise. Probability distributions have to be assigned to these parameters by the investigator before observing data. These distributions are called prior distributions and are inherently subjective. They represent the investigator's uncertainty about the true value of the unknown parameters. Assigning probabilities by degree of belief is consistent with the idea of a fixed but unknown value of the parameter. Treating the unknown parameter as a random variable does not mean that we believe the parameter is random. Rather, it expresses the state of our knowledge about the parameter.

The Bayesian approach is a scientific method in which you state a hypothesis by means of a prior probability distribution, collect and summarize relevant data, and revise your beliefs about the population parameters by learning from data that were collected. The revised distribution of the unknown parameter is called the posterior probability distribution. It is obtained by the use of Bayes' rule and reflects your new beliefs about the population parameters. By treating the unknown parameters as random variables, you are able to make direct statements about the probability that the hypothesis is true. In applications, one often wants to know this probability rather than the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the hypothesis is true. This tail probability is known as the p-value in classical statistics and its correct interpretation is misunderstood by many students and practitioners. The Bayesian posterior probability, on the other hand, cannot be misunderstood and is directly interpretable as the probability that the hypothesis is true. 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. 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 the light of these conclusions. This fact is of particular relevance in the case of medical applications.

Bayesian statistics is often used to update the degree of belief in the effectiveness of medical treatments given new clinical data and to predict election results based on the results of new opinion polls. This will be illustrated with two examples below. In these examples the Bayesian approach produces answers that are more logical and easier to understand than those produced by the frequentist approach of classical statistics. Bayesian statistics has many other applications. For example, it 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.

Example 8.13 A new treatment is tried out for a disease for which the historical success rate is 35%. The discrete uniform distribution on $0, 0.01, \ldots, 0.99$, and 1 is taken as prior for the success probability of the new treatment. The experimental design is to make exactly ten observations by treating ten patients. The experimental study yields seven successes and three failures. What is the posterior probability that the new treatment is more effective than the standard treatment?[†]

Solution. Model the unknown success probability of the new treatment by the random variable Θ. Assume that our state of knowledge about the unknown parameter is expressed by the "non-informative" prior distribution $p_0(\theta) = \frac{1}{101}$ for $\theta = 0, 0.01, \ldots, 0.99, 1$. To update the prior distribution given the observed data, we need the so-called likelihood function $L(\text{data} \mid \theta)$. This function is defined as the probability of getting the data given that the success probability has the value θ. In the particular situation of seven successes in the treatment of ten patients,

$$L(\text{data} \mid \theta) = \binom{10}{7}\theta^7(1-\theta)^3 \quad \text{for } \theta = 0, 0.01, \ldots, 0.99, 1.$$

To find the posterior probability $p(\theta) = P(\Theta = \theta \mid \text{data})$, we use Bayes' rule

$$P(\Theta = \theta \mid \text{data}) = \frac{P(\text{data} \mid \theta)p_0(\theta)}{P(\text{data})},$$

[†] This example is based on the paper by D.A. Berry, "Bayesian clinical trials," *Nature Reviews Drug Discovery* **5** (2006): 27–36.

where $P(\text{data}) = \sum_\theta P(\text{data} \mid \theta) p_0(\theta)$. Hence the posterior is given by

$$p(\theta) = \frac{L(\text{data} \mid \theta) p_0(\theta)}{\sum_\theta L(\text{data} \mid \theta) p_0(\theta)} \quad \text{for } \theta = 0, 0.01, \ldots, 0.99, 1.$$

Note that the *posterior* is proportional to *prior* $\times$ *likelihood*. The prior and the likelihood are only needed up to a multiplicative constant. Letting $\theta_i = \frac{i}{100}$ for $0 \le i \le 100$ and inserting the expressions for $L(\text{data} \mid \theta)$ and $p_0(\theta)$, we obtain the result

$$p(\theta_i) = \frac{\theta_i^7 (1 - \theta_i)^3}{\sum_{k=0}^{100} \theta_k^7 (1 - \theta_k)^3} \quad \text{for } i = 0, 1, \ldots, 100.$$

In particular, the posterior probability of the new treatment being more effective than the standard treatment is given by

$$\sum_{i=36}^{100} p(\theta_i) = 0.9866.$$

Incidentally, the posterior probability that the new treatment is not more effective than the standard treatment is 0.0134. This value is not very different from the value 0.0260 of the excess probability of obtaining seven or more successes in ten trials under the hypothesis that the new treatment causes no difference. In classical statistics, this p-value would have been calculated and on the basis of its small value the hypothesis that the new treatment causes no difference would have been rejected. The p-value is often interpreted incorrectly, whereas the Bayesian posterior probability is directly interpretable as the probability that the new treatment is effective. However, it should be said that this posterior probability depends on the prior distribution. Assuming a "non-informative" prior such as the uniform distribution, the results of the trials carry essentially all the influence in the posterior distribution. In practice, it happens in many instances that the prior can be based on earlier experimental studies or scientific theory. The Bayesian approach has the advantage that you can continuously update your beliefs as information accrues. In Problem 8.38 you are asked to update your beliefs after each trial, using the posterior obtained after the preceding trials as prior. This gives the same result as when the update takes place for the whole batch of observations. The Bayesian approach gives the user much flexibility in the experimental design. Whereas the frequentist approach requires trials to reach a prespecified sample size before stopping, the Bayesian approach allows the user to stop trials early when adding more patients will not appreciably change the conclusion.

Example 8.14 Two candidates A and B are contesting the election for 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. Let us model the fraction of the voting population in favor of candidate A by the prior distribution:

$$p_0(\theta) = \begin{cases} \frac{\theta - 0.29}{4.41} & \text{for } \theta = 0.30, \ldots, 0.50, \\ \frac{0.71 - \theta}{4.41} & \text{for } \theta = 0.51, \ldots, 0.70. \end{cases}$$

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. Using the likelihood $L(\text{data} \mid \theta) = \binom{1,000}{517}\theta^{517}(1-\theta)^{483}$, this posterior probability is calculated from

$$p(\theta) = \frac{\theta^{517}(1-\theta)^{483} p_0(\theta)}{\sum_{a=30}^{70} \left(\frac{a}{100}\right)^{517} \left(1 - \frac{a}{100}\right)^{483} p_0\left(\frac{a}{100}\right)}.$$

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 $p(0.50) = 0.1558$.

Problem 8.37 You wonder who is the better player of the tennis players Alassi and Bicker. Your prior assigns equal probabilities to the three possible values 0.4, 0.5 and 0.6 for the probability that Alassi wins a given match. Then you learn about a tournament at which a best-of-five series of matches is played between Alassi and Bicker over a number of days. In such an encounter the first player to win three matches is the overall winner. It turns out that Alassi wins the best-of-five contest. How should you update your prior? *Remark*: this problem is taken from the paper "Teaching Bayes' rule: a data-oriented approach," by J. Albert, in *The American Statistician* **51** (1997): 247–253.

Problem 8.38 Consider Example 8.13 again. Assume that the ten observations are $SSFSSFSSSF$, where S stands for success and F for failure. Update the

posterior probability of the new treatment being more effective than the standard treatment after each observation in order to see how this probability changes when additional information becomes available. What is the probability that the new treatment will be successful for the 11th patient? Also, verify that the same posterior probabilities would have been obtained if the experimental design had been to treat patients until facing the third failure and the third failure would have occurred after having treated ten patients.

Problem 8.39 A student has passed a final exam by supplying correct answers for 30 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 when you have the following prior information? There are three ways the student may have done the exam: totally unprepared, half prepared and well prepared. The three possibilities have the respective prior probabilities 0.2, 0.3 and 0.5. Any of the 50 questions is correctly answered by the student with probability 1/3 if he is totally unprepared, with probability 0.45 if he is half prepared, and with probability 0.8 if he is well prepared.

9

Basic rules for discrete 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*. Intuitively, a random variable can be thought of as a quantity whose value is not fixed. The value of a random variable is determined by the outcome of the experiment and consequently probabilities can be assigned to the possible values of the 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. In addition, we give rules for the expected value and the variance of a sum of random variables, including the square-root rule. The rules for random variables are easiest explained and understood in the context of discrete random variables. These random variables can take on only a finite or countably infinite number of values (the so-called continuous random variables that can take on a continuum of values are treated in the next chapter). To conclude this chapter, we discuss the most important discrete random variables such the binomial, the Poisson and the hypergeometric random variables.

9.1 Random variables

The concept of random variable is always a difficult concept for the beginner. Intuitively, a random variable is a function that takes on its values by chance. A random variable is not a variable in the traditional sense of the word and actually it is a little misleading to call it a variable. The convention is to use capital letters such as X, Y, and 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 the random variable X is defined as the smallest of the two numbers rolled in the experiment of rolling a fair die twice, then the random variable X assigns the numerical value $\min(i, j)$ 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 is performed, we can only describe the set of possible values of X. Illustrative examples of random variables are:

- the number of winners in a football pool next week
- the number of major hurricanes that will hit the United States next year
- the daily number of claims submitted to an insurance company
- the amount of rainfall that the city of London will receive next year
- the lifetime of a newly bought battery
- the duration of a telephone call.

The first three examples are examples of discrete random variables taking on a discrete number of values and the other three examples describe continuous random variables taking on a continuum of values.

In this chapter we consider only discrete random variables. 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 probabilities associated with these possible values are determined by the probability measure P on the sample space of the chance experiment. 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

$$P(X = x) = P(\{\omega : X(\omega) = x\}).$$

In words, $P(X = x)$ is the probability mass assigned by the probability measure P to the set of all outcomes ω for which $X(\omega) = x$. For example, consider the experiment of rolling a fair die twice and let the random variable X be defined as the smallest of the two numbers rolled. The range of X is the set $I = \{1, 2, \ldots, 6\}$. The random variable X takes on the value 1 if one of the eleven outcomes (1, 1), (1, 2), ..., (1, 6), (2, 1), (3, 1),..., (6, 1) occurs and so $P(X = 1) = \frac{11}{36}$. In the same way, $P(X = 2) = \frac{9}{36}$, $P(X = 3) = \frac{7}{36}$, $P(X = 4) = \frac{5}{36}$, $P(X = 5) = \frac{3}{36}$, and $P(X = 6) = \frac{1}{36}$.

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 $P(X = 50) = \frac{1}{10}$.

Example 9.2 You have a well-shuffled ordinary deck of 52 cards. You remove the cards one at a time until you get an ace. Let the random variable X be the number of cards removed. What is the probability mass function of X?

Solution. The range of the random variable X is the set $\{1, 2, \ldots, 49\}$. Obviously, $P(X = 1) = \frac{4}{52}$. Using Rule 8.1, we find for $i = 2, \ldots, 49$:

$$P(X = i) = \frac{48}{52} \times \cdots \times \frac{48 - (i - 2)}{52 - (i - 2)} \times \frac{4}{52 - (i - 1)}.$$

The latter example gives rise to the following important observation. Often an explicit listing of the sample space is not necessary to assign a probability distribution to a random variable. Usually the probability distribution of a random variable is modeled without worrying about the assignment of probability to an underlying sample space. 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 A fair die is tossed two times. Let the random variable X be the largest of the two outcomes. What is the probability mass function of X?

Problem 9.2 Imagine that people enter a room one by one and announce their birthdays. Let the random variable X be the number of people required to have a matching birthday. What is the probability mass function of X?

Problem 9.3 A bag contains three coins. One coin is two-headed and the other two are normal. A coin is chosen at random from the bag and is tossed twice. Let the random variable X denote the number of heads obtained. What is the probability mass function of X?

Problem 9.4 A fair die is rolled until each of the six possible outcomes has occurred. Let the random variable X be the number of rolls required. What is the probability mass function of X? *Hint*: use the answer to Problem 7.49 and the relation $P(X = i) = P(X > i - 1) - P(X > i)$.

9.2 Expected value

The most important characteristic of a random variable is its *expected value*. 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.

Definition 9.1 *The expected value of the discrete random variable X having I as its set of possible values is defined by*

$$E(X) = \sum_{x \in I} x\, P(X = x).$$

To illustrate this definition, consider again the experiment of rolling a fair die twice and let the random variable X denote the smallest of the two numbers rolled. Then,

$$E(X) = 1 \times \frac{11}{36} + 2 \times \frac{9}{36} + 3 \times \frac{7}{36} + 4 \times \frac{5}{36} + 5 \times \frac{3}{36} + 6 \times \frac{1}{36} = 2.528.$$

Before we give some other 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 $1 - 1 + 1 - 1 + \cdots$ has the sum 0 when you sum the terms according to $(1 - 1) + (1 - 1) + \cdots$, 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 $k = \pm 1, \pm 2, \ldots$ (by the celebrated result $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$, the probabilities sum to 1). The reason that $E(X)$ does not exist is the well-known fact from calculus that $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$.

Example 9.1 (continued) What is 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 9.3 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 payoff to 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 9.4 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 $\sum_{j=1}^{\infty} j x^{j-1} = 1/(1 - x)^2$ for all $0 < x < 1$ (see the Appendix). Hence the expected value of the number of trials required is $\frac{4}{3}$.

The expected value of a random variable X is also known as *expectation*, or *mean*, or *first moment* and *average value* of X. The term "average value" can be explained as follows. Consider Example 9.3. If the game is played many times in succession, then the average profit per game will approach $E(X)$ as the number of repetitions of the game increases without bound. This result is known as the law of large numbers. This law will be made more precise in Section 14.4 (see also the informal discussion in Section 2.3).

Problem 9.5 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 \$15 if two dice show the same even number and the other two dice show the same odd number. Is this a favorable game? Answer this same question for the following game in which the stake is \$2. A fair coin is tossed no more than five times. The game ends if the coin comes up tails or five straight heads appear, whichever happens first. You get a payoff of \$1 each time heads appears plus a bonus of \$25 if five heads appear in a row.

Problem 9.6 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 9.7 You throw darts at a circular target on which two concentric circles of radius 1 cm and 3 cm are drawn. The target itself has a radius of 5 cm. You receive 15 points for hitting the target inside the smaller circle, 8 points for hitting the middle annular region, and 5 points for hitting the outer annular region. The probability of hitting the target at all is 0.75. If the dart hits the target, the hitting point is a completely random point on the target. Let the

random variable X denote the number of points scored on a single throw of the dart. What is the expected value of X?

Problem 9.8 Three players, A, B and C each put ten dollars into a pot with a list on which they have 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. Suppose that players A and B 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). What the expected value of the amount that player C will receive?

Problem 9.9 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 9.10 In a charity 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 one number at random 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 and what is the expected value of the number of prize winners? What is the probability that a randomly picked contestant will be a prize winner?

Problem 9.11 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 $1 \le k \le m-1$, 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 9.12 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 9.13 Suppose that the random variable X is nonnegative and integer-valued. Verify that $E(X) = \sum_{k=0}^{\infty} P(X > k)$. Use this result to compute the expected value of the random variables from the Problems 9.2 and 9.4.

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 and are finite.

The proof is simple for the discrete case. Define the random variable Z by $Z = X + Y$. Then, for all z,

$$P(Z = z) = \sum_{x,y:\, x+y=z} P(X = x,\ Y = y),$$

where $P(X = x,\ Y = y)$ denotes the probability of the joint event that X takes on the value x and Y the value y. Hence, by $E(Z) = \sum_z z P(Z = z)$, we have $E(Z) = \sum_z z \sum_{x,y:\, x+y=z} P(X = x,\ Y = y)$. This gives

$$\begin{aligned} E(Z) &= \sum_z \sum_{x,y:\, x+y=z} (x + y) P(X = x,\ Y = y) \\ &= \sum_{x,y} (x + y) P(X = x,\ Y = y). \end{aligned}$$

Hence $E(Z) = \sum_{x,y} x P(X = x,\ Y = y) + \sum_{x,y} y P(X = x,\ Y = y)$. This expression can be rewritten as

$$E(Z) = \sum_x x \sum_y P(X = x,\ Y = y) + \sum_y y \sum_x P(X = x,\ Y = y).$$

Observe that, by Rule 7.1, the relations $\sum_y P(X = x,\ Y = y) = P(X = x)$ and $\sum_x P(X = x,\ Y = y) = P(Y = y)$ hold. This leads to

$$E(Z) = \sum_x x P(X = x) + \sum_y y P(Y = y) = E(X) + E(Y),$$

as was to be proved. It is pointed out that the various manipulations with the summations in the above analysis are justified by the assumption that $\sum_x |x| P(X = x)$ and $\sum_y |y| P(Y = y)$ are finite.

Rule 9.1 applies to any finite number of random variables $X_1, \ldots, X_n$. If $E(X_i)$ exists and is finite for all $i = 1, \ldots, n$, then a repeated application of Rule 9.1 gives

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n).$$

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. 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 so-called *indicator random variables*. These random variables can take on only the values 0 and 1.

Example 9.5 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 to accompany you further along the line until you encounter yet again a taller child or reach the end of the line, etc. Let the random variable X denote the number of children to be selected from the line. What is the expected value of X?

Solution. We can most easily compute $E(X)$ by writing

$$X = X_1 + \cdots + X_n,$$

where the indicator variable X_i is defined as

$$X_i = \begin{cases} 1 & \text{if the } i\text{th child is selected from the line} \\ 0 & \text{otherwise.} \end{cases}$$

The probability that the ith child is the tallest among the first i children equals $\frac{(i-1)!}{i!} = \frac{1}{i}$. Hence,

$$E(X_i) = 0 \times (1 - \frac{1}{i}) + 1 \times \frac{1}{i} = \frac{1}{i}, \qquad i = 1, \ldots, n.$$

This gives

$$E(X) = 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

An insightful approximation can be given to this expected value. It is known from calculus that $1 + \frac{1}{2} + \cdots + \frac{1}{n}$ can very accurately be approximated by $\ln(n) + \gamma + \frac{1}{2n}$, where $\gamma = 0.57722\ldots$ is Euler's constant.

Example 9.6 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, $P(X_i = 1) = \frac{3}{51}$ and so $E(X_i) = \frac{3}{51}$ 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 $E(X_1 + \ldots + X_{51}) = 51 \times \frac{3}{51} = 3$.

The linearity property for general random variables

The linearity property of the expected value holds for any type of random variables including continuous random variables. A continuous random variable such as the decay time of a radioactive particle can take on a continuum 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 both a discrete component and a continuous component. Think of your delay in a queue at a counter in a supermarket or the amount reimbursed 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.

For the case of discrete random variables, a proof of the linearity property of the expected value could be given by using first principles. However, the proof is more tricky for the general case. If the random variables are continuously distributed, we need the concept of joint probability density and using this concept the linearity property can be proved, see Rule 11.2 in Chapter 11. For the general case including the case of mixed random variables, the proof requires a technical machinery from measure theory. This is beyond the scope of this book.

Problem 9.14 What is the expected value of the number of times that two adjacent letters will be the same in a random permutation of the eleven letters of the word Mississippi?

Problem 9.15 What is the expected number of distinct birthdays within a randomly formed group of 100 persons? What is the expected number of children in a class with r children sharing a birthday with some child in another class with s children? Assume that the year has 365 days and that all possible birthdays are equally likely.

Problem 9.16 Consider Problem 8.16 again. Use indicator variables to find the expected value of the number of persons having knowledge of the rumor. Next consider Problem 7.52 again. What is the expected value of the number of people who survive the first round?

Problem 9.17 What is the expected number of times that two consecutive numbers will show up in a lotto drawing of six different numbers from the numbers $1, 2, \ldots, 45$?

Problem 9.18 Verify that the expected number of cycles in a random permutation of the integers $1, \ldots, n$ is about $\ln(n) + \gamma$ for n large, where $\gamma = 0.57722\ldots$ is Euler's constant. *Hint*: for any fixed k with $1 \leq k \leq n$, define $X_i = 1$ if the integer i belongs to a cycle of length k and $X_i = 0$ otherwise. Use the fact that $\sum_{i=1}^{n} X_i/k$ is the number of cycles of length k.

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,*

$$E\,[g(X)] = \sum_{x \in I} g(x)\, P(X = x)$$

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 $z_1 = g(x_1)$, $z_2 = g(x_2)$, ... 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 $E[g(X)]$ equal to $g(E(X))$. In general, $E[g(X)] \neq g(E(X))$! 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 $\text{var}(X)$ 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 $\text{var}(X)$ 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 9.7 What is the standard deviation of the total score of a roll of 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 9.8 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,

$$\text{var}(X) = E(X) = \lambda.$$

It immediately follows from $E(X) = \sum_{n=0}^{\infty} nP(X = n)$ that

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

where the third equality uses the well-known power series $e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots$ for every real number x. Next we verify that $\text{var}(X) = \lambda$. To evaluate $E(X^2)$, use the identity $k^2 = k(k-1) + k$. This gives $E(X^2) = \sum_{k=0}^{\infty} k^2 P(X = k) = \sum_{k=1}^{\infty} k(k-1)P(X = k) + \sum_{k=1}^{\infty} kP(X = k)$ and so

$$E(X^2) = \sum_{k=1}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} + E(X) = \lambda^2 \sum_{k=2}^{\infty} e^{-\lambda}\frac{\lambda^{k-2}}{(k-2)!} + \lambda.$$

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.

Example 9.5 (continued) What is the variance of the random variable X denoting the number of children selected from the line?

Solution. We again use the representation $X = X_1 + \cdots + X_n$, where $X_i = 1$ if the ith child is the tallest among the first $i-1$ children and $X_i = 0$ otherwise. Then

$$E(X^2) = \sum_{i=1}^{n} E(X_i^2) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} E(X_i X_j).$$

By $P(X_i = 1) = \frac{1}{i}$, it follows that $E(X_i^2) = \frac{1}{i}$ for all i. To find $E(X_i X_j)$ for $i \neq j$, note that $X_i X_j$ is equal to 1 if $X_i = X_j = 1$ and 0 otherwise. Hence, $E(X_i X_j) = P(X_i = 1,\ X_j = 1)$. For any $i < j$,

$$P(X_i = 1,\ X_j = 1) = \frac{(i-1)! \times 1 \times (i+1) \times \cdots \times (j-1) \times 1}{j!} = \frac{1}{i} \times \frac{1}{j}.^{\dagger}$$

This gives

$$E(X^2) = \sum_{i=1}^{n} \frac{1}{i} + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \frac{1}{i} \times \frac{1}{j}.$$

Next, using the relation $\text{var}(X) = E(X^2) - [E(X)]^2$ and noting that $E(X) = \sum_{i=1}^{n} \frac{1}{i}$ gives $[E(X)]^2 = \sum_{i=1}^{n} \frac{1}{i^2} + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \frac{1}{i} \times \frac{1}{j}$, we find

$$\text{var}(X) = \sum_{i=1}^{n} \frac{1}{i} - \sum_{i=1}^{n} \frac{1}{i^2}.$$

Both $E(X)$ and $\text{var}(X)$ increase very slowly as n gets larger. For example, $E(X) = 14.393$ and $\text{var}(X) = 12.748$ if $n = 1{,}000{,}000$.

The next example deals with the famous newsboy problem.

Example 9.9 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

† This means that $P(X_i = 1,\ X_j = 1) = P(X_i = 1)P(X_j = 1)$ for any $i < j$. More generally, it can be verified that $P(X_{i_1} = 1, \ldots, X_{i_r} = 1) = P(X_{i_1} = 1) \cdots P(X_{i_r} = 1)$ for any $i_1 < \cdots < i_r$ and $1 \leq r \leq n$.

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

$$\text{var}[g(X)] = 7900.2475 - (87.3762)^2 = 265.647.$$

Concluding, Billy's net earnings on any given morning has an expected value of 87.38 dollars and a standard deviation of $\sqrt{265.64} = 16.30$ dollars.

Problem 9.19 The number of paramedical treatments for a particular sports injury is a random variable X with probability mass function $P(X = k) = \frac{11-k}{55}$ for $1 \leq k \leq 10$. What are the expected value and standard deviation of X? An insurance policy reimburses the costs of the treatments up to a maximum of five treatments. What are the expected value and the standard deviation of the number of reimbursed treatments?

Problem 9.20 Calculate the standard deviation of the random variables from the Problems 9.2 and 9.4.

Problem 9.21 Each of three musicians chooses seven different pieces of music at random from 22 pieces, independently of each other. They have agreed to play together any piece of music that is chosen by each of the three musicians. Let the random variable X denote the number of pieces of music that are chosen by each of the three musicians. What are the expected value and the variance of X? What is $P(X \geq 1)$? *Hint*: let the random variable X_i be equal to 1 that all three musicians choose the ith piece of music and zero otherwise.

Problem 9.22 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 9.23 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.

[†] In general, the n random variables $X_1, \ldots, X_n$ are said to be independent if they satisfy the condition $P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n)$ 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 9.6 *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.

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 and are finite.

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

$$\begin{aligned} E(Z) &= \sum_z zP(Z = z) = \sum_z z \sum_{x,y\,:\,xy=z} P(X = x,\ Y = y) \\ &= \sum_z \sum_{x,y\,:\,xy=z} xyP(X = x, Y = y) \end{aligned}$$

and so

$$\begin{aligned} E(Z) &= \sum_{x,y} xyP(X = x,\ Y = y) = \sum_{x,y} xyP(X = x)P(Y = y) \\ &= \sum_{x\in I} xP(X = x) \sum_{y\in J} yP(Y = y) = E(X)E(Y). \end{aligned}$$

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*

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y).$$

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

$$\begin{aligned} &E[X^2 + 2XY + Y^2] - \mu_X^2 - 2\mu_X\mu_Y - \mu_Y^2 \\ &= E(X^2) + 2\mu_X\mu_Y + E(Y^2) - \mu_X^2 - 2\mu_X\mu_Y - \mu_Y^2, \end{aligned}$$

using the linearity property of the expectation operator and the fact that $E(XY) = E(X)E(Y)$ for independent X and Y. This gives the desired result $\text{var}(X+Y) = E(X^2) - \mu_X^2 + E(Y^2) - \mu_Y^2 = \text{var}(X) + \text{var}(Y)$.

If $X_1, X_2, \ldots, X_n$ are independent random variables, then a repeated application of Rule 9.8 gives

$$\text{var}(X_1 + X_2 + \cdots + X_n) = \text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n).$$

This result has the following important corollary.

Rule 9.9 *If the random variables $X_1, X_2, \ldots, X_n$ are independent and have the same probability distribution with standard deviation σ, then the standard deviation of the sum $X_1 + X_2 + \cdots + X_n$ is given by*

$$\sigma(X_1 + X_2 + \cdots + X_n) = \sigma\sqrt{n}.$$

In statistics the result of Rule 9.9 is usually formulated in the equivalent form $\sigma((X_1 + X_2 + \cdots + X_n)/n) = \sigma/\sqrt{n}$, see also Rule 9.4. This formulation is known as the *square-root law*.

Problem 9.24 Let X and Y be two independent random variables. What are the expected value and the variance of $X - Y$?

Problem 9.25 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 9.26 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 $\text{var}(X) = 0$.

Problem 9.27 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 9.10 *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 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 9.10 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

$$P(X+Y=k) = \sum_{j=0}^{k} e^{-\lambda}\frac{\lambda^j}{j!}e^{-\mu}\frac{\mu^{k-j}}{(k-j)!}$$
$$= \frac{e^{-(\lambda+\mu)}}{k!}\sum_{j=0}^{k}\binom{k}{j}\lambda^j\mu^{k-j},$$

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 9.28 Modify the convolution formula in Rule 9.10 when the random variables X and Y are integer-valued but not necessarily nonnegative.

Problem 9.29 You repeatedly draw a random integer from the integers $1, \ldots, 10$ until you have three different integers. Use the convolution formula to find the probability that you need r draws.

9.6 Important discrete random variables

This section deals with discrete random variables which appear frequently in applications. We give the probability mass functions of these random variables together with the corresponding formulas for the expected value and the variance. In particular, we pay attention to the binomial, the Poisson and the hypergeometric random variables and give several illustrative examples.

9.6.1 The binomial random variable

Let us first introduce the Bernoulli random variable being the building block of the binomial random variable. 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. Since $E(X) = 0 \times (1 - p) + 1 \times p = p$ and $E(X^2) = E(X) = p$, it follows that the expected value and the variance of X are given by

$$E(X) = p \quad \text{and} \quad \text{var}(X) = p(1 - p).$$

To introduce the binomial distribution, imagine that n independent repetitions of a Bernoulli trial are performed. Let the random variable X be defined as the total number of successes in the n Bernoulli trials each having the same probability p of success. Then

$$P(X = k) = \begin{cases} \binom{n}{k} p^k (1 - p)^{n-k} & \text{for } k = 0, 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

This probability mass function is said to be the binomial distribution with parameters n and p. The derivation of $P(X = k)$ is simple. The sample space of the compound experiments consists of all possible n-tuples of 0s and 1s, where a "1" means a "success" and a "0" means a "failure". Any specific element containing k ones and $n - k$ zeros gets assigned the probability $p^k(1 - p)^{n-k}$, by the independence of the trials. The number of elements containing k ones and $n - k$ zeros is $\binom{n}{k}$. This completes the explanation of the formula for $P(X = k)$. The expected value and the variance of the random variable X are given by

$$E(X) = np \quad \text{and} \quad \text{var}(X) = np(1 - p).$$

These results can be derived by working out $E(X) = \sum_{k=1}^{n} kP(X = k)$ and $E(X(X - 1)) = \sum_{k=1}^{n} k(k - 1)P(X = k)$. However, this requires quite some algebra. A simpler approach is as follows. Write the random variable X as $X = X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are independent random variables each having a Bernoulli distribution with parameter p. Next, using the fact that $E(X_i) = p$ and $\text{var}(X_i) = p(1 - p)$, an application of the Rules 9.1 and 9.8 gives the results for $E(X)$ and $\text{var}(X)$.

Example 9.11 Chuck-a-Luck is a carnival game of chance and is played with three dice. To play this game, the player chooses one number from the numbers $1, \ldots, 6$. The three dice are then rolled. If the player's number does not come up at all, the player loses \$10. If the chosen number comes up one, two, or three times, the player wins \$10, \$20, or \$30 respectively. What is the expected win for the house per wager?

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 $\frac{1}{2}$. Let the random variable X denote the number of times the customer's number comes up. The random variable X can be seen as the number of successes in $n = 3$ independent trials of a Bernoulli experiment with a probability of success of $p = \frac{1}{6}$ and thus has a binomial distribution. This gives

$$P(X = k) = \binom{3}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{3-k} \qquad \text{for } k = 0, 1, 2, 3.$$

Hence the expected win for the house per wager is given by

$$\begin{aligned} &10 \times \frac{125}{216} - 10 \times 3 \times \frac{25}{216} - 20 \times 3 \times \frac{5}{216} - 30 \times \frac{1}{216} \\ &= 100 \times \frac{17}{216} = 0.787 \text{ dollars.} \end{aligned}$$

A healthy house edge of 7.87%!

Example 9.12 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 between two successive 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 two successive inspections?

Solution. Suppose that r reserve detectors are installed. Let the random variable X denote the number of detectors that will cease to function between two consecutive inspections. Then the random variable X has a binomial distribution with parameters $n = 5 + r$ and $p = 0.05$. The probability that all detectors will cease to function between two inspections is

$$P(X > r) = \sum_{k=r+1}^{5+r} \binom{5+r}{k} 0.05^k \, 0.95^{5+r-k}.$$

This probability has the value 0.0038 for $r = 2$ and the value 0.00037 for $r = 3$. Hence three reserve detectors should be installed.

9.6.2 The Poisson random variable

A random variable X is said to have a Poisson distribution with parameter $\lambda > 0$ if

$$P(X = k) = \begin{cases} e^{-\lambda}\,\frac{\lambda^k}{k!} & \text{for } k = 0, 1, \ldots \\ 0 & \text{otherwise,} \end{cases}$$

where $e = 2.71828\ldots$ is the base of the natural logarithm. In Example 9.8 it was verified that

$$E(X) = \lambda \quad \text{and} \quad \text{var}(X) = \lambda.$$

The Poisson distribution has nearly all of its mass within three standard deviations from the expected value if λ is sufficiently large (say, $\lambda \geq 25$). The probability $P(|X - \lambda| > 3\sqrt{\lambda})$ is on the order of 2×10^{-3} if $\lambda \geq 25$. In other words, it would be exceptional for a Poisson distributed random variable to take on a value that differs more than $3\sqrt{\lambda}$ from the expected value λ. This is a very useful fact for statistical applications of the Poisson model.

The Poisson distribution can be seen as a limiting case of the binomial distribution with parameters n and p when n is very large and p is very small. To give a precise mathematical formulation of the approximation result, let X represent a binomially distributed random variable with the parameters n and p. Assume now that n becomes *very large* and p becomes *very small* while np remains equal to the constant λ. The following is then true:

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

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

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

Now let's look at the different terms separately. Take a fixed value for k, where $0 \leq k \leq n$. 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 $(1 - \lambda/n)^{-k}$. A well-known result is 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 desired result.

The Poisson distribution can be used to model many practical phenomena. The explanation is that in many practical situations you can speak of many independent repetitions of a Bernoulli trial with a very small "success" probability. Examples include the number of credit cards that are stolen yearly in a certain area, the number of damage claims filed yearly with an insurance company, the yearly number of impacts of meteorites, and so on. The Poisson model 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. The value λ of this product is usually known in practical applications and uniquely determines the Poisson distribution.

Example 9.13 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 can be calculated as

$$P(X > 150) = 1 - \sum_{k=0}^{150} e^{-100}\,\frac{100^k}{k!} = 1.23 \times 10^{-6}.$$

Not a probability the insurance executives need worry about.

Example 9.14 In the kingdom of Lightstone, the game of Lotto 6/42 is played. In Lotto 6/42 six distinct numbers out of the numbers $1, \ldots, 42$ are randomly drawn. At the time of an oil sheik's visit to Lightstone the jackpot (six winning numbers) 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 are not 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 10 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 sheik will be among the winners of the jackpot and what is the probability that the oil sheik will be the only winner? What is the probability that the sheik's initial outlay will be won back from the jackpot? What are the expected value and the standard deviation of the amount the sheik will get from the jackpot?

Solution. The probability that a particular ticket is a winning ticket for the jackpot is $p = 1/\binom{42}{6}$. Very many tickets are filled in by the sheik and the local people. The Poisson model is applicable. Let the random variable X denote the number of winning tickets for the jackpot among the 15,000,000 tickets of the sheik and the random variable Y the number of winning tickets for the jackpot among the 10,000,000 tickets of the locals. The random variables X and Y can be modeled as Poisson random variables with respective parameters

$$\lambda = \frac{15{,}000{,}000}{\binom{42}{6}} = 2.85944 \quad \text{and} \quad \mu = \frac{10{,}000{,}000}{\binom{42}{6}} = 1.90629.$$

The random variables X and Y are independent. The probability that the sheik will be among the winners of the jackpot is $P(X \geq 1) = 1 - e^{-\lambda} = 0.9427$. The probability that the sheik is the only winner of the jackpot is $P(X \geq 1,\ Y = 0) = P(X \geq 1)P(Y = 0)$ and thus is equal to $(1 - e^{-\lambda})\, e^{-\mu} = 0.1401$. To answer the last two questions, let $p(r, s)$ denote the probability that the sheik has r winning tickets for the jackpot and the locals have s winning tickets for the jackpot. Then $p(r, s) = P(X = r,\ Y = s)$ and so

$$p(r, s) = e^{-\lambda}\, \frac{\lambda^r}{r!} \times e^{-\mu}\, \frac{\mu^s}{s!} \quad \text{for } r, s \geq 0.$$

Define the set $A = \{(r, s) : \frac{r}{r+s}\, 27.5 \geq 10\}$. Then the probability that the sheik's initial outlay will be won back from the jackpot is equal to $\sum_{(r,s)\in A} p(r, s) = 0.5785$. Let the random variable W be the amount the sheik will get from the jackpot (in millions of dollars). Then the expected value $E(W)$ and the standard deviation $\sigma(W)$ are obtained by computing the values of

$$E(W) = \sum_{r,s} \frac{r}{r+s}\, 27.5\, p(r, s) \quad \text{and} \quad E(W^2) = \sum_{r,s} \left(\frac{r}{r+s}\, 27.5\right)^2 p(r, s).$$

We find that the expected value and the standard deviation of the amount the sheik will get from the jackpot are \$16,359,475 and \$7,162,762.

The Poisson model is one of the most useful probability models. It applies to 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. In surprisingly many probability problems, the Poisson model enables us to obtain practically useful approximations for probabilities that are otherwise difficult to calculate. In Section 4.2.3 this approach is discussed and illustrated with many examples.

9.6.3 The hypergeometric random variable

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 randomly select n balls out of the urn without replacing any. Define the random variable X as the number of red balls among the selected balls. Then,

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

By the convention $\binom{a}{b} = 0$ for $b > a$, we also have that $P(X = r) = 0$ for those r with $r > R$ or $n - r > W$. This probability mass function is called the hypergeometric distribution with parameters R, W and n. The explanation of the formula for $P(X = r)$ is as follows. The number of ways in which r red balls and $n - r$ white balls can be chosen from the urn is $\binom{R}{r}\binom{W}{n-r}$, while 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 $P(X = r)$. The expected value and the variance of the random variable X are given by

$$E(X) = n\frac{R}{R+W} \quad \text{and} \quad \text{var}(X) = n\frac{R}{R+W}\left(1 - \frac{R}{R+W}\right)\frac{R+W-n}{R+W-1}.$$

The simplest way to prove these results is to use indicator variables. Let the random variable X_i be equal to 1 if the ith ball drawn is red and 0 otherwise. By a symmetry argument, the random variables $X_1, \ldots, X_r$ have the same distribution, though these random variables are dependent. Obviously, $P(X_1 = 1) = \frac{R}{R+W}$ and so $E(X_1) = \frac{R}{R+W}$. This gives $E(X) = n\frac{R}{R+W}$. Using the indicator variables X_i, you are asked in Problem 11.27 to derive the formula for var(X).

The hypergeometric model is a versatile model and has many applications, particularly to lotteries (see also Section 4.3). The hypergeometric distribution with parameters R, W and n can be approximated by the binomial distribution with parameters n and $p = \frac{R}{R+W}$ if $R + W \gg n$.

Example 9.15 The New York State Lottery offers the game called Quick Draw, a variant of Keno. The game can be played in bars, restaurants, bowling areas and other places. A new game is played every four or five minutes and so a lot of games can be played in a day. A player chooses four numbers from 1 to 80. The lottery then randomly chooses twenty numbers from 1 to 80. The number of matches determines the payoff. The payoffs on a one-dollar bet are \$55 for four matches, \$5 for three matches, \$1 for two matches and \$0 otherwise. In November of 1997, the state lottery offered a promotion "Big Dipper Wednesday" where payoffs on the game were doubled on the four Wednesdays in that month. Is this a good deal for the player or just a come-on for a sucker bet?

Solution. Let us first have a look at the game with the ordinary payoffs. The hypergeometric model with $R = 4$, $W = 76$, and $n = 20$ is applicable to this game of Quick Draw. Let the random variable X indicate how many numbers are matched by a single ticket. Then,

$$P(X = k) = \frac{\binom{4}{k}\binom{76}{20-k}}{\binom{80}{20}} \quad \text{for } k = 0, 1, \ldots, 4.$$

This probability has the numerical values 0.308321, 0.432732, 0.212635, 0.043248, and 0.003063 for $k = 0, 1, \ldots, 4$. The expected payoff per one-dollar bet is $1 \times 0.212635 + 5 \times 0.043248 + 55 \times 0.003063 = 0.59735$ dollars. In other words, you would expect to lose about 40 cents per dollar bet on average; the usual house edge for this type of game. Now, what about the "Big Dipper Wednesday"? The expected payoff per one dollar bet for the game with double payoffs becomes twice as much and is 1.1947 dollars. This is more than the amount of the one-dollar bet! The player now enjoys a 19.47% edge over the house in the game with double payoffs. It seems that the New York State Lottery made a miscalculation.

Example 9.16 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 randomly chosen distinct numbers between 1 and 80, one at a time. What is the probability that more than 70 numbers must be called out before one of the players has achieved a full card? What is the probability that you will be the first player to achieve a

full card while no other player has a full card at the same time as you? What is the probability that you will be among the first players achieving a full card?

Solution. The key to solving this problem is to find the probability mass function of the random variable X counting how many numbers have to be called out before a *particular* player has achieved a full card. Denote the probability Q_k by $Q_k = P(X > k)$ for $k = 23, \ldots, 80$, where $Q_{23} = 1$ and $Q_{80} = 0$. Let us first determine Q_{70}. Note that Q_{70} is the same as the probability of getting fewer than 24 red balls when drawing 70 balls at random and without replacement from an urn with 24 red balls and 56 white balls. The latter probability is one minus the probability of getting 24 red balls when drawing 70 balls at random and without replacement from an urn with 24 red balls and 56 white balls. This gives

$$Q_{70} = 1 - \frac{\binom{56}{46}}{\binom{80}{70}} = 0.978374.$$

By the same argument, we obtain for any n with $24 \leq n \leq 79$ that

$$Q_n = 1 - \frac{\binom{56}{n-24}}{\binom{80}{n}}.$$

We are now in a position to answer the questions. The probability that more than 70 numbers must be called out before one of the players has achieved a full card is given by

$$Q_{70}^{36} = 0.4552.$$

To find the second probability, note that your probability of having a full card after n numbers have been called out is given by $Q_{n-1} - Q_n$. Hence

$$\sum_{n=24}^{79} (Q_{n-1} - Q_n)\, Q_n^{35} = 0.0228$$

gives the probability that you will be the first player to achieve a full card while no other player has a full card at the same time as you. Finally, the probability that you will be among the first players achieving a full card is given by

$$\sum_{n=24}^{80} \sum_{a=0}^{35} \binom{35}{a} [Q_{n-1} - Q_n]^{a+1} \times Q_n^{35-a} = \sum_{n=24}^{80} [Q_{n-1} - Q_n] Q_{n-1}^{35}$$

and has the numerical value 0.0342.

9.6.4 The discrete uniform random variable

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. Using the fact that $\sum_{k=1}^{n} k = \frac{1}{2}n(n+1)$ and $\sum_{k=1}^{n} k^2 = \frac{1}{6}n(n+1)(2n+1)$ for all $n \geq 1$, it is a matter of some algebra to verify that

$$E(X) = \frac{a+b}{2} \quad \text{and} \quad \text{var}(X) = \frac{(b-a+1)^2 - 1}{12}.$$

Example 9.17 Two friends lose each other while wandering through a crowded amusement park. They cannot communicate by mobile phone. Fortunately, they have foreseen that losing each other could happen and have made a prior agreement. The younger friend stays put at the entrance of one of the main attractions of the park and the older friend searches these main attractions in random order (the "wait-for-mommy" strategy). The park has 15 main attractions. What are the expected value and the standard deviation of the number of searches before the older friend finds his younger friend?

Solution. Let the random variable X denote the number of searches before the older friend finds his younger friend. The random variable X has a discrete uniform distribution on the integers $1, 2, \ldots, 15$. Hence the expected and the standard deviation of X are given by $E(X) = 8$ and $\sigma(X) = 4.32$.

9.6.5 The geometric random variable

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

$$E(X) = \frac{1}{p} \quad \text{and} \quad \text{var}(X) = \frac{1-p}{p^2}.$$

Example 9.18 Suppose one die is rolled over and over. How many rolls are needed so that the probability of rolling a six within this number of rolls is at least 50%?

Solution. Let the random variable X be defined as the number of rolls of the die until the first six appears. Then the random variable X has a geometric distribution with parameter $p = \frac{1}{6}$. We are asking for the smallest integer k for which $P(X \leq k) \geq 0.5$).[†] Using the fact that $P(X \leq k) = 1 - (1-p)^k$, it follows that the number of rolls needed is the smallest integer k that is larger than or equal to $\ln(0.5)/\ln(5/6) = 3.802$. Hence four rolls are needed.

9.6.6 The negative binomial random variable

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 by the probability p of having a success at the kth trial. The random variable X can be written as $X_1 + \cdots + X_n$, where X_i is the number of trials needed in order to go from $i-1$ to i successes and $X_1, \ldots, X_n$ are independent random variables each having a geometric distribution with parameter p. Using the fact that $E(X_i) = p$ and $\text{var}(X_i) = (1-p)/p^2$, an application of the Rules 9.1 and 9.8 gives

$$E(X) = \frac{r}{p} \quad \text{and} \quad \text{var}(X) = \frac{r(1-p)}{p^2}.$$

[†] This value of k is the median of the distribution. In general, the median of an integer-valued random variable X is any integer m with $P(X \leq m) \geq 0.5$ and $P(X \geq m) \geq 0.5$. The median of a probability distribution need not be unique.

Example 9.19 Suppose one die is rolled over and over. How many rolls are needed so that the probability of rolling a six three times within this number of rolls is at least 50%?

Solution. Let the random variable X be defined as the number of rolls of the die until a six appears for the third time. Then the random variable X has a negative binomial distribution with parameters $r = 3$ and $p = \frac{1}{6}$. The number of rolls needed is the smallest integer k for which

$$\sum_{j=3}^{k} \binom{j-1}{2} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{j-3} \geq 0.5.$$

Numerical computations give $k = 16$. Hence 16 rolls are needed.

Problem 9.30 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 9.31 It is your birthday and you are at an assembly where 500 other persons are also present. The organizers of the assembly are raffling off a prize among all of those present whose birthday falls on that particular day. What is the probability mass function of the number of other persons at the assembly who have the same birthday as you? What is the probability that you will be the winner of the prize? Can you explain why this probability can be approximated by $\frac{1}{\lambda}(1 - e^{-\lambda})$ with $\lambda = 500/365$?

Problem 9.32 On bridge night, the cards are dealt round nine times. Only twice 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 9.33 In the final of the World Series Baseball, two teams play a series consisting of at most seven games until one of the two teams has won four games. 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 of 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 9.34 How many rolls of two dice are needed so that the probability of rolling a double six within this number of rolls is at least 50%? How many rolls of two dice are needed so that the probability of rolling a double six three times within this number of rolls is at least 50%?

Problem 9.35 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 9.36 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 9.37 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 9.38 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 the exam?

Problem 9.39 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 9.40 What is the probability that a thirteen-card bridge hand contains at least k spades and what is the probability that a five-card poker hand contains at least k spades? Calculate these probabilities for the relevant values of k.

Problem 9.41 What is the probability that you hold one pair in a five-card poker hand and what is the probability that you hold two pairs?

Problem 9.42 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 9.43 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 9.44 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?

Problem 9.45 The "New Amsterdam Lottery" offers the game "Take Five." In this game, players must tick five different numbers from the numbers $1, \ldots, 39$. The lottery draws five distinct numbers from the numbers $1, \ldots, 39$. 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?

Problem 9.46 A fair coin is tossed until heads appears for the third time. Let the random variable X be the number of tails shown up to that point. What is the probability mass function of X? What are the expected value and standard deviation of X?

Problem 9.47 John and Pete are having a nice evening in the pub. They decide to play the following game in order to determine who pays for the beer. Each of them rolls two dice. The game ends if the dice total of John is the same as that of Pete; otherwise, they play a next round. Upon ending the game, John pays for the beer if the dice total is odd and otherwise Pete pays. What is the probability mass function of the number of rounds required for the game? What is the probability of John paying for the beer?

Problem 9.48 An experiment has r possible outcomes $O_1, O_2, \ldots, O_r$ with respective probabilities of $p_1, p_2, \ldots, p_r$. Suppose that n independent repetitions of the experiment are performed. Let the random variable X_i be the number of times that the outcome O_i occurs. Argue that

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_r = x_r) = \frac{n!}{x_1!x_2!\cdots x_r!} p_1^{x_1} p_2^{x_2} \cdots p_r^{x_r}$$

for all non negative integers $x_1, x_2, \ldots x_r$ with $x_1 + x_2 + \cdots x_r = n$. This distribution is called the *multinomial distribution*.

Problem 9.49 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?

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 taking on a discrete number of values and continuous random variables taking on a continuum of values. 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 of a discrete random variable.

The purpose of this chapter is to familiarize the reader with the concept of 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. In particular, the exponential density and the normal density are treated in depth. Many practical phenomena can be modeled by these distributions which

are of fundamental importance. Also, attention is given to the central limit theorem being the most important theorem of probability theory. Finally, the inverse-transformation method for simulating a random observation from a continuous random variable and the important concept of failure rate function will be discussed.

10.1 Concept of probability density

The simplest 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 makes only 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). Before defining 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,

$$F(a) = \frac{2a}{1+a} \qquad \text{for } 0 < a < 1.$$

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 two examples 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

$$P(X \le x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{1}{4}x^2 & \text{for } 0 \le x \le 2, \\ 1 & \text{for } x > 2. \end{cases}$$

The probability distribution function $P(X \le x)$ 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 $P(X \le x)$ 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.

Example 10.3 A continuous random variable X has a probability density $f(x)$ of the form $f(x) = ax + b$ for $0 < x < 1$ and $f(x) = 0$ otherwise. What conditions on the constants a and b must be satisfied? What is the cumulative probability distribution function of X?

Solution. The requirements for $f(x)$ are $ax + b \ge 0$ for $0 < x < 1$ and $\int_0^1 (ax + b)\,dx = 1$. The first requirement gives $a + b \ge 0$ and $b \ge 0$. The second requirement gives $\frac{1}{2}a + b = 1$. The cumulative probability distribution function of X is equal to

$$F(x) = \int_0^x (av + b)\,dv = \frac{1}{2}ax^2 + bx \quad \text{for } 0 \le x \le 1.$$

Further, $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x > 1$.

Problem 10.1 The lifetime of an appliance is a continuous random variable X and has a probability density $f(x)$ of the form $f(x) = c(1 + x)^{-3}$ for $x > 0$ and $f(x) = 0$ otherwise. What is the value of the constant c? Find $P(X \le 0.5)$, $P(0.5 < X \le 1.5)$ and $P(0.5 < X \le 1.5 \mid X > 0.5)$.

Problem 10.2 Let the random variable X be the portion of a flood insurance claim for flooding damage to the house. The probability density of X has the form $f(x) = c(3x^2 - 8x - 5)$ for $0 < x < 1$. What is the value of the constant c? What is the cumulative probability distribution function of X?

Problem 10.3 Sizes of insurance claims can be modeled by a continuous random variable with probability density $f(x) = \frac{1}{50}(10 - x)$ for $0 < x < 10$ and $f(x) = 0$ otherwise. What is the probability that the size of a particular claim is larger than 5 given that the size exceeds 2?

Problem 10.4 The lengths of phone calls (in minutes) made by a travel agent can be modeled as a continuous random variable with probability density $f(x) = 0.25e^{-0.25x}$ for $x > 0$. What is the probability that a particular phone call will take more than 7 minutes?

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 of a substance, 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

$$\begin{aligned} P(X = a) &= \lim_{n\to\infty} P(a - \frac{1}{n} < X \leq a) \\ &= \lim_{n\to\infty} \int_{a-\frac{1}{n}}^{a} f(x)\,dx = \int_{a}^{a} f(x)dx, \end{aligned}$$

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

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

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

$$P\left(a - \frac{1}{2}\Delta a \le X \le a + \frac{1}{2}\Delta a\right) = \int_{a-\frac{1}{2}\Delta a}^{a+\frac{1}{2}\Delta a} f(x)\,dx$$

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

$$P\left(a - \frac{1}{2}\Delta a \le X \le a + \frac{1}{2}\Delta a\right) \approx f(a)\Delta a \qquad \text{for } \Delta a \text{ small}.$$

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 it 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 10.4 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 $x > 0$,

$$P(X \leq x) = P\left(-\frac{1}{\lambda}\ln(U) \leq x\right) = P(\ln(U) \geq -\lambda x)$$
$$= P(U \geq e^{-\lambda x}) = 1 - P(U \leq e^{-\lambda x}),$$

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 10.5 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 point $x = a$, 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 then differentiated to obtain the probability density.

As pointed out before, the value of the probability density at any point a is a relative measure for the likelihood that the random variable will take on a value in the immediate neighborhood of the point a. To illustrate this, let us put the following question with regard to the last example. A point will be randomly chosen within the unit disk and you are asked to bet on the value of the distance from the chosen point to the center of the disk. You win the bet if your guess is no more than 5% off from the observed distance. Should your guess be a number close to zero or close to 1? The probability density function of the distance is $f(x) = 2x$ for $0 < x < 1$ and so your guess should be close to 1. The best value of your guess follows by maximizing $\int_{c-0.05c}^{c+0.05c} f(x)\,dx$ with respect to c. This gives $c = \frac{20}{21}$ with a win probability of 0.1814.

Problem 10.5 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 10.6 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 10.7 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 10.8 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 10.9 A point is randomly chosen inside the unit square. The random variables V and W are defined as the largest and the smallest of the two coordinates of the point. What are the probability density functions of the random variables V and W?

Problem 10.10 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? (It is assumed that the bottom of the wheel is level with the ground.) 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?

10.2 Expected value of a continuous random variable

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 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 10.5. The expected value of the distance X equals

$$E(X) = \int_0^r x\frac{2x}{r^2}\,dx = \frac{2}{3}\frac{x^3}{r^2}\bigg|_0^r = \frac{2}{3}r.$$

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,

$$\begin{aligned} E(X) &= \int_0^1 x\frac{2}{(x+1)^2}\,dx = 2\int_0^1 \frac{1}{x+1}\,dx - 2\int_0^1 \frac{1}{(x+1)^2}\,dx \\ &= 2\ln(x+1)\Big|_0^1 + 2\frac{1}{x+1}\bigg|_0^1 = 2\ln(2) - 1. \end{aligned}$$

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

$$E(Y) = \int_1^\infty y\frac{2}{(y+1)^2}\,dy = 2\ln(y+1)\Big|_1^\infty + 2\frac{1}{y+1}\bigg|_1^\infty = \infty.$$

This finding is in agreement with the result of Problem 9.11 in Section 9.2. A little calculus was enough to find a result that otherwise is difficult to obtain from a simulation study.

Problem 10.11 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 10.12 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 $a, b \geq 0$.

Problem 10.13 A point is chosen at random inside the unit square $\{(x, y) : 0 \leq x, y \leq 1\}$. What is the expected value of the distance from this point to the point (0, 0)?

Problem 10.14 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 10.15 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 10.16 Let X be a nonnegative continuous random variable with density function $f(x)$. Use an interchange of the order of integration to verify that $E(X) = \int_0^\infty P(X > u)\,du$.

10.3 Substitution rule and the 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.

Rule 10.1 *Let X be a continuous random variable with probability density $f(x)$. Then, 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.

We first give the proof for the case that the random variable $g(X)$ is nonnegative. The proof is based on the fact that

$$E(Y) = \int_0^\infty P(Y > y)\,dy$$

for any nonnegative random variable Y (see also Problem 10.16).[†] Assuming that $g(X)$ is nonnegative, we get

$$E[g(X)] = \int_0^\infty P(g(X) > y)\,dy = \int_0^\infty dy \int_{x:\, g(x)>y} f(x)\,dx$$
$$= \int_{x:\, g(x)>0} f(x)\,dx \int_0^{g(x)} dy = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

It is noted that the interchange of the order of integration is justified by the fact that the integrand is nonnegative. The proof for the general case follows by writing $g(X)$ as the difference of the two nonnegative random variables $\max(g(X), 0)$ and $-\min(g(X), 0)$.

We give two illustrative examples of the substitution rule.

Example 10.6 A warranty on an appliance specifies that an amount of \$250 will be reimbursed if the appliance fails during the first year, an amount of \$125 if it fails during the second year, and nothing if it fails after the second year. The time, measured in years, until the appliance fails has the probability density $f(x) = 0.2e^{-0.2x}$ for $x > 0$. What is the expected value of the warranty reimbursement?

Solution. Let the random variable X be the time until the failure of the appliance. Define the function $g(x)$ by $g(x) = 250$ for $0 \le x \le 1$, $g(x) = 125$ for $1 < x \le 2$, and $g(x) = 0$ for $x > 2$. Then $g(X)$ is the warranty payment. Its expected value is

$$E[g(X)] = \int_0^1 250 \times \frac{1}{5} e^{-x/5}\,dx + \int_1^2 125 \times \frac{1}{5} e^{-x/5}\,dx$$
$$= 250(1 - e^{-1/5}) + 125(e^{-1/5} - e^{-2/5}) = 63.87 \text{ dollars}.$$

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 the function $g(u)$ is defined by $g(u) = u/(1 - u)$ for $0 < u \le \frac{1}{2}$ and

[†] A probabilistic proof is as follows. Think of Y as the lifetime of some item. For any $t > 0$, let $I_t = 1$ if the lifetime of the item has not yet expired at time t and $I(t) = 0$ otherwise. Then $E(Y) = E(\int_0^\infty I_t\,dt) = \int_0^\infty E(I_t)\,dt$. By $E(I_t) = P(Y > t)$, the result now follows.

$g(u) = (1-u)/u$ for $\frac{1}{2} < u < 1$. This gives

$$E(V) = \int_0^{1/2} \frac{u}{1-u}\,du + \int_{1/2}^1 \frac{1-u}{u}\,du = 2\int_{1/2}^1 \frac{1-u}{u}\,du$$
$$= 2\ln(u) - 2u\Big|_{1/2}^1 = 2\ln(2) - 1.$$

Variance of a continuous random variable

Next we deal with the variance and the standard deviation of a continuous random variable. Recall from Chapter 9 that the variance of a random variable X is defined by $\text{var}(X) = E[(X-\mu)^2]$, where $\mu = E(X)$. If X is a continuous random variable with probability density $f(x)$, then the variance of X 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 10.5:

$$\text{var}(X) = \int_0^r x^2 \frac{2x}{r^2}\,dx - \left(\frac{2}{3}r\right)^2 = \frac{2r^2}{4} - \frac{4}{9}r^2 = \frac{1}{18}r^2.$$

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 10.7 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, $P(X \le x) = \frac{x-a}{b-a}$ for $a \le x \le b$ 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

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a}\, dx = \frac{1}{3}\frac{x^3}{b-a}\bigg|_a^b = \frac{1}{3}\frac{b^3-a^3}{b-a} = \frac{a^2+ab+b^2}{3},$$

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 10.17 A manufacturer has to make a last production run for a product that is near the end of its lifetime. The final demand for the product can be modeled as a continuous random variable X having probability density $f(x) = \frac{1}{2{,}500} x e^{-x/50}$ for $x > 0$. It is decided to make a last production run of 250 units of the product. The manufacturer earns 2 for each unit product sold but incurs a cost of 0.50 for each unit of demand occurring when being out of stock. What is the expected value of the net profit of the manufacturer? What is the probability that the manufacturer runs out of stock?

Problem 10.18 A car owner insures his car worth \$20,000 for one year under a policy with a deductible of \$1,000. There is a probability of 0.01 of a total loss of the car during the policy year and a probability of 0.02 of a repairable damage. The cost (in thousands of dollars) of a repairable damage has the probability density $f(x) = \frac{1}{200}(20 - x)$ for $0 < x < 20$. What is the expected value of the insurance payment (this payment is a mixed random variable)?

Problem 10.19 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 10.20 The lifetime (in months) of a battery is a random variable X satisfying $P(X \le x) = 0$ for $x < 5$, $P(X \le x) = [(x-5)^3 + 2(x-5)]/12$ for $5 \le x < 7$ and $P(X \le x) = 1$ for $x \ge 7$. What are the expected value and the standard deviation of X?

Problem 10.21 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 a continuous random variable X is any value m for which $P(X \leq m) = P(X \geq m) = \frac{1}{2}$.†

Problem 10.22 Consider Problem 10.10 again. Calculate the expected value and standard deviation of the height above the ground when the ferris wheel stops.

Problem 10.23 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 10.24 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 people 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 10.25 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 10.26 Let the random variables V and W be defined by $V = \sqrt{U}$ and $W = U^2$, where 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. The median is primarily used for skewed distributions such as the income distributions.

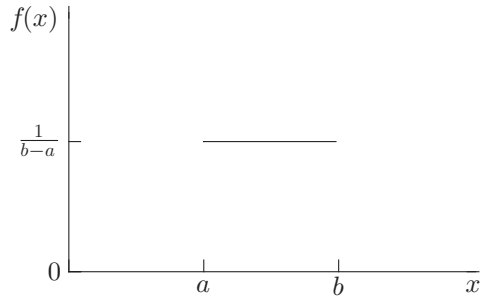


Fig. 10.1. Uniform density.

10.4 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.

10.4.1 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

$$P\left(c - \frac{1}{2}\Delta \le X \le c + \frac{1}{2}\Delta\right) = \int_{c-\frac{1}{2}\Delta}^{c+\frac{1}{2}\Delta} \frac{1}{b-a}\,dx = \frac{\Delta}{b-a},$$

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 (see Example 10.7)

$$E(X) = \frac{1}{2}(a+b) \quad \text{and} \quad \text{var}(X) = \frac{1}{12}(b-a)^2.$$

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 $x > b$, and

$$F(x) = \frac{x-a}{b-a} \qquad \text{for } a \le x \le b.$$

Problem 10.27 The lifetime of a light bulb has an uniform probability density on (2, 12). The light bulb will be replaced upon failure or upon reaching the age 10, whichever occurs first. What are the expected value and the standard deviation of the age of the light bulb at the time of replacement?

Problem 10.28 A rolling machine produces sheets of steel of different thickness. The thickness of a sheet of steel is uniformly distributed between 120 and 150 millimeters. Any sheet having a thickness of less than 125 millimeters must be scrapped. What are the expected value and the standard deviation of a non-scrapped sheet of steel?

10.4.2 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

$$f(x) = \begin{cases} h\,\frac{x-a}{m-a} & \text{for } a < x \le m \\ h\,\frac{b-x}{b-m} & \text{for } m \le x < b \\ 0 & \text{otherwise.} \end{cases}$$

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

$$h = \frac{2}{b-a}.$$

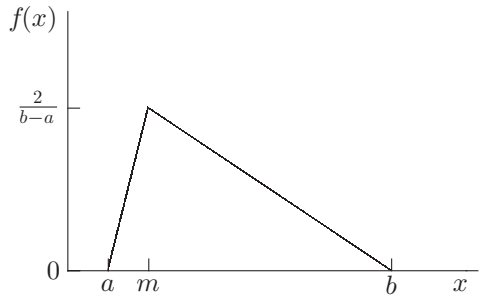


Fig. 10.2. Triangular density.

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 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 $x < a$, $F(x) = 1$ for $x > b$, and

$$F(x) = \begin{cases} \frac{(x-a)^2}{(b-a)(m-a)} & \text{for } a \le x < m \\ 1 - \frac{(b-x)^2}{(b-a)(b-m)} & \text{for } m \le x \le b. \end{cases}$$

10.4.3 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 probability model for the time until a *rare* event occurs. Examples are the time

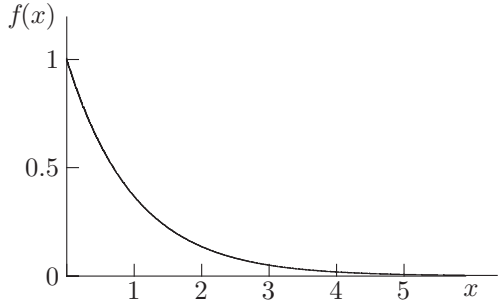


Fig. 10.3. Exponential density ($\lambda = 1$).

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

$$P(X > t + s \mid X > s) = P(X > s) \qquad \text{for all } t > 0,$$

regardless of the value of s. 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. The proof is simple. For any $x \geq 0$, we have $P(X > x) = e^{-\lambda x}$. Using the basic formula $P(A \mid B) = P(AB)/P(B)$ with $A = \{X > t + s\}$ and $B = \{X > s\}$, and noting that $P(AB) = P(A)$, we find

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

showing the memoryless property. The exponential distribution is the only continuous distribution possessing this property. The technical proof of this result is omitted.

In many applied probability problems one has to study rare events. For example, a rare event could be a system failure in reliability applications. A very useful result is that under general conditions the time until the first occurrence of a rare event is approximately exponentially distributed. To make this result plausible, consider the following example. Maintenance of an operating unit in a reliability system occurs at the scheduled times $\tau, 2\tau, ...$, where $\tau > 0$ is fixed. Each maintenance takes a negligible time and the unit is again as good as new after each maintenance. There is a given probability $p > 0$ that the unit will fail between two maintenance inspections, where p is very close to zero. Let the random variable X denote the time until the first system failure. Then, for any n, we have $P(X > n\tau) = (1 - p)^n$ and so $P(X > n\tau) \approx e^{-np}$, by $e^{-p} \approx 1 - p$ for p close to zero. Hence, taking $t = n\tau$ and replacing n by t/τ, it follows that

$$P(X > t) \approx e^{-\lambda t} \quad \text{for } t > 0,$$

where $\lambda = p/\tau$ denotes the inverse of the expected time until the first system failure. It is important to note that it suffices to know the expected value of the time until the first system failure in order to approximate the probability distribution of the random variable X.

Also, the probability of exceeding some *extreme* level is often approximately equal to an exponential tail probability, where an exponential tail probability is a probability of the form $\alpha e^{-\beta t}$ for constants $\alpha, \beta > 0$. An 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 meters. 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.

Example 10.8 A reliability system has two identical units, where one unit is in full operation and the other unit is in cold standby. The lifetime of an operating unit has an exponential density with expected value $1/\mu$. Upon failure of the operating unit, the standby unit is put into operation provided a standby unit is available. The replacement time of a failed unit is fixed and is equal to $\tau > 0$. A system failure occurs if no standby unit is available at the moment the operating unit fails. It is assumed that the probability $1 - e^{-\mu\tau}$ is close to zero, that is, the probability of an operating unit failing during the replacement time τ is very

small. What is an approximation to the probability distribution of the time until the first system failure?

Solution. Let the random variable X denote the time until the first system failure. Denote by α the expected value of X. In Example 13.11 in Section 13.3, it will be shown that

$$E(X) = \frac{2 - e^{-\mu\tau}}{\mu(1 - e^{-\mu\tau})}.$$

Under the assumption that $1 - e^{-\mu\tau}$ is very small, the occurrence of a system failure is a rare event and thus the probability distribution of X can be approximated by

$$P(X > t) \approx e^{-t/E(X)} \quad \text{for } t > 0.$$

As an illustration, take $\mu = 1$ and $\tau = 0.02$. Then $E(X) = 51.50167$. The probability $P(X > t)$ has the approximate values 0.8235, 0.6154, 0.3788, 0.1435, and 0.0206 for t =10, 25, 50, 100, and 200. The approximation is very accurate. Simulation with 100,000 runs for each t gives the simulated values 0.8232, 0.6148, 0.3785, 0.1420, and 0.0204.

Exponential distribution and its relation to the Poisson distribution

The Poisson distribution is closely related to the exponential distribution. To explain this, let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables each having the same exponential density $\lambda e^{-\lambda x}$. Define the random variable S_n by

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Then the probability density of the random variable S_n is given by

$$\frac{\lambda^n x^{n-1}}{n!} e^{-\lambda x} \quad \text{for } x > 0.$$

A proof of this result will be presented in Section 14.2. This density is called the Erlang density with parameters n and λ . The corresponding probability distribution function is given by

$$P(S_n \leq x) = 1 - \sum_{j=0}^{n-1} e^{-\lambda x} \frac{(\lambda x)^j}{j!} \quad \text{for } x > 0.$$

The easiest way to verify this formula is by differentiating its right-hand side term by term. We can now state the following important result for a sequence of events occurring in time (e.g., arrivals, or earthquakes).

Rule 10.2 *Let $X_1, X_2, \ldots$ be a sequence of independent random variables each having the same exponential density with expected value $1/\lambda$, where the random variable X_i describes the amount of time between the $(i-1)$th and ith occurrences of some specific event (the 0th occurrence of the event is at time 0).*[†] *For any $t > 0$, define the random variable $N(t)$ as the number of events occurring in $(0, t]$. Then, for any $t > 0$,*

$$P(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!} \quad \text{for } k = 0, 1, \ldots.$$

In words, $N(t)$ has a Poisson distribution with expected value λt.

The instructive proof is as follows. Fix $t > 0$. First, we verify that $P(N(t) = 0) = e^{-\lambda t}$. This follows directly from $P(N(t) = 0) = P(X_1 > t)$ and $P(X_1 > t) = e^{-\lambda t}$. Next observe that n or more events occur in $(0, t]$ if and only if the epoch of the nth occurrence of the event is before or at time t. That is,

$$P(N(t) \geq n) = P(S_n \leq t),$$

where $S_n = X_1 + \cdots + X_n$. Obviously, $P(N(t) \geq k+1) + P(N(t) = k) = P(N(t) \geq k)$ and so $P(N(t) = k$ is given by

$$P(N(t) \geq k) - P(N(t) \geq k+1) = P(S_k \leq t) - P(S_{k+1} \leq t).$$

Using the above formula for $P(S_n \leq t)$, it now follows that

$$P(N(t) = k) = 1 - \sum_{j=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!} - \left(1 - \sum_{j=0}^{k} e^{-\lambda t} \frac{(\lambda t)^j}{j!}\right) = e^{-\lambda t} \frac{(\lambda t)^k}{k!},$$

as was to be verified.

A close examination of the proof of Rule 10.2 reveals that the result of this rule can be extended as follows. For any $s, t > 0$,

$$P(N(s+t) - N(s) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!} \quad \text{for } k = 0, 1, \ldots.$$

This result is based on the fact that the amount of time measured from time point s to the first occurrence of an event has the same exponential distribution as X_1, by the lack of memory of the exponential distribution.

[†] This process is usually described as events occurring according to a Poisson process with intensity λ.

Example 10.9 In a hospital five babies are born per day on average. It is reasonable to model the times between successive arrivals of babies as independent random variables each having the same exponential distribution. Let the random variable X measure the time from midnight to the first arrival of a baby. What are the expected value and the median of X? What is the probability that more than two babies are born between twelve o'clock midnight and six o'clock in the morning?

Solution. Let us take the hour as unit of time. Then the random variable X has the exponential density $\lambda e^{-\lambda t}$ with $\lambda = \frac{5}{24}$. The expected value of X is $1/\lambda = 4.8$ hours. The median is found by solving m from $1 - e^{-\lambda m} = 0.5$ and thus is equal to $\ln(2)/\lambda = 3.33$ hours. The probability that more than two babies are born between twelve o'clock midnight and six o'clock in the morning is given by $1 - e^{-6\lambda} - 6\lambda e^{-6\lambda} - (6\lambda)^2 e^{-6\lambda}/2! = 0.1315$.

Problem 10.29 Let the random variable Y be distributed as $Y = \sqrt{X}$, where X has an exponential distribution with expected value $1/\lambda$. Verify that the density of Y is given by $2\lambda y e^{-\lambda y^2}$ for $y > 0$. This density is called the Rayleigh density What are the expected value and the variance of Y?

Problem 10.30 Limousines depart from the railway station to the airport from the early morning till late at night. The limousines leave from the railway station with independent interdeparture times that are exponentially distributed with an expected value of 20 minutes. Suppose you plan to arrive at the railway station at three o'clock in the afternoon. What are the expected value and the standard deviation of your waiting time at the railway station until a limousine leaves for the airport?

Problem 10.31 On weeknight shifts between 6 p.m and 10 p.m, 4.8 calls for medical emergencies arrive on average. It is reasonable to model the times between successive calls as independent random variables each having the same exponential distribution. Let the random variable X measure the time from 6 p.m until the first call occurs. What are the expected value and the median of X? What is the probability that the first call occurs between 6:20 p.m and 6:45 p.m? What is the probability of no calls between 7 p.m and 7:20 p.m and one or more calls between 7:20 p.m and 7:45 p.m?

Problem 10.32 On Wednesday afternoon between 1 p.m and 4:30 p.m, buses with tourists arrive in Gotham city to visit the castle in this picturesque town. The times between successive arrivals of buses are independent random variables each having an exponential distribution with an expected value of 45 minutes. Each bus stays exactly 30 minutes on the parking lot of the castle.

Explain why the number of buses on the parking lot at 4 p.m has a Poisson distribution with an expected value of $\frac{2}{3}$.

10.4.4 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

$$f(x) = \begin{cases} c\lambda^{\alpha} x^{\alpha-1} e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

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. If α is given by an integer n, the gamma density is often called the Erlang density with parameters n and λ. 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

$$E(X) = \frac{\alpha}{\lambda} \quad \text{and} \quad \text{var}(X) = \frac{\alpha}{\lambda^2}.$$

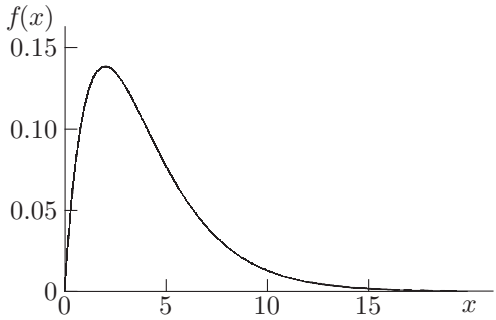


Fig. 10.4. Gamma density ($\alpha = 2.5, \lambda = 0.5$).

Problem 10.33 Use properties of the gamma function to derive $E(X)$ and $E(X^2)$ for a gamma-distributed random variable X.

10.4.5 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.

10.4.6 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 $\int_0^1 f(x)\,dx = 1$. The function $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$ is a classical function. A basic

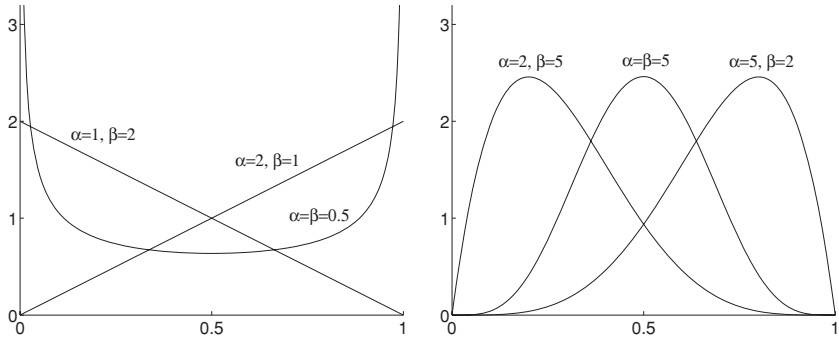


Fig. 10.5. Several beta densities.

result from calculus is that $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\,\Gamma(\alpha + \beta)$ and so

$$c = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

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

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \text{var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

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 10.34 Use properties of the beta function and the gamma function to derive the expressions for $E(X)$ and $\sigma^2(X)$ for a beta(α, β) distributed random variable X.

10.4.7 Normal density

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

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

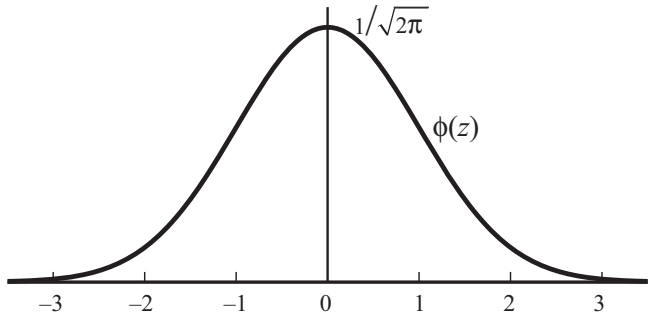


Fig. 10.6. Standard normal density

The parameter σ is a shape parameter, and the parameter μ is a scale parameter. The standard normal density with $mu = 0$ and $\sigma = 1$ is displayed in Figure 10.6. The normal density is symmetric around the point $x = \mu$. 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 density function and the cumulative probability distribution function of an $N(0, 1)$ distributed random variable Z are given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad \text{and } \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}y^2}\, dy.$$

No closed form of the cumulative distribution function $\Phi(z)$ exists. The integral for $\Phi(z)$ looks terrifying, but the integral can be approximated with extreme precision by the quotient of two suitably chosen polynomials. For all practical purposes the calculation of $\Phi(z)$ presents no difficulties at all and can be accomplished very quickly. The cumulative probability distribution function of an $N(\mu, \sigma^2)$ distributed random variable X can be calculated from

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

This very useful relation follows from the fact that

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

is an $N(0, 1)$ distributed random variable. The proof of this result goes as follows. Using the fact that $\sigma > 0$, we have

$$P(Z \leq z) = P(X \leq \mu + \sigma z) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu+\sigma z} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}\, dx.$$

By the substitution $w = (x - \mu)/\sigma$, we find

$$P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}w^2}\, dw,$$

proving the desired result. An immediate consequence of this result is

$$P(|X - \mu| > k\sigma) = 2 - 2\Phi(k) \quad \text{for any } k > 0.$$

The reader is asked to verify this very useful relation in Problem 10.35. The probability $P(|X - \mu| > k\sigma)$ has the values 0.3173, 0.0455 and 0.0027 for $k = 1$, 2 and 3. Nearly all the mass of the $N(\mu, \sigma^2)$ density is within three standard deviations of the expected value.

The normal distribution is the most important continuous distribution and has many applications. Although a normal random variable theoretically takes on values in the interval $(-\infty, \infty)$, it may still 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. The following properties contribute to the practical usefulness of the normal distribution model.

- If the random variable X has an $N(\mu, \sigma^2)$ distribution, then $aX + b$ is $N(a\mu + b, a^2\sigma^2)$ distributed for any constants a, b with $a \neq 0$.
- If X and Y are independent random variables that are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributed, then $X + Y$ is $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ distributed.

The first property is easily verified by working out the probability distribution of $Y = aX + b$, see Problem 10.36. The second property requires more advanced tools and will be proved in Section 14.2. In Chapter 14 we will also give a proof of the central limit theorem which is intrinsically linked up with the normal distribution. This theorem is the queen among the theorems in probability theory and reads as follows.

Rule 10.3 (*Central limit theorem*) *If $X_1, X_2, \ldots$ are independent and identically distributed random variables with expected value μ and standard deviation σ, then*

$$\lim_{n\to\infty} P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x) \qquad \text{for all } x.$$

In words, 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 and standard deviation σ has approximately a normal distribution with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$ for n large enough. How large n should be depends strongly on the distribution of the X_k. The first version of the central limit theorem was proved by Abraham de Moivre (1667–1754) for a special case of binomially distributed X_k. A binomial random variable X with parameters n and p can be written as the sum of n independent random variables $X_1, \ldots, X_n$, where $X_k = 1$ with probability p and $X_k = 0$ with probability $1 - p$. As a rule of thumb, the approximation of the binomial distribution by the normal distribution can be used when $np(1 - p) \geq 25$. Also, the Poisson distribution, being a limiting distribution of the binomial, can be approximated by the normal distribution when the parameter λ of the Poisson distribution satisfies $\lambda \geq 25$.

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. Suppose that the X_k are independent random variables having the expected values μ_k and the standard deviations σ_k. Then, under a weak regularity condition, the normalized random variable

$$\frac{\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k}{\sqrt{\sum_{k=1}^{n} \sigma_k^2}}$$

has also the standard normal distribution as limiting distribution. This generalized central limit theorem explains why data which are the result of many small unrelated effects are approximately normally distributed. In practice, many random phenomena, such as the rate of return on a stock, the cholesterol level of an adult male, the duration of a pregnancy, etc., are approximately normally distributed.

Several applications of the central limit theorem are given in Sections 5.4 to 5.9. The applications include the concept of confidence interval for outcomes of simulation runs and the Brownian motion process. The following two examples also illustrate the power of the central limit theorem.

Example 10.10 Your friend asserts to have rolled an average of 3.25 points per roll in 1,000 rolls of a fair die. Do you believe this?

Solution. Let the random variable X_i denote the outcome of the ith roll of the die. The random variables $X_1, \ldots, X_{1,000}$ are independent and uniformly distributed on the integers $1, \ldots, 6$, where $E(X_i) = 3.5$ and $\sigma(X_i) = 1.7078$. The total number of points to obtain in $n = 1{,}000$ rolls of a fair die is approximately normally distributed with expected value $3.5n$ and standard deviation

$1.7078\sqrt{n}$. Hence the average number points per roll in 1,000 rolls is approximately $N(\mu, \sigma^2)$ distributed with $\mu = 3.5$ and $\sigma = 1.7078/\sqrt{1{,}000} = 0.0540$. The value 3.25 lies $(3.5 - 3.25)/0.0540 = 4.629$ standard deviations below the expected value 3.5. The probability that an $N(\mu, \sigma^2)$ distributed random variable takes on a value 4.629 or more standard deviations below the expected value is $\Phi(-4.629) = 1.84 \times 10^{-7}$. The claim of your friend is highly implausible.

Example 10.11 Suppose that $X_1, \ldots, X_n$ are independent random variables that are uniformly distributed on (0,1). Approximate the probability that the rounded sum $X_1 + \cdots + X_n$ equals the sum of the rounded X_i when the rounding is to the nearest integer.[†]

Solution. Let the random variable D_i denote $D_i = X_i - \text{round}(X_i)$. Then the desired probability is given by

$$P(-\frac{1}{2} \leq S_n < \frac{1}{2}),$$

where $S_n = \sum_{i=1}^{n} D_i$. This is easily seen by noting that the rounded sum is equal to the sum of the rounded X_i only if

$$\sum_{i=1}^{n} \text{round}(X_i) - \frac{1}{2} \leq \sum_{i=1}^{n} X_i < \sum_{i=1}^{n} \text{round}(X_i) + \frac{1}{2}.$$

The random variables $D_1, \ldots, D_n$ are independent and uniformly distributed on $(-\frac{1}{2}, \frac{1}{2})$ and thus have an expected value of 0 and a standard deviation of $\sqrt{1/12}$. Using the central limit theorem, it follows that

$$P\left(-\frac{1}{2} \leq S_n < \frac{1}{2}\right) \approx \Phi\left(\sqrt{\frac{3}{n}}\right) - \Phi\left(-\sqrt{\frac{3}{n}}\right)$$

for n large enough. We find the approximate values 0.416, 0.271, 0.194, and 0.138 for $n = 10, 25, 50$ and 100. These approximations are excellent. The values obtained by one million simulation runs are 0.411, 0.269, 0.191, and 0.136. Also, the probability that the rounded sum of the amounts minus the sum of the rounded amounts is equal to the integer a can be approximated by using the central limit theorem. This probability is given by $P(a - \frac{1}{2} \leq S_n < a + \frac{1}{2})$ and can be approximated by $\Phi(\sqrt{12}(a + 0.5)/\sqrt{n}) - \Phi(\sqrt{12}(a - 0.5)/\sqrt{n})$. For $n = 25$, the approximate values are 0.215 and 0.108 for $a = 1$ and $a = 2$. The values obtained by one million simulation runs are 0.215 and 0.109.

[†] This example is based on R.B. Nelsen and J.E. Schultz, "The probability that the sum of rounds equals the round of the sum," *The College Mathematics Journal* **18** (1987): 390–396.

Example 9.15 (continued) How many bets must be placed by a player to have a probability of more than 99% of making a positive profit in the sales promotion of the lottery? What is the probability distribution of the net profit for a team of players who make 100,000 five-dollar bets?

Solution. In Example 9.15 in Section 9.6, we showed that the New York State Lottery offered the player a favorable game with the sales promotion "Big Dipper Wednesday". How could you take advantage of this offer? The answer is of course to bet as frequently as possible. The law of large numbers implies that the average profit per dollar bet will be close to 19.47 cents when a sufficiently large number of bets are made. But how many bets are necessary for the probability of a positive profit to be more than 99%? An answer to this question can be given with the help of the central limit theorem. In the game of Quick Draw a maximum of twenty individual bets can be made with a single game card. Each individual bet can be for 1, 2, 5, or 10 dollars. Let us assume that you only take betting action consisting of twenty individual $5 bets. Such a single betting action would cost $100. An individual $5 bet results in 0, 1, 2, 3, or 4 matching numbers with respective probabilities 0.308321, 0.432732, 0.212635, 0.043248, and 0.003063, where the corresponding payoffs are $0, $0, $10, $50, and $550 (see Example 9.15 in Section 9.6). The expected value and the standard deviation of the net profit resulting from an individual $5 bet are easily calculated as

$$\mu = 0.973615 \text{ dollars} \quad \text{and} \quad \sigma = 31.943315 \text{ dollars}.$$

The net profit resulting from n betting actions each consisting of twenty individual $5 bets can be seen as the sum of $20n$ independent random variables each having expected value μ and standard deviation σ. For n sufficiently large, this sum is approximately normal distributed with expected value $20n\mu$ and standard deviation $\sigma\sqrt{20n}$. The smallest value of n for which the probability of having a positive net profit after n betting actions is at least 99% follows by solving x from the equation

$$1 - \Phi\left(\frac{-20x\mu}{\sigma\sqrt{20x}}\right) = 0.99.$$

This leads to $20x\mu/(\sigma\sqrt{20x}) = 2.3264$ and so $x = 291.28$. Hence an estimated number of 292 betting actions are required to have a probability of at least 99% that the net profit is positive. Suppose a team of players could make 1,250 betting actions on each of the four Wednesdays (recall that a new game is played every four or five minutes), giving a total of 5,000 betting actions. By the central limit theorem, the probability distribution of net profit after 5,000 betting

actions (100,000 individual bets) can be very well approximated by a normal distribution with an expected value of $100{,}000\mu = 97{,}361.50$ dollars and a standard deviation of $\sigma\sqrt{100{,}000} = 10{,}101.36$ dollars. The 0.025th and 0.975th percentiles of this normal distribution are given by \$77,563 and \$117,160.[†] In other words, the net profit will be between about \$77,563 and \$117,160 with a 95% probability. The story is that students who had taken a probability course in college have applied their knowledge of probability to earn over \$100,000 playing the lottery game with double payoffs.

Problem 10.35 Verify that $P(|X - \mu| > k\sigma) = 2 - 2\Phi(k)$ for any $k > 0$ if X is $N(\mu, \sigma^2)$ distributed.

Problem 10.36 Verify that $aX + b$ is $N(a\mu + b, a^2\sigma^2)$ distributed if X is $N(\mu, \sigma^2)$ distributed.

Problem 10.37 Somebody claims to have obtained 5,250 heads in 10,000 tosses of a fair coin. Do you believe this claim?

Problem 10.38 Consider Example 9.3 again. Use the normal distribution to approximate the probability that Joe's profit after 52 weeks will be \$100 or more.

Problem 10.39 Repeat the analysis in Example 10.11 for the case that the X_i have the triangular density $f(x) = 2x$ for $0 < x < 1$ and $f(x) = 0$ otherwise.

Problem 10.40 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 $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

10.4.8 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

$$f(x) = \begin{cases} \frac{1}{\sigma x\sqrt{2\pi}}\, e^{-\frac{1}{2}[\ln(x)-\mu]^2/\sigma^2} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

[†] The pth percentile of a continuously distributed random variable X is given be any value x_p satisfying $P(X \le x_p) = p$ for $0 < p < 1$. The 0.5th percentile is usually called the median.

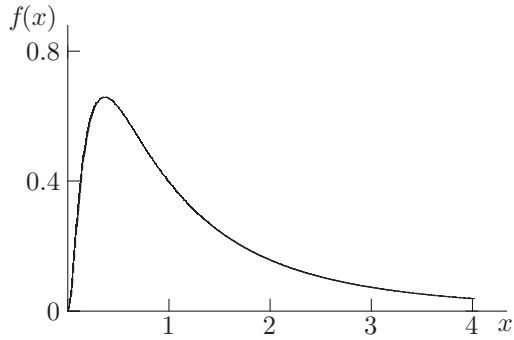


Fig. 10.7. Lognormal density ($\mu = 0, \sigma = 1$).

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 10.7. 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 10.12 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 follow closely a lognormal distribution. The lognormal distribution is also often used to model future stock prices after a long 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. This can be explained with the help of the central limit theorem, using the fact that the logarithm of a product of terms is the sum of the logarithms of the terms.

Problem 10.41 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?

10.4.9 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

$$f(x) = \frac{1}{2^{\frac{1}{2}d}\Gamma(\frac{1}{2}d)}\, x^{\frac{1}{2}d-1}\, e^{-\frac{1}{2}x} \qquad \text{for } x > 0$$

(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 $W = \sqrt{V_1^2 + \cdots + V_d^2}$ has the density function

$$f_W(w) = \frac{\sigma^{-d}}{2^{\frac{1}{2}d-1}\Gamma(\frac{1}{2}d)} w^{d-1} e^{-\frac{1}{2}w^2/\sigma^2} \qquad \text{for } w > 0.$$

The verification of this result is left as an exercise to the reader.

10.4.10 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

$$X = \frac{Z}{\sqrt{U/n}},$$

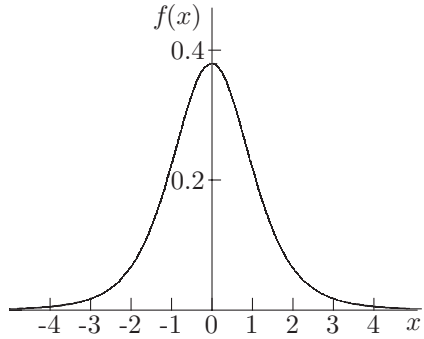


Fig. 10.8. Student-t density for $n = 5$.

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 $c = \frac{1}{\sqrt{\pi n}}\Gamma\left(\frac{1}{2}(n+1)\right) / \Gamma\left(\frac{1}{2}n\right)$. In Figure 10.8, 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 Guinness 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

$$E(X) = 0 \qquad \text{and} \qquad \text{var}(X) = \frac{n}{n-2} \quad \text{for } n > 2.$$

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 $\left(\overline{Y}(n) - \mu\right)/\sqrt{\overline{S}^2(n)/n}$ has the pleasant feature of being *robust*. This means that the statistic is not sensitive for small deviations from the normality assumption.

10.5 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 $h(r) = re^{-\frac{1}{2}r^2}$ 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 $H(r) = P(R \le r)$, 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 10.4 *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 $v(x)$.

[†] If the probability distribution function $F(x) = P(X \le x)$ of the continuous random variable X is strictly increasing and U is uniformly distributed on $(0, 1)$, then the random variable $V = F^{-1}(U)$ defined by the inverse function F^{-1} has the same distribution as X. This can be seen from $P(V \le x) = P(U \le F(x)) = F(x)$.

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 10.12 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 10.4 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 10.13 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

$$\frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}v^2/\sigma^2} \quad \text{for } v > 0.$$

The expected value of V is calculated as

$$E(V) = \int_0^\infty v \frac{2}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}v^2/\sigma^2}\, dv = \frac{-2\sigma^2}{\sigma\sqrt{2\pi}} \int_0^\infty de^{-\frac{1}{2}v^2/\sigma^2}$$
$$= \frac{-2\sigma}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2/\sigma^2}\Big|_0^\infty = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}.$$

Problem 10.42 Let the random variable X have the beta density $f(x) = 12x(1-x)^2$ for $0 < x < 1$ and $f(x) = 0$ otherwise. What is the probability density of $Y = X^2$? What are the expected value and the standard deviation of Y?

Problem 10.43 Let the variable Y be defined by $Y = 10^X$, where X is uniformly distributed on (0, 1). What is the probability density of Y? What are the expected value and the standard deviation of Y?

Simulating from continuous probability distributions

The discussion above touched upon the simulation of a random observation from a continuous probability distribution. The inverse-transformation method is a very useful method when the inversion of the probability distribution function requires little computational effort. For example, this is the case for the exponential distribution and the Weibull distribution, see Problem 10.44. However, for most probability distributions one has to resort to other methods in order to have low computing times. For the normal distribution there are several alternative methods. One of the most used methods is the *Box-Muller method* that will be discussed in Section 11.3. In more recent versions of the software tool Matlab the so-called *ziggurat method* is used for simulating random observations from the normal distribution. This sophisticated algorithm approximates the area under the standard normal curve with horizontal rectangles having equal areas; the first coordinate of a uniform point in a randomly chosen rectangle provides the required normal variate. Note that, by the result in Example 10.12, any simulation procedure for the normal distribution can also be used to simulate from the lognormal distribution.

Tailor-made algorithms to simulate from specific probability distributions are often designed on the basis of the *acceptance-rejection method*. This powerful method is based on a simple geometrical idea. A probability density $f(x)$

from which it is difficult to simulate directly is bounded by $cr(x)$, where c is a constant and $r(x)$ is a probability density from which it is easy to simulate. Details of this method will be given in Section 15.5 of Chapter 15. The acceptance-rejection method can equally well be used for both continuous and discrete distributions. In practice one uses this method to simulate from the gamma distribution and the Poisson distribution among others. Another general method to simulate from discrete probability distributions can be found in Section 2.9.

Problem 10.44 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 10.45 Give a simple method to simulate from the gamma distribution with parameters α and λ when the shape parameter α is equal to the integer n. *Hint*: the sum of n independent random variables each having an exponential density with parameter λ has a gamma density with parameters n and λ.

Problem 10.46 Let the probability density of continuous random variable X be given by the triangular density with parameters a, b and m, where $a < m \leq b$. Verify that you can simulate a random observation x from the triangular density by generating a random number u from $(0,\ 1)$ and setting $x := a + (b - a)\sqrt{m_0 u}$ if $u \leq m_0$ and $x := a + (b-a)[1 - \sqrt{(1-m_0)(1-u)}]$ if $u > m_0$, where $m_0 = (m-a)/(b-a)$. *Hint*: consider the normalized variable $V = (X-a)/(b-a)$ and invert the probability distribution function of V.

Problem 10.47 Suppose that the positive random variable X has as probability density the so-called hyperexponential density $p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}$ for $x > 0$, where $0 < p < 1$. How would you simulate the random variable X?

10.6 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 life time 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 agreement with the memoryless property discussed in Section 10.4.3. 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 10.4.5. 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 10.5 *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 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. The chapter treats joint and marginal densities, along with covariance and correlation. Also, the transformation rule for jointly distributed random variables and regression to the mean are discussed.

11.1 Joint probability mass function

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. The quantity $P(X = x,\ Y = y)$ is the probability assigned by P to the intersection of the two sets $A = \{\omega : X(\omega) = x\}$ and $B = \{\omega : Y(\omega) = y\}$, with ω representing an element of the sample space. Define the marginal probability mass functions of the

Table 11.1. *The joint probability mass function* $p(x, y)$.

x \ 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

random variables X and Y by

$$p_X(x) = P(X = x) \quad \text{and} \quad p_Y(y) = P(Y = y).$$

The marginal probability mass functions can be obtained from the joint probability mass function by

$$p_X(x) = \sum_y P(X = x,\ Y = y), \quad p_Y(y) = \sum_x P(X = x,\ Y = y).$$

These relations follow from the result that $P(A) = \sum_{i=1}^{n} P(A_i)$ if the event A is the union of mutually exclusive events $A_1, A_2, \ldots, A_n$.

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 a 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?

Problem 11.3 You choose three different numbers at random from the numbers $1, 2, \ldots, 10$. Let X be the smallest of these three numbers and Y the largest. What is the joint probability mass function of X and Y? What are the marginal distributions of X and Y and what is the probability mass function of $Y - X$?

Problem 11.4 You repeatedly draw a number at random from the numbers $1, 2, \ldots, 10$. Let X be the number of draws until the number 1 appears and Y the number of draws until the number 10 appears. What is the joint probability mass function of X and Y? What are the probability mass functions of $\min(X, Y)$ and $\max(X, Y)$?

Problem 11.5 You repeatedly toss two fair coins until both coins show heads. Let X and Y denote the number of heads resulting from the tosses of the first and the second coin respectively. What is the joint probability mass function of X and Y and what are the marginal distributions of X and Y? What is $P(X = Y)$? *Hint*: evaluate $P(X = i,\ Y = j,\ N = n)$, where N is the number of tosses until both coins show heads. Use the identity $\sum_{k=0}^{\infty} \binom{m+k}{m} x^k = 1/(1-x)^{m+1}$ for $|x| < 1$.

11.2 Joint probability density function

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

$$P\,(X \le a,\ Y \le b) = \frac{b}{2\pi}\frac{a^2}{r^2} \qquad \text{for } 0 \le a \le r \text{ and } 0 \le b \le 2\pi.$$

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 a same sample space with probability measure P. The *joint cumulative probability distribution function* of X and Y is defined by $P(X \le x,\ Y \le y)$ for all x, y, where $P(X \le x,\ Y \le y)$ is a shorthand for $P(\{\omega :\ X(\omega) \le x \text{ and } Y(\omega) \le y\})$ 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 $P(X \le a,\ Y \le b)$ allows for the representation*

$$P(X \le a,\ Y \le b) = \int_{x=-\infty}^{a}\int_{y=-\infty}^{b} f(x,\ y)\,dx\,dy, \qquad -\infty < a, b < \infty,$$

where the function $f(x, y)$ satisfies

$$f(x, y) \ge 0 \qquad \textit{for all } x, y \quad \textit{and} \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dxdy = 1.$$

Just as in the one-dimensional case, $f(a, b)$ allows for the interpretation:

$$f(a, b)\,\Delta a\,\Delta b \approx P\left(a - \frac{1}{2}\Delta a \le X \le a + \frac{1}{2}\Delta a, b - \frac{1}{2}\Delta b \le Y \le b + \frac{1}{2}\Delta b\right)$$

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 $P(X \le x,\ Y \le y)$ with respect to x and y. It then follows from

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} P(X \le x,\ Y \le y)$$

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 $f(x, y)\Delta x \Delta y$ provided that Δx and Δy are very 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, and 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 the spins,

$$\begin{aligned} &P(x < X \leq x + \Delta x,\ y < Y \leq y + \Delta y) \\ &= 6\frac{(x + \Delta x - x)}{2\pi r}\frac{(y + \Delta y - y)}{2\pi r}\frac{(y - x)}{2\pi r}. \end{aligned}$$

Hence, the joint probability density function of X and Y is given by

$$f(x, y) = \begin{cases} \frac{6(y-x)}{(2\pi r)^3} & \text{for } 0 < x < y < 2\pi r \\ 0 & \text{otherwise.} \end{cases}$$

The following general rule applies to any two random variables X and Y having a joint probability density function $f(x, y)$.

Rule 11.1 *For any neat region C in the two dimensional plane*,

$$P((X, Y) \in C) = \iint_C f(x, y)\, dx\, dy.$$

This is very useful result. It is important to note that 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. Rule 11.1 can be used to determine the probability distribution function of any function $g(X, Y)$ of X and Y. To illustrate this, we show 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

$$P(Z \le z) = \iint_{\substack{(x, y):\\ x+y \le z}} f(x, y)\, dx\, dy = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{z-x} f(x, y)\, dx\, dy$$
$$= \int_{v=-\infty}^{z} \int_{u=-\infty}^{\infty} f(u, v - u)\, du\, dv,$$

using the change of variables $u = x$ and $v = x + y$. Next, differentiation of $P(Z \le z)$ yields the convolution formula for $f_Z(z)$. If the random variables X and Y are nonnegative, the convolution formula reduces to

$$f_Z(z) = \int_0^z f(u, z - u)\, du \qquad \text{for } z > 0.$$

Another illustration of Rule 11.1 is given in the following two examples.

Example 11.4 An electronic system has two crucial components. The electronic system goes down if either of two components fails. The joint density function of the lifetimes X and Y of two components is given by $f(x, y) = \frac{1}{12}(1 + x + y)$ for $0 < x, y < 2$ and $f(x, y) = 0$ otherwise, where the lifetimes are measured in years. What is the probability that the electronic system will go down in its first year of operation?

Solution. We are asking for the probability $P(\min(X, Y) \le 1)$. To compute this probability, consider the complementary probability $P(\min(X, Y) > 1)$ and note that $\min(X, Y) > 1$ only if both $X > 1$ and $Y > 1$. Applying

Rule 11.1, we find that $P(\min(X, Y) \geq 1)$ is given by

$$\int_1^2 \int_1^2 \frac{1}{12}(1 + x + y)\, dx\, dy = \frac{1}{12} \int_1^2 dx \int_1^2 (1 + x + y)\, dy$$
$$= \frac{1}{12} \int_1^2 (1 + x + 3/2)\, dx = \frac{1}{3}.$$

Hence the desired probability is $\frac{2}{3}$.

Example 11.5 Let the random variable X be a number chosen at random from $(0, 1)$ and Y a number chosen at random from $(0, X)$. Take X and Y as the lengths of the sides of a rectangle. What is the probability density of the area of the rectangle?

Solution. Let $f(x, y)$ denote the joint probability density of the random variables X and Y. Then, $f(x, y) = \frac{1}{x}$ for $0 < x < 1$ and $0 < y < x$; otherwise, $f(x, y) = 0$ (verify this by evaluating $P(x < X \leq x + \Delta x,\ y < Y \leq y + \Delta y)$). Denote the random variable Z by $Z = XY$. By Rule 11.1,

$$P(Z \leq z) = \iint\limits_{\substack{(x, y):\\ xy \leq z}} f(x, y)\, dx\, dy = \int_0^1 dx \int_0^{\min(z/x,\, x)} \frac{1}{x}\, dy$$
$$= \int_0^{\sqrt{z}} dx \int_0^x \frac{1}{x}\, dy + \int_{\sqrt{z}}^1 dx \int_0^{z/x} \frac{1}{x}\, dy.$$

This leads to $P(Z \leq z) = 2\sqrt{z} - z$ for $0 \leq z \leq 1$. Hence the probability density $f(z)$ of the area of the rectangle is $f(z) = \frac{1}{\sqrt{z}} - 1$ for $0 < z < 1$ and $f(z) = 0$ otherwise.

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 $f(x, y)$.

Problem 11.6 Let the joint probability density function of the random variables X and Y be given by $f(x, y) = cxe^{-2x(1+y)}$ for $x, y > 0$ and $f(x, y) = 0$ otherwise. Determine the constant c. What is the probability density of XY?

Problem 11.7 Let the joint probability density function of the random variables X and Y be given by $f(x, y) = c\sqrt{x + y}$ for $0 < x, y < 1$ and $f(x, y) = 0$ otherwise. Determine the constant c. What is the probability density of $X + Y$?

Problem 11.8 Let the joint probability density function of the random variables X and Y be given by $f(x, y) = ce^{-2x}$ for $0 < y \leq x < \infty$ and $f(x, y) = 0$ otherwise. Determine the constant c. What is the probability density of $X - Y$?

Problem 11.9 Independently of each other, two points are chosen at random in the interval (0, 1). Use the solution method of Example 11.3 to answer the following questions. 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 11.10 Let X and Y be two random variables with a joint probability density $f(x, y) = \frac{1}{(x+y)^3}$ for $x, y > c$ and $f(x, y) = 0$ otherwise. Determine the constant c. What is the probability density of $\min(X, Y)$?

Problem 11.11 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 $V = \max(X, Y)$, $W = \min(X, Y)$ and $Z = V - W$. *Hint*: use the fact that $Z = |X - Y|$.

11.3 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

$$P(X \le a) = \lim_{b\to\infty} P(X \le a,\ Y \le b) = \int_{-\infty}^{a} \left[\int_{-\infty}^{\infty} f(x, y)\, dy \right] dx.$$

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 11.6 A point (X, Y) is chosen at random inside the unit circle. What is the marginal density of X?

Solution. Denote by $C = \{(x, y) \mid x^2 + y^2 \le 1\}$ 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 $P(x < X \le x + \Delta x)$ as the area of a vertical strip in the unit circle.

Problem 11.12 The joint density of the random variables X and Y is given by $f(x, y) = 4xe^{-2x(1+y)}$ for $x, y > 0$. What are the marginal densities of X and Y?

Problem 11.13 Let the joint density function $f(x, y)$ of the random variables X and Y be equal to $4e^{-2x}$ for $0 < y \leq x < \infty$ and 0 otherwise. What are the marginal densities of X and Y?

Problem 11.14 Let the random variable X be the portion of a flood insurance claim for flooding damage to the house and Y the portion of the claim for flooding damage to the rest of the property. The joint density function of X and Y is given by $f(x, y) = 3 - 2x - y$ for $0 < x, y < 1$ and $x + y < 1$. What are the marginal densities of X and Y?

Problem 11.15 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 $x = \frac{1}{2}$?

Problem 11.16 The joint density of the random variables X and Y is given by $f(x, y) = \frac{c}{x}$ for (x, y) in the triangle with vertices $(0, 0)$, $(1, 0)$ and $(1, 1)$ and $f(x, y) = 0$ otherwise. Find the value of the constant c. What are the marginal densities of X and Y?

11.3.1 Independence of jointly distributed random variables

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 11.2 *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.

In the following example we give a very important result for the exponential distribution.

Example 11.7 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. It will be shown that

$$P(\min(X, Y) \le z) = 1 - e^{-(\alpha+\beta)z} \quad \text{for } z \ge 0 \quad \text{and} \quad P(X < Y) = \frac{\alpha}{\alpha + \beta}.$$

In other words, $\min(X, Y)$ is exponentially distributed with expected value $1/(\alpha + \beta)$. The proof is simple. Noting that $P(\min(X, Y) \le z) = 1 - P(X > z,\ Y > z)$, we have

$$P(\min(X, Y) \le z) = 1 - \int_{x=z}^{\infty} \int_{y=z}^{\infty} f_X(x) f_Y(y)\,dx\,dy.$$

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.

A useful relation can be given between the exponential distribution and the normal distribution.

Example 11.8 Let X and Y be two independent random variables each having the standard normal distribution. What is the probability distribution function of $X^2 + Y^2$? How could you use this result to prove that the variance of an $N(\mu, \sigma^2)$ distributed random variable is σ^2?

Solution. Since X and Y are independent random variables with densities $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ and $f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$, the joint density of X and Y is given by

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

For any given $z > 0$,

$$P(X^2 + Y^2 \le z) = \iint_D f(x, y)\, dx\, dy = \iint_D \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}\, dx\, dy,$$

where $D = \{(x, y) : x^2 + y^2 \le z\}$ is a disk of radius $\sqrt{z}$. In the Cartesian (rectangular) coordinates x and y the disk D is described by the inequalities $-\sqrt{z} \le x \le \sqrt{z}$ and $-\sqrt{z - x^2} \le y \le \sqrt{z - x^2}$. This gives an unpleasant integral to compute. However, a disk of radius $\sqrt{z}$ can be defined in polar coordinates r and θ by $0 \le \theta \le 2\pi$ and $0 \le r \le \sqrt{z}$, where r (the distance) and θ (the angle) are defined by $x = r\cos(\theta)$ and $= r\sin(\theta)$. When changing to polar coordinates, the integral gets constant limits of integration and becomes easier to compute. The term $x^2 + y^2$ is converted into $r^2\cos^2(\theta) + r^2\sin^2(\theta) = r^2$, using the identity $\cos^2(\theta) + \sin^2(\theta) = 1$. Some care is required with $dx\, dy$ when changing to polar coordinates. An infinitesimal small area element $\Delta x\, \Delta y$ in Cartesian coordinates is not converted into $\Delta r\, \Delta\theta$ but becomes $r\, \Delta r\, \Delta\theta$ in polar coordinates. Thus, by changing to polar coordinates,

$$\begin{aligned} P(X^2 + Y^2 \le z) &= \iint_D \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}\, dx\, dy \\ &= \int_0^{2\pi} \int_0^{\sqrt{z}} \frac{1}{2\pi} e^{-\frac{1}{2}r^2}\, r\, dr\, d\theta = \int_0^{\sqrt{z}} e^{-\frac{1}{2}r^2}\, r\, dr. \end{aligned}$$

By the change of variable $u = r^2$, the integral $\int_0^{\sqrt{z}} e^{-\frac{1}{2}r^2}\, r\, dr$ can be computed as $\int_0^z \frac{1}{2} e^{-\frac{1}{2}u}\, du = 1 - e^{-\frac{1}{2}z}$. Hence

$$P(X^2 + Y^2 \le z) = 1 - e^{-\frac{1}{2}z} \qquad \text{for } z > 0.$$

That is, $X^2 + Y^2$ has an exponential distribution with expected value 2. This result implies that $E(X^2) = 1$ for an $N(0, 1)$ distributed random variable and so $E[(\frac{V-\mu}{\sigma})^2] = 1$ for an $N(\mu, \sigma^2)$ distributed random variable V. This in turn gives $E[(V - \mu)^2] = \sigma^2$, proving that an $N(\mu, \sigma^2)$ distributed random variable has variance σ^2.

The above result that $X^2 + Y^2$ has an exponential distribution with parameter $\lambda = \frac{1}{2}$ if X and Y are independent $N(0, 1)$ random variables is useful for

simulation purposes. It can be used to simulate a random observation from a chi-square random variable. Recall that chi-square random variable with n degrees of freedom is distributed as $X_1^2 + X_2^2 + \cdots + X_n^2$ with $X_1, X_2, \ldots, X_n$ independent $N(0, 1)$ random variables, see Section 10.4.9. Using Example 10.4, we have that $-2\ln(u)$ is a random observation from the exponential distribution with mean 2 when u is a random number from (0, 1). Hence, for $n = 2k$, it follows that $-2[\ln(u_1) + \cdots + \ln(u_k)] = -2\ln(u_1 \times \cdots \times u_k)$ is a random observation from the chi-square distribution with $2k$ degrees of freedom when $u_1, \ldots, u_k$ are random numbers from (0, 1) that are independently generated from each other. For the case that $n = 2k + 1$, we need an additional random observation from the standard normal distribution.

Problem 11.17 Let X and Y be two independent random variables each having the uniform distribution on (0, 1). Take X and Y as the lengths of the sides of a rectangle. What are the probability density and the expected value of the area of the rectangle?

Problem 11.18 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**) What is the probability density of Z?
(**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?

Problem 11.19 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 $V = X^2 + Y^2$ when $c = 0$ and $d = 1$? Use the latter density to calculate the expected value of the distance of a point chosen at random inside the unit square to any vertex of the unit square.

11.3.2 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 the following.

Rule 11.3 *If the random variables X and Y have a joint probability density function $f(x, y)$, then*

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy$$

for any function $g(x, y)$ provided that the integral is well-defined.

An easy consequence of Rule 11.3 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

$$\begin{aligned}
&\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by) f(x, y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} axf(x, y)\, dx\, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} byf(x, y)\, dx\, dy \\
&= \int_{x=-\infty}^{\infty} ax\, dx \int_{y=-\infty}^{\infty} f(x, y)\, dy + \int_{y=-\infty}^{\infty} by\, dy \int_{x=-\infty}^{\infty} f(x, y)\, dx \\
&= a \int_{-\infty}^{\infty} x f_X(x)\, dx + b \int_{-\infty}^{\infty} y f_Y(y)\, dy,
\end{aligned}$$

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 2.17: 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 $0 < x, y < 1$. By the substitution rule, $E(|X - Y|) = \int_0^1 \int_0^1 |x - y|\, dx\, dy$ and so

$$\begin{aligned}
E(|X - Y|) &= \int_0^1 dx \left[\int_0^x (x - y)\, dy + \int_x^1 (y - x)\, dy \right] \\
&= \int_0^1 \left[\frac{1}{2}x^2 + \frac{1}{2} - \frac{1}{2}x^2 - x(1 - x) \right] dx = \frac{1}{3}.
\end{aligned}$$

Hence, the answer to the question is $\frac{1}{3}$.

As another illustration of Rule 11.3, 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

$$\begin{aligned} E[g(X,Y)] &= \int_0^r \int_0^{2\pi} \{x|\cos(y)| + x|\sin(y)|\}\frac{x}{\pi r^2}\,dx\,dy \\ &= \frac{1}{\pi r^2}\int_0^{2\pi} |\cos(y)|\,dy \int_0^r x^2\,dx + \frac{1}{\pi r^2}\int_0^{2\pi} |\sin(y)|\,dy \int_0^r x^2\,dx \\ &= \frac{r^3}{3\pi r^2}\left[\int_0^{2\pi} |\cos(y)|\,dy + \int_0^{2\pi} |\sin(y)|\,dy\right] = \frac{8r}{3\pi}. \end{aligned}$$

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

11.4 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 11.4 *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 11.4 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 11.2 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 11.4, 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 11.2 again, it now follows that $V = X + Y$ and $W = X - Y$ are $N(0, 2)$ distributed and independent.

11.4.1 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 11.2, 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

$$f_V(v) = \frac{1}{2\pi}\int_0^{2\pi} ve^{-\frac{1}{2}v^2}\,dw = ve^{-\frac{1}{2}v^2}, \qquad 0 < v < \infty$$

and

$$f_W(w) = \frac{1}{2\pi}\int_0^{\infty} ve^{-\frac{1}{2}v^2}\,dv = \frac{1}{2\pi}, \qquad 0 < w < 2\pi.$$

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 10.5 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 11.20 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 11.21 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 11.22 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 $(0, \infty)$.

11.5 Covariance and correlation coefficient

Let the random variables X and Y be defined on a 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. By $\text{var}(X + Y) = E[\{X + Y - E(X + Y)\}^2]$

and the linearity of the expectation operator, $\text{var}(X+Y)$ can be written as

$$E[(X-E(X))^2 + 2(X-E(X))(Y-E(Y)) + (Y-E(Y))^2].$$

Using the linearity of the expectation operator again, we next find that

$$\text{var}(X+Y) = \text{var}(X) + 2E\big[\big(X-E(X)\big)\big(Y-E(Y)\big)\big] + \text{var}(Y).$$

This leads to the following general definition.

Definition 11.2 *The covariance* $\text{cov}(X,Y)$ *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 $\text{cov}(X,Y)$ 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 $\text{cov}(X,Y)$ has as direct consequence:

Rule 11.5 *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, $\text{cov}(X,Y)$ 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 $\text{var}(X+Y)$, we find the general rule:

Rule 11.6 *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*

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y).$$

The units of $\text{cov}(X, Y)$ 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

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}},$$

provided that $\text{var}(X) > 0$ and $\text{var}(Y) > 0$. 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 11.24. 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 $\rho(X, Y) = 1$, 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 11.23 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 11.24 Verify that

$$\text{var}(aX + b) = a^2\text{var}(X) \quad \text{and} \quad \text{cov}(aX, bY) = ab\text{cov}(X, Y)$$

for any constants a, b. Next, evaluate the variance of the random variable $Z = Y/\sqrt{\text{var}(Y)} - \rho(X, Y)X/\sqrt{\text{var}(X)}$ to prove that $-1 \leq \rho(X, Y) \leq 1$. Also, for any constants a, b, c, and d, verify that $\text{cov}(aX + bY, cV + dW)$ can be worked out as $ac\text{cov}(X, V) + ad\text{cov}(X, W) + bc\text{cov}(Y, V) + bd\text{cov}(Y, W)$.

Problem 11.25 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 millimeters 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 millimeters. 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 millimeters.

Problem 11.26 Let the random variables $X_1, \ldots, X_n$ be defined on a common probability space. Prove that

$$\text{var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \text{var}(X_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{cov}(X_i, X_j).$$

Next, evaluate $\text{var}(\sum_{i=1}^{n} t_i X_i)$ 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 σ_{ij}=cov(X_i, X_j). In other words, the covariance matrix $C = (\sigma_{ij})$ is positive semi-definite.

Problem 11.27 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 ball selected is red and X_i equals 0 otherwise.

Problem 11.28 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 11.29 You roll a pair of dice. What is the correlation coefficient of the high and low points rolled?

Problem 11.30 What is the correlation coefficient of the Cartesian coordinates of a point picked at random in the unit circle?

Problem 11.31 Let X be a randomly chosen integer from $1, \ldots, 100$ and Y a randomly chosen integer from $1, \ldots, X$. What is the correlation coefficient $\rho(X, Y)$?

Problem 11.32 Let X be a randomly chosen number from the interval $(0,\ 1)$ and Y a randomly chosen number from the interval $(0,\ X)$. What is the correlation coefficient $\rho(X, Y)$?

Problem 11.33 You generate three random numbers from $(0, 1)$. Let X be the smallest of these three numbers and Y the largest. What is the correlation coefficient $\rho(X, Y)$?

11.5.1 Linear predictor and regression

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

$$E\left[(Y - \alpha - \beta X)^2\right]$$

minimal? The answer to this question is

$$y = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X),$$

where $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X = \sqrt{\text{var}(X)}$, $\sigma_Y = \sqrt{\text{var}(Y)}$, 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

$$\begin{aligned}
&E[\{Y - \mu_Y - \beta(X - \mu_X) + \mu_Y - \alpha - \beta\mu_X\}^2] \\
&= E[\{Y - \mu_Y - \beta(X - \mu_X)\}^2] + (\mu_Y - \alpha - \beta\mu_X)^2 \\
&\quad + 2(\mu_Y - \alpha - \beta\mu_X)E[Y - \mu_Y - \beta(X - \mu_X)] \\
&= \sigma_Y^2 + \beta^2\sigma_X^2 - 2\beta\rho_{XY}\sigma_X\sigma_Y + (\mu_Y - \alpha - \beta\mu_X)^2.
\end{aligned}$$

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

$$E\left[(Y - \alpha - \beta X)^2\right] = \sigma_Y^2(1 - \rho_{XY}^2).$$

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.

The purpose of this chapter is to give a first introduction to the multivariate normal distribution and the multidimensional central limit theorem. Several practical applications will be discussed, including the drunkard's walk in higher dimensions and the chi-square test.

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 the bivariate normal density $f(x, y)$. By $E(X) = E(Y) = 0$, we have $\text{cov}(X, Y) = E(XY)$ and so $\text{cov}(X, Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x, y)\, dx\, dy$. Letting $\tau^2 = (1-\rho^2)$, it now follows that

$$\begin{aligned}\text{cov}(X, Y) &= \int_{x=-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\, dx \int_{y=-\infty}^{\infty} y\frac{1}{\tau\sqrt{2\pi}}\, e^{-\frac{1}{2}(y-\rho x)^2/\tau^2}\, dy \\ &= \int_{-\infty}^{\infty} \rho x^2 \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}\, dx = \rho,\end{aligned}$$

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

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]/(1-\rho^2)}.$$

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 11.2, 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. The following characterization of the bivariate normal distribution can be given.

Rule 12.3 *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.*[†]

[†] 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.

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 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 Let the random variables X and Y have a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, and let $\Phi(x)$ denote the standard normal distribution function. Use the decomposition of the standard bivariate normal density to verify that $P(X \le a,\ Y \le b)$ can be calculated as

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(a-\mu_1)/\sigma_1} \Phi\left(\frac{(b-\mu_2)/\sigma_2 - \rho x}{\sqrt{1-\rho^2}} \right) e^{-\frac{1}{2}x^2}\, dx.$$

Next verify that $P(X \le \mu_1,\ Y \le \mu_2) = \frac{1}{4} + \frac{1}{2\pi}\text{arctg}(a)$ with $a = \rho/\sqrt{1-\rho^2}$.

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

$$S_{n1} = X_1 + \cdots + X_n \quad \text{and} \quad S_{n2} = Y_1 + \cdots + Y_n.$$

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 for one-dimensional random variables (see Section 10.4.7) will not be surprised to learn that the two-dimensional version of the central limit theorem applies to the random vector $(S_{n1}, S_{n2}) = (X_1 + \cdots + X_n, Y_1 + \cdots + Y_n)$. In general form, the two-dimensional version of the central limit theorem

reads as:

$$\lim_{n\to\infty} P\left(\frac{S_{n1}-n\mu_1}{\sigma_1\sqrt{n}} \le x,\ \frac{S_{n2}-n\mu_2}{\sigma_2\sqrt{n}} \le y\right)$$
$$= \frac{1}{2\pi\sqrt{(1-\rho^2)}} \int_{-\infty}^{x}\int_{-\infty}^{y} e^{-\frac{1}{2}(v^2-2\rho vw+w^2)/(1-\rho^2)}\, dv\, dw,$$

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

$$\sigma_1^2 = E\left[\cos^2(\Theta)\right] = \int_0^{2\pi} \cos^2(\theta) f(\theta)\, d\theta = \frac{1}{2\pi}\int_0^{2\pi} \cos^2(\theta)\, d\theta.$$

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 trigonometry, we obtain $\sigma_1^2 + \sigma_2^2 = 1$. Hence, for reasons of symmetry, $\sigma_1^2 = \sigma_2^2 = \frac{1}{2}$. Finally,

$$\operatorname{cov}(X_1, Y_1) = E\left[(\cos(\Theta) - 0)(\sin(\Theta) - 0)\right] = \frac{1}{2\pi}\int_0^{2\pi} \cos(\theta)\sin(\theta)\, d\theta.$$

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 $0 \le x \le \frac{\pi}{2}$ and $\pi \le x \le \frac{3\pi}{2}$. This verifies that $\rho = 0$.

Next we can formulate two interesting results using the two-dimensional central limit theorem. The first result states that

$$P\left(S_{n1} \le x,\ S_{n2} \le y\right) \approx \frac{1}{\pi n}\int_{-\infty}^{x}\int_{-\infty}^{y} e^{-(t^2+u^2)/n}\, dt\, du$$

for n large. In other words, the position of the particle after n steps has approximately the bivariate normal density function

$$\phi_n(x, y) = \frac{1}{\pi n} e^{-(x^2+y^2)/n}$$

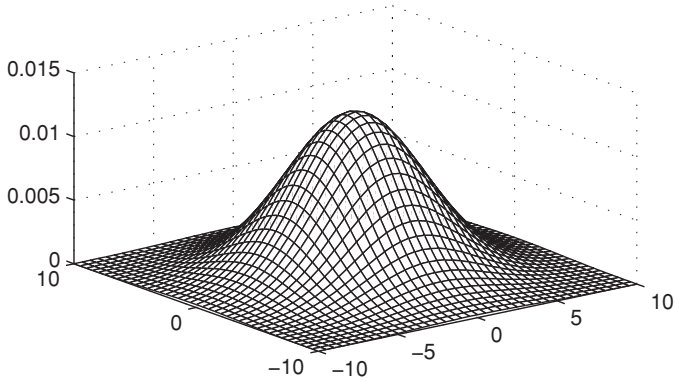


Fig. 12.1. The density of the particle's position after 25 steps.

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.

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

$$P(S_{n1} \le x,\ S_{n2} \le y) \approx \frac{1}{2\pi} \int_{-\infty}^{x/\sqrt{n/2}} \int_{-\infty}^{y/\sqrt{n/2}} e^{-\frac{1}{2}(v^2+w^2)}\, dv\, dw$$

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 11.3 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

$$\begin{aligned}
&\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sqrt{x^2+y^2} e^{-(x^2+y^2)/n}\, dx\, dy \\
&= \int_0^{\infty} \int_0^{2\pi} \sqrt{r^2\cos^2(\theta) + r^2\sin^2(\theta)} e^{-(r^2\cos^2(\theta)+r^2\sin^2(\theta))/n} r\, dr\, d\theta \\
&= \int_0^{\infty} \int_0^{2\pi} r^2 e^{-r^2/n}\, dr\, d\theta = 2\pi \int_0^{\infty} r^2 e^{-r^2/n}\, dr.
\end{aligned}$$

Obviously,

$$\begin{aligned}
\int_0^{\infty} r^2 e^{-r^2/n}\, dr &= -\frac{n}{2} \int_0^{\infty} r de^{-r^2/n} = -\frac{n}{2} r e^{-r^2/n} \Big|_0^{\infty} + \frac{n}{2} \int_0^{\infty} e^{-r^2/n}\, dr \\
&= \frac{1}{2}\frac{n}{2}\sqrt{n/2}\sqrt{2\pi} \int_{-\infty}^{\infty} \frac{1}{\sqrt{n/2}\sqrt{2\pi}} e^{-\frac{1}{2}r^2/(\frac{1}{2}n)}\, dr = \frac{1}{4} n\sqrt{n\pi},
\end{aligned}$$

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

$$P(D_n \le u) \approx \frac{2}{n} \int_0^u r e^{-r^2/n}\, dr \qquad \text{for } u > 0.$$

Hence, $P(D_n \leq u) \approx 1 - e^{-u^2/n}$ 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 $\frac{4}{\sqrt{2\pi n}} e^{-\frac{1}{4}r^2/n} \left[\Phi(\frac{r}{\sqrt{n}}) - \Phi(\frac{-r}{\sqrt{n}})\right]$ 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:

$$P(D_n \leq u) \approx \int_0^{u/\sqrt{n}} \frac{d^{\frac{1}{2}d}}{2^{\frac{1}{2}d-1}\Gamma(\frac{1}{2}d)} e^{-\frac{1}{2}dr^2} r^{d-1}\, dr \quad \text{for } u > 0,$$

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 10.4.9). It is matter of some algebra to derive from the approximate density of D_n that

$$E(D_n) \approx \frac{\alpha_d}{d^{\frac{1}{2}} 2^{\frac{1}{2}d-1}\Gamma(\frac{1}{2}d)} \sqrt{n},$$

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

$$\alpha_m = (m-1)\alpha_{m-2} \quad \text{for } m = 2, 3, \ldots$$

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

$$\mu_i = E(X_i) \quad \text{and} \quad \sigma_{ij} = \operatorname{cov}(X_i, X_j).$$

Note that $\sigma_{ii} = \operatorname{var}(X_i)$. 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 11.26).

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 $Z_1, \ldots, Z_d$ are independent random variables each having the standard normal distribution (the same $Z_1, \ldots, Z_d$ apply to each of the S_i). 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*

$$\mu = (n\mu_1^{(0)}, \ldots, n\mu_d^{(0)}) \quad \textit{and} \quad \mathbf{C} = (n\sigma_{ij}^{(0)})$$

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 large n. 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 $X_{1A}, \ldots, X_{nA}$,

$$\sigma^2(S_{nA}) = \sigma^2\left(\sum_{i=1}^{n} X_{iA}\right) = \sum_{i=1}^{n} \sigma^2(X_{iA}) = \sum_{i=1}^{n} p_{iA}(1 - p_{iA}).$$

Noting that $\operatorname{cov}(S_{nA}, S_{nB}) = E[(\sum_{i=1}^{n} X_{iA})(\sum_{j=1}^{n} X_{jB})] - E(\sum_{i=1}^{n} X_{iA}) \times E(\sum_{j=1}^{n} X_{jB})$, it follows from the independence of X_{iA} and X_{jB} for $j \neq i$ and the fact that X_{iA} and X_{iB} cannot both be positive that

$$\begin{aligned}\operatorname{cov}(S_{nA}, S_{nB}) &= \sum_{i=1}^{n} \sum_{j \neq i} E(X_{iA} X_{jB}) - \left(\sum_{i=1}^{n} p_{iA}\right)\left(\sum_{j=1}^{n} p_{jB}\right) \\ &= \sum_{i=1}^{n} \sum_{j \neq i} p_{iA} p_{jB} - \sum_{i=1}^{n} \sum_{j=1}^{n} p_{iA} p_{jB} = -\sum_{i=1}^{n} p_{iA} p_{iB}.\end{aligned}$$

Similarly, the other terms in matrix $\mathbf{C}$ are explained.

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

$$S_A = \text{the number of votes on party } A \text{ when the 1,000 voters actually vote on election day.}$$

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

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)

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 $\mathbf{U}\mathbf{D}^{1/2}$ 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

$$\begin{aligned} &P(\text{party } A \text{ becomes the largest party}) = 0.123\,(\pm 0.002) \\ &P(\text{party } B \text{ becomes the largest party}) = 0.877\,(\pm 0.002) \\ &P(\text{parties } A \text{ and } C \text{ get the majority of the votes}) = 0.855\,(\pm 0.002), \end{aligned}$$

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 3.7 for more details on this powerful method). Performing 100,000 simulation runs, we obtain the values 0.120 ($\pm$ 0.002), 0.872 ($\pm$ 0.002), and 0.851 ($\pm$ 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

$$\mu_j^{(0)} = \frac{r}{s},\ \sigma_{jj}^{(0)} = \frac{r}{s}(1 - \frac{r}{s}) \quad \text{and} \quad \sigma_{ij}^{(0)} = -\frac{r(s-r)}{s^2(s-1)} \qquad \text{for } i \neq j.$$

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 $\max_{1 \leq j \leq s} S_{nj} - \min_{1 \leq j \leq s} S_{nj}$ 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 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.

In the above discussion, we assumed that the probabilities $p_1, \ldots, p_d$ of the possible outcomes $1, \ldots, d$ are beforehand known. In applications, however, you have sometimes to estimate one or more parameters in order to get the probabilities $p_1, \ldots, p_d$ (see the Problems 12.14 and 12.15). Then you must lower the number of degrees of freedom of the chi-square test statistic by one for each parameter estimated from the data.

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 In Von Bortkiewicz's classical study on the distribution of 196 soldiers kicked to death by horses among 14 Prussian cavalry corps over the 20 years 1875 to 1894, the data are as follows. In 144 corps-years no deaths occurred, 91 corps-years had one death, 32 corps-years had two deaths, 11 corps-years had three deaths, and two corps-years had four deaths. Use a chi-square test to investigate how closely the observed frequencies conform to the expected Poisson frequencies. Assume that the probability of death is constant for all years and corps.

Problem 12.15 A total of 64 games were played during the World Cup Soccer 2010 in South Africa. The number of goals per game was distributed as follows. There were seven games with zero goals, 17 games with one goal, 13 games with two goals, 14 games with three goals, seven games with four goals, five games with five goals, and one game with seven goals. Use the chi-square test to find whether the distribution of the number of goals per game can be described by a Poisson distribution.

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.

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.

Problem 12.17 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 two winners of the jackpot, six winners of the second prize, nine 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.

13

Conditioning by random variables

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. The law of conditional expectation and several practical applications of this law are discussed. In the final section we explain Bayesian inference for continuous models and give several statistical applications.

13.1 Conditional distributions

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? We first answer this question for the discrete case. Conditioning on a discrete random variable is nothing else than conditioning on an event having nonzero probability. The analysis for the continuous case involves some technical subtleties, because the probability that a continuous random variable will take on a particular value is always zero.

13.1.1 Conditional probability mass function

Let X and Y be two discrete random variables with joint probability mass function $p(x, y) = P(X = x,\ Y = y)$. The *conditional probability mass function* of X given that $Y = y$ is denoted and defined by

$$P(X = x \mid Y = y) = \frac{P(X = x,\ Y = y)}{P(Y = y)}$$

for any fixed y with $P(Y = y) > 0$. 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 $B = \{\omega :\ Y(\omega) = y\}$ with ω denoting an element of the sample space. Note that, for any *fixed* y, the function $P(X = x \mid Y = y)$ satisfies

$$P(X = x \mid Y = y) \geq 0 \quad \text{for all } x \quad \text{and} \quad \sum_x P(X = x \mid Y = y) = 1,$$

showing that $P(X = x \mid Y = y)$ as function of x is indeed a probability mass function. The notation $p_X(x \mid y)$ is often used for $P(X = x \mid Y = y)$.

Using the representation

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

and the fact that $\sum_y P(X = x,\ Y = y) = P(X = x)$, the unconditional probability $P(X = x)$ can be calculated from

$$P(X = x) = \sum_y P(X = x \mid Y = y)P(Y = y).$$

This is the law of conditional probability stated in terms of discrete random variables.

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 $p_Y(y \mid 3) = P(Y = y \mid X = 3)$:

$$p_Y(6 \mid 3) = \frac{1/36}{7/36} = \frac{1}{7},\ p_Y(7 \mid 3) = p_Y(8 \mid 3) = p_Y(9 \mid 3) = \frac{2/36}{7/36} = \frac{2}{7}$$
$$p_Y(y \mid 3) = 0 \qquad \text{for } y = 2, 3, 4, 5, 10, 11, 12.$$

Problem 13.1 You repeatedly draw a number at random from the numbers $1, 2, \ldots, 10$ with replacement. Let X be the number of draws until the number 1 appears for the first time and Y the number of draws until the number 10 appears for the first time. What is the conditional mass function $P(X = x \mid Y = y)$?

Problem 13.2 Three different numbers are chosen at random from the numbers $1, 2, \ldots, 10$. Let X be the smallest of these three numbers and Y the largest. What are the conditional mass functions $P(X = x \mid Y = y)$ and $P(Y = y \mid X = x)$?

Problem 13.3 A fair die is rolled until a six appears. Each time the die is rolled a fair coin is tossed. Let the random variables X and Y denote the number of rolls of the die and the number of heads from the tosses of the coin. What is the conditional mass function $P(X = x \mid Y = y)$. *Hint*: use the identity $\sum_{n=r}^{\infty} \binom{n}{r} a^n = a^r/(1-a)^{r+1}$ for $0 < a < 1$ to find $P(Y = j)$ when $j \neq 0$.

Problem 13.4 Two dice are rolled. Let the random variable X be the smallest of the two outcomes and let Y be the largest of the two outcomes. What are the conditional mass functions $P(X = x \mid Y = y)$ and $P(Y = y \mid X = x)$?

Problem 13.5 You simultaneously roll 24 dice. Next you roll only those dice that showed the face value six in the first roll. Let the random variable X denote the number of sixes in the first roll and Y the number of sixes in the second roll. What is the conditional mass function $P(X = x \mid Y = y)$?

Problem 13.6 Let X denote the number of hearts and Y the number of diamonds in a bridge hand. What are the conditional mass functions $P(X = x \mid Y = y)$ and $P(Y = y \mid X = x)$?

13.1.2 Conditional probability density function

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 $\Delta b > 0$,

$$\begin{aligned} &P\left(a - \frac{1}{2}\Delta a \le X \le a + \frac{1}{2}\Delta a \,\middle|\, b - \frac{1}{2}\Delta b \le Y \le b + \frac{1}{2}\Delta b\right) \\ &= \frac{P(a - \frac{1}{2}\Delta a \le X \le a + \frac{1}{2}\Delta a, b - \frac{1}{2}\Delta b \le Y \le b + \frac{1}{2}\Delta b)}{P(b - \frac{1}{2}\Delta b \le Y \le b + \frac{1}{2}\Delta b)} \\ &\approx \frac{f(a, b)\Delta a \Delta b}{f_Y(b)\Delta b} = \frac{f(a, b)}{f_Y(b)}\Delta a \end{aligned}$$

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 $Y = y$ is defined by*

$$f_X(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \qquad -\infty < x < \infty$$

for any fixed y with $f_Y(y) > 0$.

Note that, for any *fixed* y, the function $f_X(x \mid y)$ satisfies

$$f_X(x \mid y) \ge 0 \quad \text{for all } x \quad \text{and} \quad \int_{-\infty}^{\infty} f_X(x \mid y)\, dx = 1,$$

showing that $f_X(x \mid y)$ as function of x is indeed a probability density function. Similarly, the conditional probability density function of the random variable Y given that $X = x$ is defined by $f_Y(y \mid x) = \frac{f(x,y)}{f_X(x)}$ for any fixed x with $f_X(x) > 0$.

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 the 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 concept of conditional probability distribution function is defined as follows. For any fixed y with $f_Y(y) > 0$, the conditional probability that the random variable X takes on a value smaller than or equal to x given that $Y = y$ is denoted by $P(X \leq x \mid Y = y)$ and is defined by

$$P(X \leq x \mid Y = y) = \int_{-\infty}^{x} f_X(u \mid y)\, du.$$

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 11.6, 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 y with $-1 < y < 1$,

$$f_X(x \mid y) = \begin{cases} \frac{1}{2\sqrt{1-y^2}} & \text{for } -\sqrt{1-y^2} < x < \sqrt{1-y^2} \\ 0 & \text{otherwise.} \end{cases}$$

In other words, the conditional distribution of X given that $Y = y$ is the uniform distribution on the interval $(-\sqrt{1-y^2}, \sqrt{1-y^2})$. The same distribution as that of the x-coordinate of a randomly chosen point of the horizontal chord through the point $(0, y)$. This chord has length $2\sqrt{1-y^2}$, 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 $f_X(x \mid y)$ of X given that $Y = y$ is the

$$N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)\right)$$

density and the conditional probability density $f_Y(y \mid x)$ of Y given that $X = x$ is the

$$N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2)\right)$$

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 11.6 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 to be discussed in Section 15.5.

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, 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

$$P\left(X \le \frac{1}{2},\ Y \le \frac{1}{2}\right) = \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}} f(x, y)\, dx\, dy = \int_0^{\frac{1}{2}} \frac{dx}{1-x} \int_0^{\frac{1}{2}} dy$$
$$= \frac{1}{2} \int_{\frac{1}{2}}^{1} \frac{du}{u} = \frac{1}{2} \ln(2) = 0.3466.$$

Example 13.5 For an insurance company, let the random variable X be the losses under collision insurance and Y the losses under liability insurance. The two random variables X and Y have the joint probability density function $f(x, y) = \frac{1}{24}(1.5x^2 + 4y)$ for $0 < x, y < 2$ and $f(x, y) = 0$ otherwise, where the losses are expressed in units of one million dollars. What is the conditional probability that the loss under collision insurance is less than 1 given that the total loss is 1.5?

Solution. Define the random variables V and W by $V = X$ and $W = X + Y$. To answer the question, we need the density function $f_W(w)$ of W and the joint probability density function $f_{V,W}(v, w)$ of V and W. The conditional density of V given that $W = w$ is $f_V(v \mid w) = f_{V,W}(v, w)/f_W(w)$ and so the desired probability is equal to

$$\int_0^1 \frac{f_{V,W}(v, 1.5)}{f_W(1.5)} dv.$$

The density function of $W = X + Y$ is only needed for $0 < w < 2$ and so it suffices to determine $P(X + Y \leq w)$ for $0 \leq w \leq 2$. Then we have

$$P(X + Y \leq w) = \int_0^w dx \int_0^{w-x} \frac{1}{24}(1.5x^2 + 4y)\, dy = \frac{w^4}{192} + \frac{w^3}{36}.$$

This gives

$$f_W(w) = \frac{w^3}{48} + \frac{w^2}{12} \quad \text{for } 0 < w < 2.$$

It suffices to determine $f_{V,W}(v, w)$ for $0 < v < w < 2$. Then we have

$$\begin{aligned} P(V \leq v,\ W \leq w) &= \int_0^v dx \int_0^{w-x} \frac{1}{24}(1.5x^2 + 4y)\, dy \\ &= \frac{1}{24}\left[\frac{1}{2}wv^3 - \frac{3}{8}v^4 + 2\left(w^2v - wv^2 + \frac{1}{3}v^3\right)\right]. \end{aligned}$$

This gives

$$f_{V,W}(v, w) = \frac{1}{24}(1.5v^2 + 4(w - v)) \quad \text{for } 0 < v < w < 2.$$

The probability that the loss under collision insurance is less than 1 given that the total loss is 1.5 can now be computed as

$$\int_0^1 \frac{f_{V,W}(v, 1.5)}{f_W(1.5)} dv = \int_0^1 \frac{(1.5v^2 + 4(1.5 - v))/24}{1.5^3/48 + 1.5^2/12} = 0.7273.$$

Problem 13.7 Let X and Y be two continuous random variables with the joint probability density $f(x, y) = xe^{-x(y+1)}$ for $x, y > 0$ and $f(x, y) = 0$ otherwise. What are the conditional probability densities $f_X(x \mid y)$ and $f_Y(y \mid x)$? What is the probability that Y is larger than 1 given that $X = 1$?

Problem 13.8 Let X and Y be two continuous random variables with the joint probability density $f(x, y) = \frac{2}{7}(2x + 5y)$ for $0 < x, y < 1$ and $f(x, y) = 0$ otherwise. What are the conditional probability densities $f_X(x \mid y)$ and $f_Y(y \mid x)$? What is the probability that X is larger than 0.5 given that $Y = 0.2$?

Problem 13.9 Let the discrete random variable X and the continuous random variable Y be defined on the same sample space, where the probability mass function $p_X(x) = P(X = x)$ and the conditional probability density function $f_Y(y \mid x) = \frac{d}{dy}P(Y \leq y \mid X = x)$ are given. It is assumed that the function $f_Y(y \mid x)$ is continuous in y for any x. How would you compute the conditional probability mass function of X given the observed value $Y = y$? *Hint*: Evaluate $P(X = x \mid y - \frac{1}{2}\Delta y \leq Y \leq y + \frac{1}{2}\Delta y)$ for Δy small.

Problem 13.10 A receiver gets as input a random signal that is represented by a discrete random variable X, where X takes on the value $+1$ with probability p and the value -1 with probability $1 - p$. The output Y is a continuous random variable which is equal to the input X plus random noise, where the random noise has an $N(0, \sigma^2)$ distribution. You can only observe the output. What is the conditional probability of $X = 1$ given the observed value $Y = y$?

Problem 13.11 A point X is randomly chosen in the interval $(0, 1)$ and next a point Y is randomly chosen in $(1 - X, 1)$. What are the probabilities $P(X + Y > 1.5)$ and $P(Y > 0.5)$?

Problem 13.12 Let X and Y be two independent random variables each having an exponential density with expected value $1/\lambda$. Show that the conditional density of X given that $X + Y = u$ is the uniform density on $(0, u)$. *Hint*: use Rule 11.4 with $V = X$ and $W = X + Y$.

13.2 Law of conditional probability for random variables

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 probability* is a special case of Rule 8.2 in Chapter 8. In the situation of continuous random variables X and Y, the continuous analog of the law of conditional probability 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 probability 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 probability 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 13.6 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 following example deals with the continuous version of the game of chance discussed in Example 8.7.

Example 13.7 Each of two players A and B in turn draws a random number once or twice, where the random number comes from the uniform distribution on (0, 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. It is obvious how player B acts. Player B stops after the first draw if the score of this draw is larger than the final score of player A and continues otherwise. In analyzing the 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

$$S(a) = \text{the winning probability of player } A \text{ if player } A \text{ stops after the first draw given that } U = a.$$

Also, define $C(a)$ as the winning probability of player A if player A continues for a second draw given that $U = a$. It will be shown that there is a number a^* such that $C(a) > S(a)$ for all $0 \le a < a^*$, $C(a^*) = S(a^*)$ and $C(a) < S(a)$ for all $a^* < a \le 1$. 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 probability, the overall winning probability of player A under this strategy is given by

$$P_A = \int_0^{a^*} C(a) f_U(a)\, da + \int_{a^*}^1 S(a) f_U(a)\, da,$$

where $f_U(a) = 1$ for all $0 \le a \le 1$. It remains to find $S(a)$, $C(a)$ 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 probability, we then find

$$\begin{aligned} 1 - S(a) &= \int_a^1 1 \times f_Y(y)\, dy + \int_0^a [1 - y - (a - y)] f_Y(y)\, dy \\ &= 1 - a + (1 - a)a = 1 - a^2, \end{aligned}$$

showing that $S(a) = a^2$. To obtain $C(a)$, 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 $S(x) = x^2$ (only the final score of player A matters for player B). Thus,

$$\begin{aligned} C(a) &= \int_0^{1-a} (a + v)^2 f_V(v)\, dv + \int_{1-a}^1 0 \times f_V(v)\, dv \\ &= \int_a^1 w^2\, dw = \frac{1}{3}(1 - a^3). \end{aligned}$$

The function $S(a) - C(a)$ is negative for $a = 0$ and positive for $a = 1$. Also, this continuous function is increasing on $(0, 1)$ (the derivative is positive). This proves that the optimal stopping level for the first player is the unique solution

to the equation

$$a^2 = \frac{1}{3}(1 - a^3)$$

on (0, 1). The solution of this equation can be numerically calculated as

$$a^* = 0.53209.$$

The winning probability of player A can next be calculated as $P_A = 0.4538$.

A challenging problem is to analyze the three-player game. The reader is asked to do this in Problem 13.21 and to verify that the optimal stopping level for the first player is the solution $a^* = 0.64865$ to the equation $a^4 = \frac{1}{5}(1 - a^5)$. The optimal strategy for the second player in the three-player game is to stop after the first draw only if the score of this draw exceeds both the final score of the first player and the stopping level 0.53209 from the two-player game. The solution of the continuous game provides an excellent approximation to the optimal stopping levels for the players in a television game show that involves spinning a wheel with the money values $1, 2, \ldots, R$ on it. In the game show each of three players in turn spins the wheel once or twice and the goal is to have the highest total score without going over R.

Conditional probability and rejection sampling

The properties of conditional probability can be exploited to simulate from "difficult" probability densities for which the inverse-transformation method cannot be used.

Rule 13.2 *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 a probability density $g(y)$. Let $f(y)$ be another probability density function such that, for some constant c, $f(y) \leq cg(y)$ for all y. Then*

$$P\left(Y \leq x \mid U \leq \frac{f(Y)}{cg(Y)}\right) = \int_{-\infty}^{x} f(y)dy \qquad \text{for all } x.$$

The technical proof is as follows. By conditioning on Y and using the fact that $P(U \leq u) = u$ for $0 < u < 1$, it follows that

$$P\left(U \leq \frac{f(Y)}{cg(Y)}\right) = \int_{-\infty}^{+\infty} \frac{f(y)}{cg(y)} g(y)dy = \frac{1}{c}.$$

Also, by conditioning on Y,

$$P\left(Y \le x,\ U \le \frac{f(Y)}{cg(Y)}\right) = \int_{-\infty}^{x} \frac{f(y)}{cg(y)} g(y)dy = \frac{1}{c}\int_{-\infty}^{x} f(y)dy.$$

Using the fact that

$$P\left(Y \le x \mid U \le \frac{f(Y)}{cg(Y)}\right) = P\left(Y \le x,\ U \le \frac{f(Y)}{cg(Y)}\right) \Big/ P\left(U \le \frac{f(Y)}{cg(Y)}\right),$$

the desired result follows.

Rule 13.2 underlies the so-called acceptance-rejection method for simulating from a "difficult" density $f(x)$ via an "easy" density $g(y)$. This method goes as follows. *Step* 1. Generate Y having density $g(y)$ and generate a random number U. *Step* 2. If $U \le \frac{f(Y)}{cg(Y)}$, 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 $U \times cg(Y) \le f(Y)$. In the proof of Rule 13.2 we found that $P\left(U \le \frac{f(Y)}{cg(Y)}\right) = \frac{1}{c}$, which shows that the number of iterations of the acceptance-rejection method is geometrically distributed with an expected value of c. The constant c satisfies $c \ge 1$, as follows by integrating both sides of $f(x) \le cg(x)$ over x and noting that the densities $f(x)$ and $g(x)$ integrate to 1. We come back to this method in Section 15.5.

Problem 13.13 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 13.14 Every morning there is a limousine service from Gotham City to the airport. The van has capacity for 10 passengers. The number of reservations for a trip follows a discrete uniform distribution between 5 and 10. The probability that a person with a reservation shows up for the trip is 0.80. Passengers without reservations are also accepted as long as the van is not full. The number of walk-up passengers for a trip is independent of the number of reservations and has a discrete probability distribution with $p_0 = 0.25,\ p_1 = 0.45,\ p_2 = 0.20$, and $p_3 = 0.10$. For a given trip, let the random variable V denote the number of passengers with reservations who show up and let W denote the number of walk-up passengers who are accepted. What are the probability mass functions of V and W?

Problem 13.15 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 probability to obtain the probability distribution of the number of red balls among the selected balls. Does the result surprise you? Can you give a direct probabilistic argument for the result obtained?

Problem 13.16 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 13.17 Use twice the law of conditional probability 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 13.18 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 the inter-arrival times of buses number 3 are independent random variables having an exponential distribution 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 probability 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 13.19 Let the random variable Z have the standard normal density. Verify that the half-normal distributed random variable $|Z|$ has the probability density $f(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ for $x > 0$ and propose a "simple" density $g(x)$ to simulate from the density $f(x)$. What is the constant c for your choice of $g(x)$? What density $g(x)$ would you propose to simulate from the density $f(x) = \frac{2}{\pi}\sqrt{1 - x^2}$ for $-1 < x < 1$?

Problem 13.20 Consider the following two stochastic two-person games in which the players have no information about each other's actions.
(**a**) Assume now in the stochastic game of Example 13.7 that each player must act without knowing what the other player has done. What strategy should player A follow to ensure that his probability of winning is at least 50%,

regardless of what player B will do? Assume that the player whose total score is closest to 1 is the winner if the scores of both players exceed 1. *Hint*: define $P(a, b)$ as the win probability of player A when players A and B use threshold values a and b for stopping after the first draw.
(**b**) Two players A and B each receive a random number between 0 and 1. Each player gets a once-only opportunity to discard his number and receive a new random number between 0 and 1. This choice is made without knowing the other player's number or whether the other player chose to replace his number. The player ending up with the higher number wins. What strategy should player A follow to ensure that his probability of winning is at least 50%, regardless of what player B will do?

Problem 13.21 (difficult) Consider the three-player variant of Example 13.7. What is the optimal strategy for the first player A and what is the overall win probability of player A? Next use the probability distribution function of the final score of player A to find the overall win probabilities of the players B and C.

13.3 Law of conditional expectation

In the case that the random variables X and Y have a discrete distribution, the *conditional expectation* of X given that $Y = y$ is defined by

$$E(X \mid Y = b) = \sum_x x \; P(X = x \mid Y = y)$$

for each y with $P(Y = y) > 0$ (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 $Y = y$ is defined by

$$E(X \mid Y = y) = \int_{-\infty}^{\infty} x \; f_X(x \mid y)\, dx.$$

for each y with $f_Y(y) > 0$ (assuming that the integral is well-defined).

Just as the law of conditional probability directly follows from the definition of the conditional distribution of X given that $Y = y$, the law of conditional expectation is a direct consequence of the definition of $E(X \mid Y = y)$. In the discrete case the *law of conditional expectation* 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 expectation 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 13.8 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

$$E(X \mid Y = y) = \int_{10}^{y} x \frac{2(x-10)}{(y-10)^2}\, dx = 10 + \frac{2}{3}(y-10) \qquad \text{for } 10 < y < 30.$$

Further, using the general formula $P(X \in A \mid Y = y) = \int_A f_X(x \mid y)\, dx$,

$$P(X > 20 \mid Y = 25) = \int_{20}^{25} \frac{2(x-10)}{(25-10)^2}\, dx = \frac{5}{9}.$$

Problem 13.22 You choose three different numbers at random from the numbers $1, 2, \ldots, 100$. Let X be the smallest of these three numbers and Y the largest. What are the conditional expected values $E(X \mid Y = j)$ and $E(Y \mid X = i)$?

Problem 13.23 You generate three random numbers from $(0, 1)$. Let X be the smallest of these three numbers and Y the largest. What are the conditional expected values $E(X \mid Y = y)$ and $E(Y \mid X = x)$?

Problem 13.24 Verify that $E(X \mid Y = y) = E(X)$ for all y if X and Y are independent.

Problem 13.25 Let (X, Y) have a bivariate normal density with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, where $\sigma_1^2 = \sigma_2^2$. What is $E(X \mid X + Y = v)$? *Hint*: use the result of Problem 12.7.

Problem 13.26 Let the continuous random variables X and X have a joint density function $f(x, y)$. What is the natural definition for $E(Y \mid a < X < b)$? Next show that $E(Y \mid a < X < b)$ equals $\frac{\rho}{\sqrt{2\pi}}(e^{-\frac{1}{2}a^2} - e^{-\frac{1}{2}b^2})/[\Phi(b) - \Phi(a)]$ if (X, Y) has a standard bivariate normal density with correlation coefficient ρ.

Problem 13.27 Let the joint density function $f(x, y)$ of the random variables X and Y be equal to $4e^{-2x}$ for $0 < y \leq x < \infty$ and 0 otherwise. What are the conditional expected values $E(X \mid Y = y)$ and $E(Y \mid X = x)$?

Problem 13.28 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%?

13.3.1 The regression curve and conditional expectation

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 the case that we can observe the random variable X but we want to know the dependent random

[†] 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 $(x, y) = (0,0)$.

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 expectation, $E[(Y - g(X))^2]$ 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 with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, it follows from the result in Example 13.3 that

$$m(x) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

In this case the optimal prediction function $m(x)$ coincides with the best linear prediction function discussed in Section 11.5. The best linear prediction function uses only the expected values, the variances, and the correlation coefficient of the random variables X and Y.

Example 13.9 In a classical study on the heights of fathers and their grown sons, Sir Francis Galton (1822–1911) measured the heights of 1078 fathers and sons, and found that the sons averaged one inch taller than the fathers. Let the random variable X represent the fathers' heights in inches and Y represent the sons' heights in inches. Galton modeled the distribution of (X, Y) by a bivariate normal distribution with parameters

$$E(X) = 67.7, E(Y) = 68.7, \sigma(X) = 2.7, \sigma(Y) = 2.7, \rho(X, Y) = 0.5.$$

Using the above results, the prediction for Y given that $X = x$ is given by

$$E(Y \mid X = x) = 68.7 + \frac{0.5 \times 2.7}{2.7}(x - 67.7) = 34.85 + 0.5x$$

for each x. The line $y = 34.85 + 0.5x$ is midway between the line $y = x + 1$ and the line $y = 68.7$. For example, if the father is very tall and has a height of 73.1 inches, the son is not predicted to be 74.1 inches but only 71.4 inches tall. Galton called this phenomenon regression to the mean.

13.4 Conditional expectation as a computational tool

In applied probability problems, the law of conditional expectation 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 13.10 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 $E(\text{insurance payout}) = \$86{,}466.50$.

Example 13.11 A reliability system has two identical units, where one unit is in full operation and the other unit is in cold standby. The lifetime of an operating unit has an exponential density with expected value $1/\mu$. Upon failure of the operating unit, the standby unit is put into operation provided a standby unit is available. The replacement time of a failed unit is fixed and is equal to $\tau > 0$. A system failure occurs if no standby unit is available at the moment the operating unit fails. What is the expected value of the time until the first system failure?

Solution. Let the random variable X denote the time until the first system failure. By conditioning on the lifetime of the standby unit which is put into operation upon failure of the operating unit and using the memoryless property of the exponential distribution , we find

$$E(X) = \frac{1}{\mu} + \int_0^\tau x\mu e^{-\mu x} dx + e^{-\mu\tau}(\tau + E(X)).$$

Using the fact that $\int_0^\tau x\mu e^{-\mu x} dx = (1/\mu)(1 - e^{-\mu\tau} - \mu\tau e^{-\mu\tau})$, we next find that

$$E(X) = \frac{2 - e^{-\mu\tau}}{\mu(1 - e^{-\mu\tau})}.$$

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 are central to Chapter 15 on Markov chains.

Example 13.12 In any drawing of the Lotto 6/45 six different numbers are chosen at random from the numbers $1, \ldots, 45$. 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}, we use a recurrence relation for the μ_i. By conditioning on the result of the

[†] 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.

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 $\mu_{45} = 31.497$.

Conditional probability and optimization

The law of conditional expectation is a key ingredient of the recursive approach of stochastic dynamic programming. This approach is a computational tool for optimization problems in which a sequence of interrelated decisions must be made in order to maximize reward or minimize cost. As an example, consider the game of rolling a fair die at most five times. You may stop whenever you want and receive as a reward the number shown on the die at the time you stop. What is the stopping rule that maximizes your expected payoff in this optimal stopping game? To answer this question, the idea is to consider a sequence of nested problems having planning horizons of increasing length. For the one-roll problem in which only one roll is permitted, the solution is trivial. You stop after the first roll and your expected payoff is $1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5$. In the two-roll problem, you stop after the first roll if the outcome of this roll is larger than the expected value 3.5 of the amount you get if you do not stop but continue with what is a one-roll game. Hence, in the two-roll problem, you stop if the first roll gives a 4, 5, or 6; otherwise, you continue. The expected payoff in the two-roll game is $\frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 + \frac{3}{6} \times 3.5 = 4.25$. Next consider the three-roll problem. If the first roll in the three-roll problem gives an outcome larger than 4.25, then you stop; otherwise, you do not stop and continue with what is a two-roll game. Hence the expected payoff in the three-roll problem is $\frac{1}{6} \times 5 + \frac{1}{6} \times 6 + \frac{4}{6} \times 4.25 = 4.67$. Knowing this expected payoff, we can solve the four-roll problem. In the four-roll problem you stop after the first roll if this roll gives a 5 or 6; otherwise, you continue. The expected payoff in the four-roll problem is $\frac{1}{6} \times 5 + \frac{1}{6} \times 6 + \frac{4}{6} \times 4.6667 = 4.944$. Finally, we find the optimal strategy for the original five-roll problem. In this problem you stop after the roll if this roll gives a 5 or 6; otherwise, you continue. The maximal expected payoff in the original problem is $\frac{1}{6} \times 5 + \frac{1}{6} \times 6 + \frac{4}{6} \times 4.944 = 5.129$.

The above method of backward induction is called the method of dynamic programming. It decomposes the original problem in a series of nested problems having planning horizons of increasing length. Each nested problem is simple to solve and the solutions of the nested problems are linked by a recursion. The

above argument can be formalized as follows. For $k = 1, 2, \ldots, 5$, define

$$f_k(i) = \text{the maximal expected payoff if still } k \text{ rolls are permitted and the outcome of the last roll is } i,$$

where $i = 0, 1, \ldots, 6$. This function is called the value-function. It enables us to compute the desired maximal expected payoff $f_5(0)$ and the optimal strategy for achieving this expected payoff in the five-roll problem. This is done by applying the recursion

$$f_k(i) = \max\left\{i,\ \frac{1}{6}\sum_{j=1}^{6} f_{k-1}(j)\right\} \quad \text{for } i = 0, 1, \ldots, 6,$$

where k runs from 1 to 5 and the recursion is initialized with $f_0(j) = j$.

The method of backward induction is very versatile, and does not require that the outcomes of the successive experiments are independent of each other. As an example, take the following game. You take cards, one at a time, from a thoroughly shuffled deck of 26 red and 26 black cards. You may stop whenever you want and your payoff is the number of red cards drawn minus the number of black cards drawn. What is the maximal expected value of the payoff? The approach is again to decompose the original problem into a sequence of smaller nested problems. Define the value function $E(r, b)$ as the maximal expected payoff you can still achieve if r red cards and b black cards are left in the deck. Using conditional expectations, we can establish the recursive scheme

$$E(r, b) = \max\left\{b - r,\ \frac{r}{r+b}E(r-1, b) + \frac{b}{r+b}E(r, b-1)\right\}.$$

The desired maximal expected $E(26, 26)$ is obtained by "backward" calculations starting with $E(r, 0) = 0$ and $E(0, b) = b$. The maximal expected payoff is $E(26,\ 26) = 2.6245$. The optimal decisions in the various states can be summarized through threshold values β_k: stop if the number of red cards drawn minus the number of black cards drawn is β_k or more after the kth draw; otherwise, continue. The numerical values of the β_k are $\beta_1 = 2$, $\beta_2 = 3$, $\beta_3 = 4$, $\beta_4 = 5$, $\beta_5 = 6$, $\beta_6 = 5$, $\beta_7 = 6$, $\beta_8 = 7$, $\beta_9 = 6$, $\beta_{2m} = 5$ and $\beta_{2m+1} = 4$ for $5 \le m \le 11$, $\beta_{2m} = 3$ and $\beta_{2m+1} = 4$ for $12 \le m \le 16$, $\beta_{2m} = 3$ and $\beta_{2m+1} = 2$ for $17 \le m \le 21$, $\beta_{44} = 1$, $\beta_{45} = 2$, $\beta_{46} = 1$, $\beta_{47} = 2$, $\beta_{48} = 1$, $\beta_{49} = 0$, $\beta_{50} = 1$, $\beta_{51} = 0$. In the problems below we give several other problems that can be solved by the method of backward induction.

Problem 13.29 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 13.30 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 13.31 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 expectation to find these expected values.

Problem 13.32 Let's return to Problem 13.18. Use the law of conditional expectation to verify that the expected value of your waiting time until the next bus arrival is equal to $\frac{15}{e}$.

Problem 13.33 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 expectation 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 13.34 Five fair coins are tossed together. This is repeated until four or more heads come up. Let X be the number of times that two or more heads come up and Y the number of times that two or more tails come up. What are $E(X)$ and $E(Y)$?

Problem 13.35 You roll a fair die until two consecutive sixes have been obtained. What is the expected value of the number of rolls?

Problem 13.36 Consider Problem 2.44 again. Calculate the probability mass function of the number of points gained in a single turn under the hold-at-20 rule. *Hint*: derive a recursion for the probability that the turn total will take on the intermediate value i in a given turn, where $2 \leq i \leq 19$. Next, use a recursion to find the expected value of the number of turns needed to reach a total score of 100 or more points under the hold-at-20 rule. You are also asked to find the expected value of the number of turns needed to reach a total score of 100 or more points under the five-dice rule in the game of fast Pig.[†]

Problem 13.37 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 you 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 13.38 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 number of drawings until you have a random number below a?

Problem 13.39 Consider Problem 13.33 again but assume now that you want to find the expected payoff of an optimal stopping rule. What is the maximal expected payoff for $n = 25$, 50, 100 and 1,000?

Problem 13.40 Use the optimization method of backward induction for the following two dice games.
(a) Consider a game of rolling two fair dice at most six times. You may stop whenever you want and your payoff is the score of the two dice upon stopping. What is the maximal expected payoff?
(b) The dice game of threes starts by rolling five dice. After each roll you must freeze at least one of the dice rolled. A frozen die counts as zero if the face

[†] The game of Pig is extensively studied in T. Neller and C.G.M. Presser, Optimal play of the dice game Pig, *The UMAP Journal* **25** (2004): 25–47. For the case of two players who in turn roll the die, it is computed in the paper how the first player should act to maximize the probability of reaching 100 or more points. The computations are based on the optimization method of dynamic progranmming.

value is three; otherwise, it counts at face value. You keep rolling the unfrozen dice until all five dice are frozen. You want to minimize the expected value of the total score of the frozen dice. What is the minimal expected value?

13.5 Bayesian statistics – continuous case

Bayesian statistics takes a fundamentally different viewpoint from classical statistics. Whereas the frequentist view of probability and of statistical inference is based on the idea of a random experiment that can be repeated very often, the Bayesian view of probability and of statistical inference is based on personal assessments of probability and on data of a single performance of a random experiment. In practice, many statisticians use ideas from both methods. Unlike the frequentist approach, the Bayesian approach to statistical inference treats population parameters not as fixed, unknown constants but as random variables – subject to change as additional data arise. Probability distributions have to be assigned to these parameters by the investigator before observing data. These distributions are called prior distributions and are inherently subjective. They represent the investigator's uncertainty about the true value of the unknown parameters. Assigning probabilities by degree of belief is consistent with the idea of a fixed but unknown value of the parameter. Thus, in the Bayesian approach, you state a hypothesis by means of a prior probability distribution and you revise your beliefs about the population parameters by learning from data that were collected. The posterior probability distribution reflects your new beliefs about the population parameters. Bayes' rule is the recipe for computing the posterior distribution based on the prior distribution and the observed data. In Section 8.4 we considered two examples of a random experiment in which the hypotheses were represented by a parameter Θ having a discrete prior distribution $\{p_0(\theta_i),\ i = 1, \ldots, n\}$. The goal was to update this prior distribution after having obtained data x from the random experiment. Letting the likelihood $L(x \mid \theta)$ denote the conditional probability of obtaining the data x given the value θ of the parameter Θ, the prior distribution was updated according to Bayes' rule:

$$p(\theta_i \mid x) = \frac{L(x \mid \theta_i)p_0(\theta_i)}{\sum_{k=1}^{n} L(x \mid \theta_k)p_0(\theta_k)} \quad \text{for } i = 1, \ldots, n.$$

This is how it all works for discrete problems. To illustrate this, consider the following experiment. Suppose that there is reason to believe that a coin might be slightly biased towards heads. To test this, you decide to throw the coin 1,000 times. Before performing the experiment, you express your uncertainty

about the unbiasedness of the coin by assuming that the probability of getting heads in a single toss of the coin can take on the values $\theta = 0.50$, 0.51, and 0.52 with respective prior probabilities $p_0(\theta) = \frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{6}$. Next the experiment is performed and 541 heads are obtained in 1,000 tosses of the coin. The likelihood of getting 541 heads in 1,000 tosses is $L(541 \mid \theta) = \binom{1,000}{541}\theta^{541}(1-\theta)^{459}$ for $\theta = 0.50$, 0.51, and 0.52. This leads to the posterior probability $p(\theta \mid 541) = 0.1282$ for $\theta = 0.50$. In other words, your posterior belief that the coin is fair equals 0.1282. In classical statistics, one would compute the probability of getting 541 or more heads in 1,000 tosses of the coin under the hypothesis that the coin is fair. This tail probability is equal to 0.0052 and is called the p-value. Many classical statisticians would consider this small p-value as significant evidence that the coin is biased towards heads. However, your subjective Bayesian probability of 0.1282 for the hypothesis of a fair coin is not strong enough evidence for such a conclusion. In general, the p-value is typically much smaller than the Bayesian probability of the hypothesis. This can be explained as follows. The p-value is based on the set of all possible observations that cast as much or more doubt on the hypothesis than the actual observations do. It is not possible to base the p-value only on the actual data, because it frequently happens that all individual outcomes have such small probabilities that every outcome would look significant. The inclusion of unobserved data means that the resulting p-value may greatly exaggerate the strength of evidence against the hypothesis. In the Bayesian approach only the actual data obtained matter. However, in order to give a Bayesian conclusion about the truth of the hypothesis by using only the actual data, you must first choose a prior distribution for the truth of the hypothesis. In many situations you can choose priors on the basis of earlier experimental results or results predicted by scientific theory.

Bayes' rule for the continuous case

What is the formulation of Bayes' rule for the continuous case? For that purpose we go back to results in Section 13.1. Let X and Y be two continuous random variables with joint density $f(x, y)$ and marginal densities $f_X(x)$ and $f_Y(y)$. Denote by $f_X(x \mid y)$ the conditional density of X given $Y = y$ and by $f_Y(y \mid x)$ the conditional density of Y given $X = x$. Using the relations $f(x, y) = f_Y(y \mid x) f_X(x)$ and $f(x, y) = f_X(x \mid y) f_Y(y)$ together with $f_X(x) = \int f_X(x \mid y) f_Y(y)\, dy$, it follows for any fixed x that

$$f_Y(y \mid x) = \frac{f_X(x \mid y) f_Y(y)}{\int f_X(x \mid y) f_Y(y)\, dy} \quad \text{for all } y.$$

In continuous Bayesian problems, the hypotheses are represented by a continuous parameter Θ having a prior density $f_0(\theta)$. Let the likelihood function $L(x \mid \theta)$ represent the conditional density of the data x given θ. Then, given data x, the prior density of Θ is updated to the posterior density

$$f(\theta \mid x) = \frac{L(x \mid \theta) f_0(\theta)}{\int_\Theta L(x \mid \theta) f_0(\theta)\, d\theta} \quad \text{for all } \theta.$$

The denominator of this formula is not a function of θ. In the literature the Bayes formula is often written as

$$f(\theta \mid x) \propto L(x \mid \theta) f_0(\theta) \quad \text{for all } \theta,$$

where the symbol $\propto$ stands for proportionality and the proportionality constant is such that the conditional density $f(\theta \mid x)$ integrates to 1 for fixed x. In the case that the elements of Θ are high-dimensional vectors it is usually not computationally feasible to find the proportionality constant by numerical integration. Also, in the high-dimensional case, it will typically be very difficult to obtain the marginal density of any component of Θ by numerical integration. One then has to resort to Markov chain Monte Carlo simulation to find the posterior density. This method has enormously increased the applicability of Bayesian analysis and it will be discussed in Chapter 15.

Bayes' rule for the mixed case

In mixed Bayesian problems the update of the prior density $f_0(\theta)$ is based on the discrete outcome E of a random experiment rather than on an outcome of a continuous variable. The adaptation of Bayes' formula is obvious. Letting the likelihood $L(E \mid \theta)$ denote the conditional probability of the outcome E given θ, the prior density is updated to the posterior density $f(\theta \mid E)$ defined by

$$f(\theta \mid E) \propto L(E \mid \theta) f_0(\theta) \quad \text{for all } \theta,$$

where the proportionality constant is such that $f(\theta \mid E)$ integrates to 1. The motivation of this definition proceeds along similar lines to the motivation of Definition 13.1. Suppose that the prior density of the success probability in a Bernoulli trial is given by a beta(α, β) density and that n independent repetitions of the Bernoulli trials have led to s successes and $r = n - s$ failures. Then the above relation gives that the posterior density is proportional to $\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^s(1-\theta)^r = \theta^{\alpha+s-1}(1-\theta)^{\beta+r-1}$ and thus has a beta$(\alpha + s, \beta + r)$ density. If the posterior density follows the same form as the prior density, the prior and the likelihood are said to be conjugate. The property of conjugateness is helpful in Bayesian computations.

In the following example we give an application of a mixed Bayesian problem. This example is a continuous version of Problem 8.37 and illustrates that Bayesian analysis is particularly useful for less repeatable experiments.

Example 13.13 You wonder who is the better player of the tennis players Alassi and Bicker. You have a continuous prior density $f_0(\theta)$ for the probability θ that Alassi is the better player. Then you learn about a tournament at which a best-of-five series of matches is played between Alassi and Bicker over a number of days. In such an encounter the first player to win three matches is the overall winner. It turns out that Alassi wins the best-of-five contest. How should you update your prior density $f_0(\theta)$?

Solution. The prior density $f_0(\theta)$ expresses your uncertainty about the probability θ of Alassi being the better player. A typical approach for overcoming concerns about the subjective nature of prior densities is to consider several choices for the prior density. We will consider the following two choices for the prior density $f_0(\theta)$.

Case A. $f_0(\theta)$ is the uniform density on $(0.4,\ 0.6)$, that is,

$$f_0(\theta) = 5 \quad \text{for } 0.4 < \theta < 0.6.$$

Case B. $f_0(\theta)$ is the triangular density on $(0.4,\ 0.6)$ with $\theta = 0.5$ as its mode, that is,

$$f_0(\theta) = \begin{cases} 100(\theta - 0.4) & \text{for } 0.4 < \theta \leq 0.5 \\ 100(0.6 - \theta) & \text{for } 0.5 < \theta < 0.6. \end{cases}$$

How should you adjust the prior when you have learned about the outcome E that Alassi has won the best-of-five contest? For that we need the conditional probability $L(E \mid \theta)$ that Alassi is the first player to win three matches in the contest when the probability of Alassi winning any given match has the fixed value θ. Assuming independence between the matches, it is easily verified that

$$L(E \mid \theta) = \theta^3 + \binom{3}{2}\theta^2(1-\theta)\theta + \binom{4}{2}\theta^2(1-\theta)^2\theta.$$

On the basis of the information that Alassi has won the best-of-five contest the prior density $f_0(\theta)$ is updated to $f(\theta \mid E)$, where the posterior density $f(\theta \mid E)$ is proportional to $L(E \mid \theta) f(\theta)$. In this example the proportionality constant is easily determined by numerical integration. The proportionality constant has the value 2 both for Case A and Case B. The posterior density is now completely specified. We can now answer the question how the prior probability of Alassi being the better player changes when we learn that Alassi has won the best-of-five contest. The prior probability that Alassi is the better player is equal to

0.5 for both Case A and Case B. The posterior probability of Alassi being the better player is given by $\int_{0.5}^{0.6} f(\theta \mid E)\, d\theta$ and has the value 0.5925 for Case A and the value 0.5620 for Case B.

Suppose that not only are you told that Alassi won the best-of-five contest but you are also informed that Alassi won this contest by winning the first three matches. Then the formula for $L(E \mid \theta)$ becomes $L(E \mid \theta) = \theta^3$ and the calculations yield that the posterior probability of Alassi being the better player has the value 0.6452 for Case A and the value 0.5984 for Case B.

In physics, a fundamental problem is a probabilistic assessment of the true value of a physical quantity from experimental data. A question such as "what is the probability that the true value of the mass of the top quark is between two specified bounds?" is a natural question for physicists. The following example shows how useful Bayesian analysis can be for the probabilistic assessment of the true value of a physical quantity. Also, in this example we encounter the concept of a Bayesian confidence interval. Such an interval is a probability interval for the true value of an unknown quantity.

Example 13.14 You want to know more precisely the true weight of a particular steel beam. The weights of the beams from the supplier of your beam are uniformly distributed between 71 and 75 kg. Your steel beam is weighed several times on a scale. The scale is known to give results that are normally distributed without bias but with a standard deviation of $\sqrt{2}$ kg. The ten measurements of the weighing process are 72.883, 71.145, 73.677, 73.907, 71.879, 75.184, 75.182, 73.447, 73.963 and 73.747 kg. How should you update the prior density of the weight of your steel beam?

Solution. The prior density $f_0(\theta)$ of the weight of the beam is given by $f_0(\theta) = 0.25$ for $71 < \theta < 75$. For ease of notation, denote the ten measurements of the weight by $x_1, \ldots, x_{10}$ and denote by $\mathbf{x}$ the data vector $\mathbf{x} = (x_1, \ldots, x_{10})$. The likelihood density function $L(\mathbf{x} \mid \theta)$ is given by $h(x_1 \mid \theta) \times \cdots \times h(x_{10} \mid \theta)$, where $h(x \mid \theta)$ is the $N(\theta, 2)$ density $(\sqrt{2}\sqrt{2\pi})^{-1} e^{-\frac{1}{2}(x-\theta)^2/2}$. This leads to

$$L(\mathbf{x} \mid \theta) = (2\sqrt{\pi})^{-10}\, e^{-\frac{1}{2}\sum_{i=1}^{10}(x_i-\theta)^2/2} \quad \text{for all } \theta.$$

The posterior density $f(\theta \mid \mathbf{x})$ is proportional to $L(\mathbf{x} \mid \theta) f(\theta)$. Using numerical integration, we find the value 4.407×10^7 for the proportionality constant. This completes the specification of the posterior density $f(\theta \mid \mathbf{x})$ of the weight of the beam. The graph of $f(\theta \mid \mathbf{x})$ is displayed in Figure 13.1. Setting the derivative of $f(\theta \mid \mathbf{x})$ equal to zero yields that the posterior density reaches its maximum at $\theta = (1/10)\sum_{i=1}^{10} x_i = 73.501$. The mode of the posterior density is called the Bayesian maximum a posteriori estimate. Next we determine an interval which contains 95% (say) of the area under the curve of the posterior density.

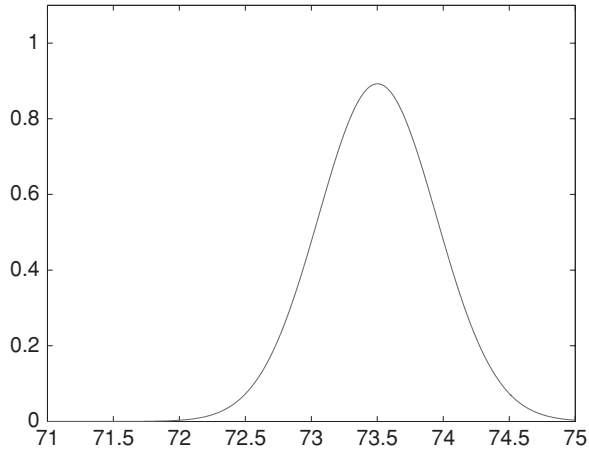


Fig. 13.1. The posterior density of the weight of the beam

A Bayesian 95% confidence interval for θ could be any interval (θ_1, θ_2) for which

$$\int_{\theta_1}^{\theta_2} f(\theta \mid \mathbf{x})\, d\theta = 0.95.$$

Among all such intervals, one might choose the one determined by the 0.025 percentile and the 0.975 percentile of $f(\theta \mid \mathbf{x})$. In other words, for $p = 0.025$ and $p = 0.975$, one numerically solves for α in the equation

$$\int_{71}^{\alpha} f(\theta \mid \mathbf{x})\, d\theta = p.$$

This gives the 0.025 percentile 72.625 and the 0.975 percentile 74.375, and so (72.625, 74.375) is a Bayesian 95% confidence interval for the weight of the beam. The Bayesian confidence interval tells you that you believe the true weight of the beam lies between 72.625 and 74.375 with probability 0.95. This is an interpretation which is in accordance with how most people would interpret a confidence interval. A frequentist confidence interval cannot be interpreted in this way but refers to the statistical properties of the way the interval is constructed, see Section 5.7.

Problem 13.41 Verify that the likelihood function $L(\mathbf{x} \mid \theta)$ from Example 13.14 is proportional to $e^{-\frac{1}{2}(\theta-\overline{x})^2/0.2}$ for all θ and so the posterior density requires the data $x_1, \ldots, x_{10}$ only through their average $\overline{x} = (1/10)\sum_{i=1}^{10} x_i$. *Hint*: write $x_i - \theta$ as $x_i - \overline{x} + \overline{x} - \theta$ and work out $\sum_{i=1}^{10}(x_i - \theta)^2$. Calculate a Bayesian

95% confidence interval for the true weight of the beam when the prior density of the weight is a normal density with an expected value of 73 kg and a standard deviation of 0.7 kg. What value of θ maximizes the posterior density?

Problem 13.42 Consider the continuous version of Example 8.13. Take for your prior density both the uniform density on (0.2, 0.8) and the triangular density on (0.2, 0.8) with $\theta = 0.50$ as mode. What value of θ maximizes the posterior density and what is a Bayesian 95% confidence interval for the posterior probability that the new treatment is successful?

Problem 13.43 An astronomer will perform an experiment to measure the distance of a star from the Earth. The error in the measurement of the true value of the distance is given by an $N(0, \sigma^2)$ density with $\sigma = 20$ light years. The astronomer's prior belief about the true value of the distance is described by a normal distribution with an expected value of 150 light years and a standard deviation of 25 light years. The measurement gives a distance of 140 light years. Verify that the posterior density is a normal density. What is the Bayesian maximum a posteriori probability estimate for the distance of the star and what is a 95% Bayesian confidence interval for the distance?

Problem 13.44 A pollster will conduct a poll attempting to predict whether the Liberal Party or the National Party would win the coming election in a country with two parties. On the basis of a previous poll, the pollster's prior belief about the proportion of Liberal voters is a beta(a, b) density with $a = 474$ and $b = 501$. In the new poll 527 voters are for the Liberal Party and 573 voters are for the National Party. What is a Bayesian 95% confidence interval for the proportion of Liberal voters? What is the posterior probability that the Liberal Party will win the election, assuming that the poll is representative of the population at the time of the election?

Problem 13.45 The lifetime of a light bulb has an exponential density. The parameter of this density is a fixed but unknown constant. Your prior belief about the true parameter value is described by a gamma density with shape parameter α and scale parameter λ. You decide to observe m light bulbs over the time interval $(0, T)$. It appears that r of the m light bulbs have failed at the respective times $t_1, \ldots, t_r$, while the other $m - r$ light bulbs are still functioning at time T. What is the posterior density of the true parameter value of the lifetime density?

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. This chapter also gives a proof of the strong law of large numbers, using moment-generating functions together with so-called Chernoff bounds. Finally, the strong law of large numbers is used to establish the powerful renewal-reward theorem for stochastic processes having the property that the process probabilistically restarts itself at certain points in time.

14.1 Generating functions

We first introduce the concept of generating function 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

$$G_X(z) = E\left(z^X\right),$$

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,

$$E(X) = G'_X(1) \quad \text{and} \quad E\left(X^2\right) = G''_X(1) + G'_X(1).$$

Example 14.1 Suppose that the random variable X has a Poisson distribution with expected value λ. Verify that the generating function of X is given by

$$G_X(z) = e^{-\lambda(1-z)}, \qquad |z| \leq 1.$$

What are the expected value and the standard deviation of X?

Solution. Applying the definition of generating function and using the series expansion $e^x = \sum_{n=0}^{\infty} x^n/n!$, we find

$$\sum_{k=0}^{\infty} z^k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda z)^k}{k!} = e^{-\lambda} e^{\lambda z},$$

as was to be verified. 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 9.8.

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

$$E\left(z^{X+Y}\right) = E\left(z^X z^Y\right) = E\left(z^X\right) E\left(z^Y\right),$$

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 $X + Y$?

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. Let the random variable X be defined as the number of records. Argue that X can be represented as the sum of independent indicator variables (see also Example 9.5 in Section 9.4). Next determine the generating function of X. What are the expected value and variance of X?

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 You first roll a single die. Next you simultaneously roll as many dice as the face value shown up on the first roll and sum the face values shown up on the simultaneous roll. What is the generating function of the sum and how do you compute the probability mass function of the sum? Also, answer these questions when you sum only the face values that are larger than the face value shown up on the first roll.

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. For example, in the coupon collector's problem from Section 3.2, consider the random variable X representing the number of bags of chips that must be purchased to get a complete set of n distinct flippos. This random variable X can be written as the sum of n independent random variables that are geometrically distributed with respective parameters $a_1, \ldots, a_n$, where $a_i = \frac{n-i+1}{n}$ for all i. It is left to the reader to verify that

$$G_X(z) = \frac{a_1 a_2 \cdots a_n z^n}{(1 - z + p_1 z)(1 - z + p_2 z) \cdots (1 - z + p_n z)}.$$

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.

The following example shows that the probability distribution of the waiting time for success runs in Bernoulli trials can also be studied through the generating function approach.

Example 14.3 For a sequence of independent Bernoulli trials with success probability p, let the random variable X be the number of trials it would take to have a run of r successes in a row. What are the expected value and the variance of X?

Solution. For ease of presentation, let us first take $r = 3$. Also, let X_1, X_2 and X_3 be random variables having the same distribution as X. By appropriate conditioning on the outcomes of the first three trials, it follows that the waiting time X is distributed as $1 + X_1$ with probability q, as $2 + X_2$ with probability pq, as $3 + X_3$ with probability p^2q, and as 3 with probability p^3. Here $q = 1 - p$. This leads to

$$E(z^X) = qzE(z^X) + pqz^2E(z^X) + p^2qz^3E(z^X) + p^3z^3.$$

Hence $G_X(z) = E(z^X)$ is given by $G_X(z) = p^3z^3/(1 - (qz + pqz^2 + p^2qz^3))$. Noting that $qz(1 + pz + p^2z^2) = qz(1 - p^3z^3)/(1 - pz)$, it is a matter of simple algebra to rewrite the generating function $G_X(z)$ as

$$G_X(z) = \frac{p^3z^3(1 - pz)}{1 - z + qp^3z^4}.$$

The above derivation immediately extends to the general case of a run of r successes in a row. We then have

$$G_X(z) = \frac{p^rz^r(1 - pz)}{1 - z + qp^rz^{r+1}}.$$

Using the formulas $E(X) = G'_X(1)$ and $E(X^2) = G''_X(1) + G'_X(1)$, it is a matter of algebra to conclude that

$$E(X) = \frac{1 - p^r}{qp^r} \quad \text{and} \quad \text{var}(X) = \frac{1 - p^{2r+1} - qp^r(2r + 1)}{q^2p^{2r}}.$$

As illustration, consider the coin-tossing experiment with $p = q = 0.5$. Then, $E(X) = 62$ and $\sqrt{\text{var}(X)} = 58.22$ for $r = 5$. The smallest n for which $P(X \leq n) \geq 0.5$ is $n = 44$ when $r = 5$. If $p = 0.4463$ and $q = 0.5537$, as is the case in a variant of the casino game of baccarat, then $E(X) = 100.19$ and $\sqrt{\text{var}(X)} = 101.95$ for $r = 5$. The smallest n for which $P(X \leq n) \geq 0.5$ is $n = 71$ in the game of baccarat with $r = 5$.

As in the coupon collector's problem, the probability distribution of the waiting time X until the first run of r successes in a row can be numerically obtained by inversion of the generating function. The distribution of X also determines the probability of seeing a run of r successes in a row somewhere during n trials of the Bernoulli experiment. This probability is given by $P(X \leq n)$.

Both for the coupon collector's problem and the problem of success runs in Bernoulli trials, the most simple way to compute the probability distribution of the random variable X is through matrix multiplications of the transition matrix of a so-called absorbing Markov chain. This approach will be discussed in Section 15.3.

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 $E(X)$?

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 reactions. 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 probability. 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 probability.

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

$$M_X(t) = 1 + t E(X) + t^2 \frac{E(X^2)}{2!} + t^3 \frac{E(X^3)}{3!} + \cdots$$

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 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

$$E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E(e^{tX})E(e^{tY}).$$

Example 14.4 Suppose that the random variable X has an $N(\mu, \sigma^2)$ distribution. Verify that

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}, \qquad -\infty < t < \infty.$$

What are the expected value and variance of the random variable X?

Solution. 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

$$E(e^{tX}) = E(e^{t(\mu+\sigma Z)}) = e^{t\mu} E(e^{t\sigma Z}) = e^{t\mu} e^{\frac{1}{2}\sigma^2 t^2}.$$

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*

$$\mu = a_1\mu_1 + \cdots + a_n\mu_n \quad \text{and} \quad \sigma^2 = a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2.$$

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

$$\begin{aligned} E[e^{t(a_1X_1+a_2X_2)}] &= E(e^{ta_1X_1})E(e^{ta_2X_2}) \\ &= e^{\mu_1a_1t+\frac{1}{2}\sigma_1^2(a_1t)^2}e^{\mu_2a_2t+\frac{1}{2}\sigma_2^2(a_2t)^2} \\ &= e^{(a_1\mu_1+a_2\mu_2)t+\frac{1}{2}(a_1^2\sigma_1^2+a_2^2\sigma_2^2)t^2}, \end{aligned}$$

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 holds for the gamma distribution (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. Verify that

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha, \qquad t < \lambda.$$

Solution. 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.

An exponential random variable with parameter λ has the moment-generating function $\lambda/(\lambda - t)$ and so the moment-generating function of the sum of n independent exponential random variables each having the same parameter λ has the moment-generating function $[\lambda/(\lambda - t)]^n$. It now follows from the result in Example 14.5 that the sum of n independent exponential random variables each having the same parameter λ is gamma distributed with shape parameter n and scale parameter λ. This particular gamma distribution is often called the Erlang distribution with parameters n and λ.

Rule 14.5 *Let $Z_1, \ldots, Z_n$ be independent random variables each having a standard normal distribution. Define the chi-square 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

$$\begin{aligned} E(e^{tZ^2}) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx^2} e^{-\frac{1}{2}x^2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2t)x^2}\, dx \\ &= \frac{1}{\sqrt{1-2t}} \frac{1}{(1/\sqrt{1-2t})\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2/(1/\sqrt{1-2t})^2}\, dx \\ &= \frac{1}{\sqrt{1-2t}}, \qquad t < \frac{1}{2}. \end{aligned}$$

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 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 λ. Denote by α_i the shape

parameter of the density of X_i. Verify that $X_1 + \cdots + X_n$ has a gamma density with shape parameter $\alpha_1 + \cdots + \alpha_n$ and scale parameter λ.

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.[†]

(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?

14.3 Chernoff bound

Let X be a random variable for which the moment-generating function $M_X(t)$ exists. The *Chernoff bound* states that

$$P(X \geq c) \leq \min_{t>0} \left[e^{-ct} M_X(t)\right] \qquad \text{for any constant } c,$$

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 Chernoff bound follows directly from *Markov's inequality*, which states that

$$P(U \geq a) \leq \frac{1}{a} E(U) \qquad \text{for any constant } a > 0$$

[†] 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.

when U is a nonnegative random variable. To get the Chernoff bound from this inequality, note that

$$P(X \geq c) = P(tX \geq tc) = P(e^{tX} \geq e^{tc}) \qquad \text{for any } t > 0.$$

Applying Markov's inequality with $U = e^{tX}$ and $a = e^{ct} > 0$, we get

$$P(X \geq c) \leq e^{-ct} M_X(t) \quad \text{for any } t > 0,$$

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 directly be obtained 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 Suppose that the random variable X has the standard normal distribution. Verify that

$$P(X \geq c) \leq e^{-\frac{1}{2}c^2} \qquad \text{for any constant } c > 0.$$

Solution. To verify this result, we use the Chernoff bound. 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 $P(X \leq c) \leq \min_{t<0}\left[e^{-ct} M_X(t)\right]$ 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

$$P\big(X \geq (1 + \delta)np\big) \leq \left[\left(\frac{p}{a}\right)^a \left(\frac{1 - p}{1 - a}\right)^{1-a}\right]^n,$$

where $a = (1 + \delta)p$. (*Remark*: the upper bound can be shown to be smaller than or equal to $e^{-2p^2\delta^2 n}$.)

14.4 Strong law of large numbers revisited

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

$$P\left(\{\omega : \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k(\omega) = \mu\}\right) = 1,$$

where the symbol ω represents an outcome in the underlying sample space on which the process $\{X_1, X_2, \ldots\}$ is defined. This type of convergence is called convergence with probability one or almost sure convergence. The terminology of an "almost sure" event means that realizations not in this event are theoretically possible but will not happen in reality. To prove the strong law of large numbers, 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

$$P\left(\frac{1}{n}\sum_{k=1}^{n} X_k \geq \mu + \epsilon\right) \leq \min_{t>0} e^{-n(\mu+\epsilon)t}[M_X(t)]^n = \min_{t>0}[e^{-(\mu+\epsilon)t}M_X(t)]^n$$

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 $0 \le \lambda < 1$. In the same way it can be verified that $0 \le \eta < 1$, 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

$$P\left(\frac{1}{n}\sum_{k=1}^{n} X_k \le \mu - \epsilon\right) \le \eta^n \qquad \text{for } n = 1, 2, \ldots.$$

Thus, letting $A_n = \{\omega : |\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) - \mu| \ge \epsilon\}$, we have

$$\sum_{n=1}^{\infty} P(A_n) \le \sum_{n=1}^{\infty} (\lambda^n + \eta^n) = \frac{\lambda}{1-\lambda} + \frac{\eta}{1-\eta} < \infty.$$

Invoking now the Borel–Cantelli lemma (see Problem 7.19 in Chapter 7), we find that $P(C_\epsilon) = 0$, where

$$C_\epsilon = \left\{\omega : |\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) - \mu| \ge \epsilon \text{ for infinitely many values of } n\right\}.$$

In other words, letting

$$\overline{C_\epsilon} = \left\{\omega : |\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) - \mu| < \epsilon \text{ for all } n \text{ sufficiently large}\right\},$$

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 $(\epsilon_k, k \ge 1)$ 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.

As the above proof clearly demonstrates, we can conclude from the *strong* law of large numbers that, no 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

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right| \ge \epsilon\right) = 0$$

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!).

The above proof of the strong law of large numbers used the assumption that the moment-generating function of the X_k exists. However, this assumption can be dropped. It suffices to require that the expected value of the X_k exists and is finite. If the X_k are nonnegative, it is not even required that the expected value of the X_k is finite. The proofs of these results are beyond the scope of this book. A nice application of the strong law of large numbers is the so-called renewal-reward theorem. This theorem is extremely useful in applied probability.

Renewal-reward theorem

Let us first introduce the notion of a stochastic process. A *stochastic process* $\{S_t,\ t \in \Gamma\}$ is a collection of random variables S_t, indexed by an ordered time variable $t \in \Gamma$. The process is said to be a continuous-time process if Γ is the set of nonnegative reals and is said to be a discrete-time process if Γ is the set of nonnegative integers. Many stochastic processes have the property of regenerating themselves at certain points in time so that the continuation of the process after any regeneration epoch is a probabilistic replica of the process starting at time zero and is independent of the behavior of the process before the regeneration epoch. To illustrate this, consider the following example. A light bulb having a continuously distributed lifetime is replaced by a new one upon failure or upon reaching the age T, whichever occurs first. For any $t \geq 0$, let the random variable S_t denote the age of the bulb in use at time t. Then the process $\{S_t\}$ is a continuous-time stochastic process that regenerates itself each time a new bulb is installed.

Let us now consider a regenerative stochastic process $\{S_t\}$. For clarity of presentation, we take a continuous-time process and assume that the process starts in the regeneration state at time zero. The time interval between two successive regeneration epochs is called a *cycle*. Let the random variables $L_1, L_2, \ldots$ denote the lengths of the successive cycles. These random variables are independent and identically distributed. It is assumed that the expected value of the L_k are finite and positive. Imagine now that rewards are earned

[†] 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.

or costs are incurred during the cycles. In the replacement example, you may think of a fixed cost $c_1 > 0$ for a planned replacement and a fixed cost $c_2 > c_1$ for a failure replacement. Let the random variable R_n denote the reward earned during the nth cycle. The assumption is made that $R_1, R_2, \ldots$ are independent and identically distributed random variables, where the expected value of the R_k exists and is finite. For any $t > 0$, define the random variable

$$R(t) = \text{the cumulative reward earned up to time } t.$$

Rule 14.6 (*Renewal-reward theorem*) *With probability* 1,

$$\lim_{t\to\infty} \frac{R(t)}{t} = \frac{E(R_1)}{E(L_1)}.$$

In words, with probability 1, *the long-run average reward per unit time equals the expected reward in a cycle divided by the expected length of a cycle.*

The proof of this important theorem goes as follows. For any $t > 0$, define the random variable $N(t)$ as the number of cycles completed up to time t. Then, with probability 1,

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{E(L_1)}.$$

To see this, note that $L_1 + \cdots + L_{N(t)} \le t < L_1 + \cdots + L_{N(t)+1}$, by the definition of $N(t)$. Also, with probability 1, the random variable $N(t)$ tends to ∞ as $t \to \infty$, being a corollary of $P(L_1 + \cdots + L_n < \infty) = 1$ for all $n \ge 1$. The limit result $\lim_{t\to\infty} N(t)/t = 1/E(L_1)$ now follows by letting $t \to \infty$ in

$$\frac{L_1 + \cdots + L_{N(t)}}{N(t)} \le \frac{t}{N(t)} < \frac{L_1 + \cdots + L_{N(t)+1}}{N(t)}$$

and applying the strong law of large numbers to the sequence $\{L_k\}$. The remainder of the proof now proceeds as follows. Assuming for ease that the rewards are nonnegative, we have for any $t > 0$

$$\sum_{k=1}^{N(t)} R_k \le R(t) \le \sum_{k=1}^{N(t)+1} R_k.$$

This inequality implies

$$\frac{\sum_{k=1}^{N(t)} R_k}{N(t)} \times \frac{N(t)}{t} \le \frac{R(t)}{t} \le \frac{\sum_{k=1}^{N(t)+1} R_k}{N(t)} \times \frac{N(t)}{t}.$$

Letting $t \to \infty$, applying the strong law of large numbers to the R_k and using $\lim_{t\to\infty} N(t)/t = 1/E(L_1)$, we obtain the desired result. If the rewards R_n can take on both positive and negative values, use the decomposition

$R_n = R_n^+ - R_n^-$ with $R_n^+ = \max(R_n, 0)$ and $R_n^- = -\min(R_n, 0)$. Next repeat the above proof for both the nonnegative R_n^+ and the nonnegative R_n^-.

Problem 14.19 What is the long-run average cost in the replacement example discussed above?

Problem 14.20 Customer orders arrive at a production facility, where the interarrival times of successive orders are independent random variables having the same continuous distribution with expected value η. Processing of the orders starts as soon as N orders have accumulated. All orders are simultaneously processed and the processing time is negligible. A fixed cost of $K > 0$ is incurred for each production set-up. Also, for each order there is a holding cost at a rate of $h > 0$ per unit time the order is waiting to be processed. What are the regeneration epochs of the process describing the number of orders present? What is the long-run average cost per unit time?

Problem 14.21 A communication channel is alternately on and off. An on-time starts at time zero. The on-times are independent random variables and are beta distributed with density $f(x) = 6x(1-x)$ for $0 < x < 1$. Any off-time is dependent on the preceding on-time and is equal to $x^2\sqrt{x}$ if the realized value of the on-time is x. What are the regeneration epochs? What is the long-run fraction of time the channel is on? *Hint*: assume that a reward at rate 1 is earned while the channel is on.

14.5 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

$$\lim_{n\to\infty} P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2}\, dy \qquad \text{for all } x.$$

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

$$U_i = \frac{X_i - \mu}{\sigma}, \qquad i = 1, 2, \ldots.$$

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 $M_{Z_n}(t) = E\left(e^{tZ_n}\right)$ 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,

$$\lim_{n\to\infty} M_{Z_n}(t) = E\left(e^{tZ}\right)$$

when Z is a standard normal random variable. This result implies

$$\lim_{n\to\infty} P\left(Z_n \le x\right) = P(Z \le x) \qquad \text{for all } x,$$

by 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

$$E\left(e^{tZ_n}\right) = E\left(e^{t(U_1+\cdots+U_n)/\sqrt{n}}\right) = E\left(e^{(t/\sqrt{n})U_1}\right)\cdots E\left(e^{(t/\sqrt{n})U_n}\right)$$

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 obtain the desired result

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{\frac{1}{2}t^2}.$$

14.6 Law of the iterated logarithm

The strong law of large numbers and the central limit theorem are the most important and fundamental limit theorems in probability theory. A good candidate for the third place is the law of the iterated logarithm. To formulate this law, let $X_1, X_2, \ldots$ be a sequence of independent random variables having a common probability distribution with finite mean μ and finite standard deviation σ. For $n = 1, 2, \ldots$, the partial sum S_n is defined by

$$S_n = \sum_{k=1}^{n}(X_k - \mu).$$

The central limit theorem tells us that suitably normalized sums of independent random variables can be approximated by a normal distribution. That is, the random variable $S_n/(\sigma\sqrt{n})$ is approximately $N(0, 1)$ distributed for n large. The values of $S_n/(\sigma\sqrt{n})$ as a function of n fluctuate, but fluctuate slowly because they are heavily correlated. However, $S_n/(\sigma\sqrt{n})$ as a function of n may take on, and will take on, arbitrarily large positive and negative values, though the occurrence of this is a rare event. The law of the iterated logarithm provides precise bounds on the fluctuations of $S_n/(\sigma\sqrt{n})$ as n grows. The name of this law comes from the iterated (double) logarithm that appears in the expression for the bounds. For almost every possible realization ω of the stochastic process $\{S_n\}$, there is a finite integer $n(\omega)$ so that

$$-\sqrt{2\ln(\ln(n))} < \frac{S_n(\omega)}{\sigma\sqrt{n}} < \sqrt{2\ln(\ln(n))} \qquad \text{for all } n \geq n(\omega).$$

Further, for any $c \in (-\sqrt{2}, \sqrt{2})$, the sequence $\{S_n(\omega)/[\sigma\sqrt{2n\ln(\ln(n))}]\}$ takes on infinitely often a value between c and $\sqrt{2}$ and infinitely often a value between c and $-\sqrt{2}$. The set of all realizations ω for which these statements are true has probability one. The function $\sqrt{2\ln(\ln(n))}$ bordering the fluctuations of the $S_n/(\sigma\sqrt{n})$ is not bounded but increases incredibly slowly as n grows. The function has the value 2.292 when $n = 10^6$ and this value has only increased to 2.505 when $n = 10^{10}$. The proof of the law of the iterated logarithm is quite

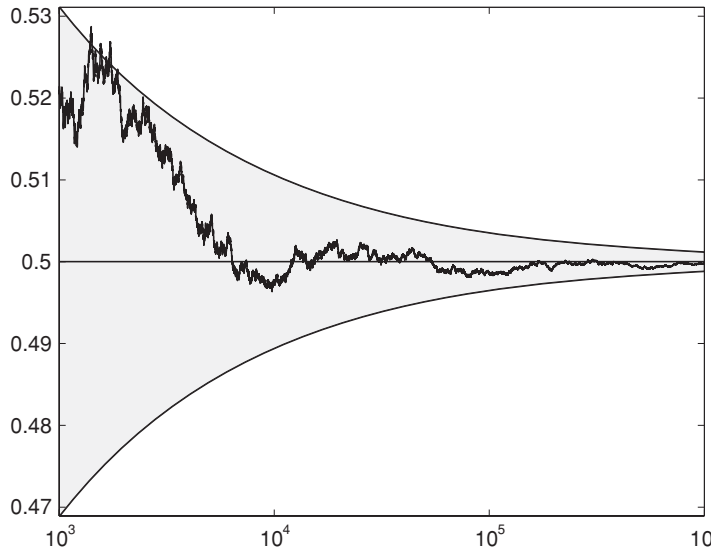


Fig. 14.1. The law of iterated algorithm in action.

technical and will not be given. The law provides a beautiful and fascinating result.

As a concrete example, consider the coin-tossing experiment. What can be said about the fluctuations of the random process that describes the proportion of heads obtained when a fair coin is repeatedly tossed? The coin-tossing experiment can be modeled by a sequence of independent random variables $X_1, X_2, \ldots$ with $P(X_i = 0) = P(X_i = 1) = 0.5$ for all i. The expected value and the standard deviation of the X_i are given by $\mu = 0.5$ and $\sigma = 0.5$. The proportion of heads after n tosses is

$$Y_n = \frac{X_1 + \cdots + X_n}{n}.$$

Then, for almost every realization ω of the coin-tossing process, there is a finite integer $n(\omega)$ so that

$$\frac{1}{2} - \frac{1}{2\sqrt{n}}\sqrt{2\ln(\ln(n))} < Y_n(\omega) < \frac{1}{2} + \frac{1}{2\sqrt{n}}\sqrt{2\ln(\ln(n))} \quad \text{for all } n \geq n(\omega),$$

by the law of the iterated logarithm and the fact that $n(Y_n - \mu) = S_n$. As an illustration, Figure 14.1 displays the result of a simulation of the coin-tossing process. The relative frequency of heads is simulated for 1,000,000 tosses of a fair coin. Beginning with the 1,000th toss, the relative frequency of heads is plotted against a log scale for the number of tosses. The simulated sample

path nicely demonstrates that the relative frequency stays within the parabolic bounds of the iterated law of the logarithm when the number of tosses increases.

The law of the iterated algorithm is crucial for establishing the form of the best strategy in the following optimal stopping problem for the coin-tossing process. Toss a fair coin repeatedly and stop whenever you want, receiving as a reward the proportion of heads obtained at the time you stop. What is the best strategy to maximize your expected payoff? It can be proved that the optimal stopping rule is characterized by integers $\beta_1, \beta_2, \ldots$ such that you stop after the nth toss when the number of heads minus the number of tails is larger than or equal to β_n. Obviously, $\beta_1 = 1$. It has also been proved that $\beta_n/\sqrt{n}$ tends to the constant $0.83992\ldots$ as n gets very large. The approximation $\beta_n \approx 0.83992\sqrt{n}$ is already a useful approximation for moderate values of n. How to get the true values of the β_n is still an open problem. However, numerical approximation methods can be applied by putting a bound on the number of tosses allowed. To give a few values, numerical studies suggest that $\beta_5 = 3$, $\beta_{10} = 4$, $\beta_{25} = 5$, $\beta_{50} = 6$, and $\beta_{100} = 8$. Also, numerical investigations indicate that the maximal expected payoff has the value 0.79295.

15

Discrete-time 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 price of IBM stock at the end of successive trading days. Many such systems have the property that the current state alone contains sufficient information to give the probability distribution of the next state. 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 to analyze 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 time-dependent 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. Next, 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. Finally, we discuss the method of Markov chain Monte Carlo simulation which has revolutionized the field of Bayesian statistics and many other areas of science.

15.1 Markov chain 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 steps. 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

$$I = \{(x, y) : x, y \text{ integer and } -L \leq x, y \leq L\},$$

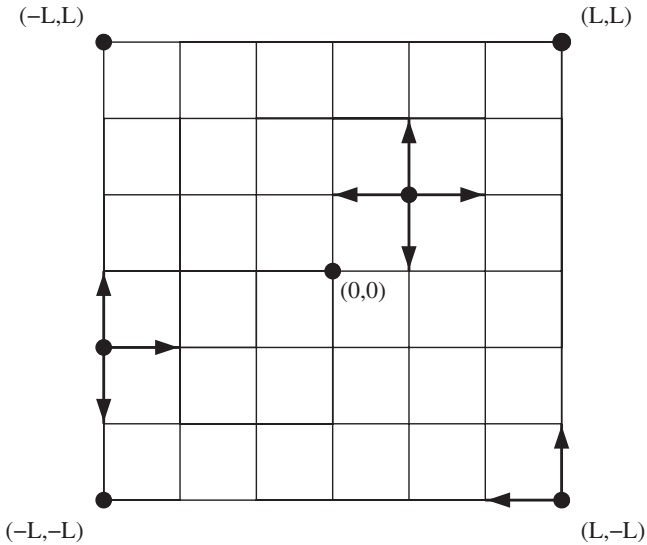


Fig. 15.1. The drunkard's walk.

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 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 $t = n$. The sequence $X_0, X_1, \ldots$ is called a discrete-time stochastic process. 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*

$$P(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n)$$
$$= P(X_{n+1} = i_{n+1} \mid X_n = i_n).$$

The term $P(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n)$ 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 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. At any time $t = n$ the process essentially forgets how it got into the state X_n. It is not true that the state X_{n+1} at the next time point $t = n + 1$ is independent of $X_0, \ldots, X_{n-1}$, but all of the dependency is captured by X_n.

The Markov chain approach is a very powerful tool in applied probability. Using the concept of state and choosing the state in an appropriate way, numerous probability problems can be solved by Markov chain methods.

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 licence 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 licence 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 third example deals with an inventory problem and the modeling of this problem is more involved than that of the previous two problems.

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 case 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 Time-dependent 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 time-dependent 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 probability and invoking the Markovian property, we have that $p_{ij}^{(n)} = P(X_n = j \mid X_0 = i)$ is equal to

$$\begin{aligned} &\sum_{k \in I} P(X_n = j \mid X_0 = i, X_{n-1} = k) P(X_{n-1} = k \mid X_0 = i) \\ &= \sum_{k \in I} P(X_n = j \mid X_{n-1} = k) P(X_{n-1} = k \mid X_0 = i). \end{aligned}$$

By the assumption of time homogeneity, the last expression is equal to $\sum_{k \in I} p_{kj} p_{ik}^{(n-1)}$, proving the desired result.

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

$$\mathbf{P}^{(n)} = \mathbf{P} \times \mathbf{P} \times \cdots \times \mathbf{P} = \mathbf{P}^n.$$

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

$$\begin{array}{lc} \text{from\textbackslash to} & \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{array}{l} 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} 0.70 & 0.10 & 0.20 \\ 0.50 & 0.25 & 0.25 \\ 0.40 & 0.30 & 0.30 \end{pmatrix}. \end{array}$$

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

$$\sigma^2[V_{ij}(n)] = \sum_{t=1}^{n} p_{ij}^{(t)}(1 - p_{ij}^{(t)}) + 2\sum_{t=1}^{n}\sum_{u=t+1}^{n} [p_{ij}^{(t)} p_{jj}^{(u-t)} - p_{ij}^{(t)} p_{ij}^{(u)}].$$

Hint: use the fact that $P(I_t = 1,\ I_u = 1 \mid X_0 = i) = p_{ij}^{(t)} p_{jj}^{(u-t)}$ 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

[†] 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 $n \geq 1$.

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 to analyze this problem (and many other problems in applied probability!). It leads to an algorithmic 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 does one 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

$$P(R > r) = 1 - p_{06}^{(r)} \qquad \text{for } r = 1, 2, \ldots.$$

Put it differently, $p_{06}^{(r)}$ is the probability of reaching the absorbing state 6 in r or less steps. The numerical of the probability $P(R > r)$ is 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 $j = 1, \ldots, 6$ gives 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.

Absorbing Markov chains are very useful to analyze success runs. This is illustrated with the following example (see the Problems 15.15 and 15.19 for success runs in Bernoulli trials).

Example 15.7 Chess grandmaster Boris Karparoff has entered a match with chess computer Deep Blue. The match will continue until either Karparoff or Deep Blue has won two consecutive games. For the first game as well as any game ending in a draw, it holds that the next game 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 game 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 match will last for longer than 10 games? What is the probability that Karparoff will be the final winner, and what is the expected value of the duration of the match?

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, 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

$$\begin{array}{l} \text{from /to} \\ 0 \\ (1,K) \\ (1,D) \\ (2,K) \\ (2,D) \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & (1,K) & (1,D) & (2,K) & (2,D) \end{array} \\ \begin{pmatrix} 0.3 & 0.4 & 0.3 & 0 & 0 \\ 0.25 & 0 & 0.25 & 0.5 & 0 \\ 0.2 & 0.3 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

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 $P(X_r \neq (2, K), (2, D) \mid X_0 = 0)$ is equal to the probability $1 - P(X_r = (2, K) \mid X_0 = 0) - P(X_r = (2, D) \mid X_0 = 0)$, 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 probability, 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.

To find the expected value of the duration of the match, we can use the relation $E(L) = \sum_{r=0}^{\infty} P(L > r)$, see Problem 9.13. However, a more elegant the approach is to set 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 expectation, 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 $\mu_0 = \sum_{r=1}^{\infty} r\, P(L = r)$?

In the following two examples we use absorbing Markov chains to shed light on remarkable events in the game of craps and in the lotto.

Example 15.8 In Section 3.1.5 we have given the story of a woman who rolled the dice 154 consecutive times in a single turn of the casino game of craps before she "sevened out" by rolling a seven. What is the probability of 154 or more craps rolls in a single turn of the shooter? What is the expected number of craps rolls in a single turn?

Solution. Let us repeat the rules for the shooter rolling the two dice in craps. The shooter's turn begins with come-out rolls. These rolls continue until the dice add up to 4, 5, 6, 8, 9, or 10, which establishes the shooter's "point". Once this point is established, the game enters a second stage. The shooter rolls until throwing either the point, which ends the second stage and begins a new set of come-out rolls, or a seven, which ends the shooter's turn. Note that it takes a seven to end the shooter's turn, but the turn cannot end during a come-out roll. Some reflections show that the process of rolling the dice in a single turn can be described by an absorbing Markov chain. The shooter can be in any of five

possible states in craps before the next roll. State 0 means that the shooter is in the stage of come-out rolls, state 1 means that the shooter's point is 4 or 10, state 2 means that the shooter's point is 5 or 9, state 3 means that the shooter's point is 6 or 8, and state 4 means that the shooter's turn has ended by sevening out. The state 4 is absorbing. It readily follows from the rules for the shooter that the matrix $\mathbf{P} = (p_{ij})$ of the one-step transition probabilities of the Markov chain is given by (verify!)

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 12/36 & 6/36 & 8/36 & 10/36 & 0 \\ 3/36 & 27/36 & 0 & 0 & 6/36 \\ 4/36 & 0 & 26/36 & 0 & 6/36 \\ 5/36 & 0 & 0 & 25/36 & 6/36 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

The probability of not having sevened out after 153 rolls can be computed as $1 - p_{04}^{(153)}$ and is equal to is 1.789×10^{-10}. This is a probability of 1 in 5.59 billion. To find the expected number of dice rolls before sevening out, let μ_i be the expected number of remaining dice rolls before sevening out when the current state of the process is i. By the same arguments as in Example 15.7,

$$\begin{aligned} \mu_0 &= 1 + (12/36)\mu_0 + (6/36)\mu_1 + (8/36)\mu_2 + (10/36)\mu_3 \\ \mu_1 &= 1 + (3/36)\mu_0 + (27/36)\mu_1 + (6/36)\mu_4 \\ \mu_2 &= 1 + (4/36)\mu_0 + (26/36)\mu_2 + (6/36)\mu_4 \\ \mu_3 &= 1 + (5/36)\mu_0 + (25/36)\mu_3 + (6/36)\mu_4, \end{aligned}$$

where $\mu_4 = 0$. The solution of these linear equations is given by $\mu_0 = 8.526$, $\mu_1 = 6.842$, $\mu_2 = 7.010$, and $\mu_3 = 7.148$. Hence the expected number of craps rolls in a single turn is 8.526.

Example 15.9 In the first 1,240 draws of the UK National Lottery a record gap of length 72 appeared on 4th November 2000. The number 17 did not appear for 72 consecutive draws. In each draw of the lottery six different numbers are drawn from the numbers $1, 2, \ldots, 49$. What is the probability that some number will not appear during 72 or more consecutive draws in the next 1,240 draws of the lottery 6/49?

Solution. We first calculate the probability that a *particular* number r will not appear during 72 or more consecutive draws in the next 1,240 draws of the lottery 6/49. This can be done by using an absorbing Markov chain with state spate $I = \{0, 1, \ldots, 72\}$, where state i indicates the number of draws since the particular number r appeared for the last time. The state 72 is taken as an

absorbing state. The one-step transition probabilities of the Markov chain are given by

$$p_{i0} = \frac{6}{49} \quad \text{and} \quad p_{i,i+1} = \frac{43}{49} \quad \text{for } i = 0, 1, \ldots, 71,$$

using the fact that $\binom{1}{1}\binom{48}{5}/\binom{49}{6}$ gives the probability a particular number will appear in a given draw. Further, $p_{72,72} = 1$ and the other p_{ij} are zero. The probability of a particular number r not appearing during 72 or more consecutive draws in the next 1,240 draws is given by $p_{0,72}^{(1,240)}$. This probability has the value $p = 0.011797$. Next the desired probability can be approximated as follows. Define A_k as the event that there is *not* a gap of length 72 or more for the particular number k in 1,240 draws of the lottery. Then the desired probability is equal to $1 - P(A_1 A_2 \cdots A_{49})$. The events $A_1, A_2, \ldots, A_{49}$ are not independent but the dependence is weak enough to justify the approximation

$$P(A_1 A_2 \cdots A_{49}) \approx P(A_1)P(A_2) \cdots P(A_{49}) = (1 - p)^{49} = 0.5591.$$

This gives the approximate value 0.4409 for the desired probability. It appears that this is an excellent approximation. A simulation study with 150,000 runs of 1,240 draws gives the simulated value 0.4424.

Tabu probability

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 $\mathbf{P}$ in Example 15.4 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 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? Do you think the following game is fair? 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.

Problem 15.16 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 15.17 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 You have a grid with a total of 36 squares numbered as (i, j) for $i, j = 1, \ldots, 6$. You repeatedly roll two fair dice. If a roll of the dice gives the two numbers a and b, then you are allowed to scratch off one of the squares (a, b) or (b, a). How could you compute the probability mass function and the expected value of the number of rolls until you have scratched off all the 36 squares?

Problem 15.19 On August 18, 1913, black came up twenty-six times in a row on a roulette wheel in the casino of Monte Carlo. What is the probability that in 5 million spins of a European roulette wheel the same color will come up

26 or more times in a row? 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. At Caesar's Palace in Las Vegas on July 14, 2000, the number 7 came up six times in a row at roulette wheel #211. What is the probability that in 5 million spins of a American roulette wheel any of the numbers $1, 2, \ldots, 36$ will come up six or more times in a row? In American roulette the wheel is divided in 38 sections, numbered as $1, 2, \ldots, 36, 0$ and 00.

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.

Problem 15.21 A fair die is repeatedly rolled. What is the expected value of the number of rolls until a run of six different outcomes appears and what is the probability of getting such a run within 100 rolls of the die? What are the answers when restricting to the run 123456?

Problem 15.22 Solve Problem 2.18 from Chapter 2 by using an absorbing Markov chain. In playing the game Yahtzee, what is the probability of rolling five of a kind within three rolls of the dice if you want to get Yahtzee?

Problem 15.23 A fair die is rolled until the total score of the rolls exceeds 100. What is the probability mass function of the final score?

Problem 15.24 Three equally matched opponents decide to have a ping-pong tournament. Two people play against each other in each game. Drawing lots, it is determined who are playing the first game. The winner of a game stays on and plays against the person not active in that game. The games continue until somebody has won two games in a row. What is the probability that the person not active in the first game is the ultimate winner?

Problem 15.25 A queue of 50 people is waiting at a box office in order to buy a ticket. The tickets cost five dollars each. For any person, there is a probability of $\frac{1}{2}$ that she/he will pay with a five-dollar note and a probability of $\frac{1}{2}$ that she/he will pay with a ten-dollar 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?

Problem 15.26 Consider the single-player versions of the game of Pig and the game of Fast Pig from Problem 2.44. Both for the hold-at-20 rule and the five-dice rule, calculate the probability mass function of the number of turns needed to reach 100 points. *Hint*: calculate first the probability mass function of the number of points gained in a single turn.

Problem 15.27 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 15.28 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 rather 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*

$$p_{ij} = 0 \qquad \text{for } i \in C \text{ and } j \notin C,$$

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*

$$\pi_j = \sum_{k\in I} \pi_k p_{kj} \qquad \text{for } j \in I \quad \text{and} \quad \sum_{j\in I} \pi_j = 1.$$

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*

$$\eta_j = \sum_{k\in I} \eta_k p_{kj} \qquad \text{for all } j \in I.$$

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*

$$\pi_j = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} \qquad \text{for all } j \in I,$$

with the averaging limit being independent of the starting state i.

(b) *The* π_j *are the unique solution to the linear equations*

$$\pi_j = \sum_{k\in I} \pi_k p_{kj} \qquad \text{for } j \in I \quad \text{and} \quad \sum_{j\in I} \pi_j = 1,$$

(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 can be based on Rule 14.6.

In case 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 case 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,

$$\pi_j = \frac{1}{\mu_{jj}},$$

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. This result can be obtained from the discrete-time version of Rule 14.6, taking as cycle the time interval between two successive visits to state j and assuming that a reward of 1 is earned each time state j is visited.

The above definition of transient state and recurrent state applies to any Markov chain. If the Markov chain has a finite state space, then it is not difficult to show that the set of recurrent states is not empty and that 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 15.29 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 15.30 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 15.31 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 15.32 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 Boris will win a game after having won the previous game?

Problem 15.33 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 15.34 Consider Problem 2.42 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 twelve states.

Problem 15.35 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 15.36 Consider Problem 15.4 again. The local entrepreneur 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 μ_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 15.37 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 15.38 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 just ended year 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 case 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.

15.5 Markov chain Monte Carlo simulation

Markov chain Monte Carlo (MCMC) methods are powerful simulation techniques to sample from a multivariate probability density that is known up to a multiplicative constant, where the constant is very difficult to compute. In Bayesian statistics one often encounters the situation that obtaining the posterior density requires the computation of a multiplicative constant in the form of a high-dimensional integral, see also Section 13.5. MCMC methods are very useful for obtaining this constant by random sampling and have greatly enhanced the applicability of Bayesian approaches in practice. MCMC methods are so-named because the sequence of simulated values forms a Markov chain that has the required probability distribution as equilibrium distribution. MCMC has revolutionized applied mathematics and is used in many areas of science including statistics, computer science, physics and biology.

This section begins with a re-examination of classical simulation procedures before moving on to the modern computationally intensive technique of Markov chain Monte Carlo simulation. The hit-or-miss method and the acceptance-rejection method can be seen as predecessors of MCMC methods and will be briefly discussed in Section 15.5.1. Ideas of these methods also appear in MCMC methods. Next we discuss in Section 15.5.2 the concept of reversible Markov chains. This concept is the key to the construction of a Markov chain that has a given probability distribution as equilibrium distribution. A very versatile MCMC method is the Metropolis–Hastings algorithm that will be considered in Section 15.5.3. This algorithm gives rise to the Gibbs sampling algorithm as a special case. The Gibbs sampling algorithm is one of the best known MCMC methods and will be discussed in Section 15.5.4.

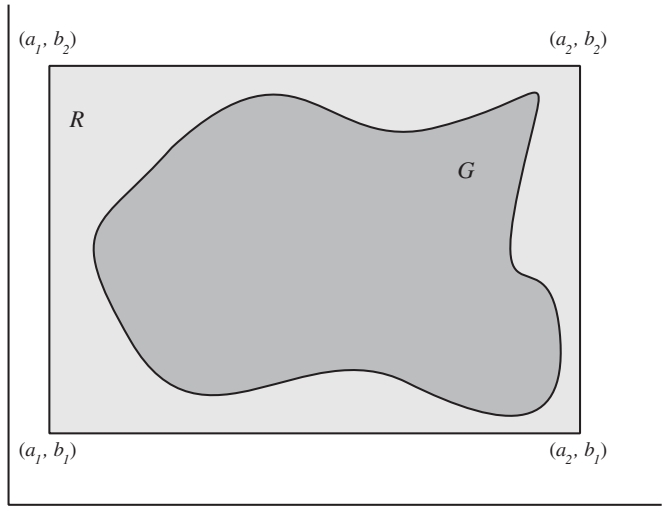


Fig. 15.2. Hit-or-miss method.

15.5.1 Classical simulation methods

The hit-or-miss method generates random points in a bounded region and was already discussed in Section 2.9. It can be used to estimate the volume of a domain in a high-dimensional space. Let G be a bounded region in which it is difficult to generate a random point directly. Suppose that an easy test is available to verify whether a given point belongs to G or not. For ease of presentation, the hit-or-miss method is described on the basis of a bounded region G in the two-dimensional plane.

Hit-or-miss method

Step 0. Choose a rectangle R that envelops the region G, see Figure 15.2.
Step 1. Generate a random point (x, y) in the rectangle R: using two random numbers u_1 and u_2 from (0, 1), let $x := a_1 + (a_2 - a_1)u_1$ and $y := b_1 + (b_2 - b_1)u_2$.
Step 2. If the point (x, y) belongs to G, accept the point as a random point in G; otherwise, repeat step 1.

The expected number of iterations of steps 1 and 2 is equal to the ratio of the area of R and the area of G. Hence it is important to choose the rectangle R as minimal as possible. The generalization of the hit-or-miss method to higher dimensional spaces is obvious.

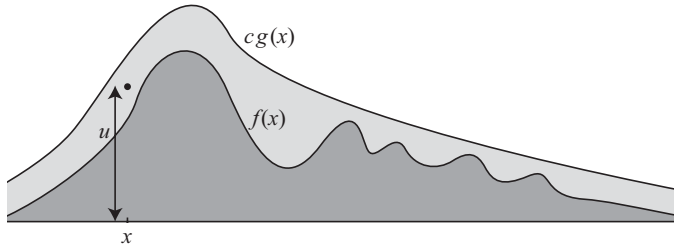


Fig. 15.3. Acceptance-rejection method.

Acceptance-rejection method

The acceptance-rejection method is an extension of the hit-or-miss method. It is a technique to simulate from a general probability density $f(x)$ that is difficult to sample from directly. Instead of sampling directly from the density $f(x)$, one uses an envelope density from which sampling is easier. Let $g(x)$ be a probability density that can be simulated by some known method and suppose there is a known constant c such that

$$f(x) \leq cg(x) \qquad \text{for all } x.$$

Note that c must be at least 1 since both $f(x)$ and $g(x)$ integrate to 1. The acceptance-rejection method goes as follows.

Step 1. Generate a candidate x from $g(x)$ and a random number u from (0, 1).
Step 2. If $u \leq \frac{f(x)}{cg(x)}$, accept x as a sample from $f(x)$; otherwise, repeat step 1.

In Rule 13.2 we have shown that the method gives a random sample from the target density $f(x)$. Intuitively, the method works because it generates randomly a point $(x, ucg(x))$ in a region covering $f(x)$ and then only keeps points in the required region under $f(x)$. The acceptance-rejection method is illustrated in Figure 15.3. The particular sample shown in the figure will be rejected. Since the expected number of iterations of steps 1 and 2 to obtain an accepted draw is equal to c, the method is only attractive when c is not too large. The acceptance-rejection method has been described for the probability density $f(x)$ of a continuous random variable, but translates literally to the case of a probability mass function $\{p(x)\}$ of a discrete random variable.

To illustrate the method, let the probability density $f(x)$ be concentrated on a finite interval (a, b) and suppose that $f(x)$ is continuous on (a, b). Choose as the candidate density the uniform density $g(x) = 1/(b - a)$ on (a, b). Letting m denote the maximum of $f(x)$ on (a, b), we have $f(x) \leq cg(x)$ for all x with $c = m(b - a)$. The acceptance-rejection method is then as follows.

Step 1. Generate random numbers u_1 and u_2 from $(0, 1)$ and let $x := a + (b - a)u_1$.
Step 2. If $u_2 \leq f(x)/m$, accept x as a sample from $f(x)$; otherwise, return to step 1.

The acceptance-rejection method is a very useful technique for simulating from one-dimensional densities. However, in high-dimensional problems the method is not practical. Apart from the difficulty of finding a high-dimensional candidate density, the bounding constant c will typically be very large.

Importance sampling

Let $f(x)$ be the probability density of a random variable X and suppose we want to estimate $E[a(X)] = \int a(x)f(x)dx$ for some function $a(x)$. Importance sampling is a useful simulation method for doing this when it is difficult to sample directly from $f(x)$ (or when the variance of the random variable $a(X)$ is very large). This method can also be used when $f(x)$ is only known up to a multiplicative constant. Choose a probability density $g(x)$ that can be simulated by some known method, where $g(x) > 0$ if $f(x) > 0$. Let the random variable Y be distributed according the density $g(x)$. Then

$$\int a(x)f(x)dx = \int \left[a(y)\frac{f(y)}{g(y)}\right] g(y)dy = E_Y\left[a(Y)\frac{f(Y)}{g(Y)}\right].$$

This forms the basis for importance sampling. Let $y_1, \ldots, y_n$ be independent samples from $g(y)$ with n large. Then, by the strong law of large numbers,

$$\int a(x)f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} w_i a(y_i) \quad \text{with } w_i = \frac{f(y_i)}{g(y_i)}.$$

An alternative formulation of importance sampling is to use

$$\int a(x)f(x)dx \approx \frac{\sum_{i=1}^{n} w_i a(y_i)}{\sum_{i=1}^{n} w_i}.$$

Why does this work? The explanation is simple. By the strong law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} w_i \approx \int [f(y)/g(y)]g(y)\,dy = \int f(y)\,dy$ and so $\frac{1}{n}\sum_{i=1}^{n} w_i \approx 1$ for n large. In the alternative formulation it suffices to know the weights w_i up to a multiplicative constant and so this formulation can be used when $f(x)$ is only known up to proportionality.

Importance sampling requires much care in the choice of the proposal density $g(x)$. If the proposal density $g(x)$ is small in a region where $|a(x)f(x)|$ is large, then the estimate of $E[a(X)]$ may be drastically wrong, even after many points

y_i have been generated. In high-dimensional problems it is even more difficult to find an appropriate proposal density.

15.5.2 Reversible Markov chains

Let $\{X_n\}$ be a Markov chain with finite state space I and having the property that any state can be reached from any other state, that is, for any two states i and j there is an $n \geq 1$ such that $p_{ij}^{(n)} > 0$. A Markov chain with this property is said to be *irreducible*. The irreducible Markov chain $\{X_n\}$ has unique equilibrium distribution $\{\pi_j\}$, see Rule 15.5 in Section 15.4. In this section we also argued that $\pi_j p_{jk}$ can be interpreted as the long-run average number of transitions per unit time from state j to state k. Many Markov chains have the property that the rate of transitions from j to k is equal to the rate of transitions from k to j for any $j, k \in I$. An example of such a Markov chain is a random walk on the number line with transitions $+1$ or -1 at each step. In this Markov chain the rate of transitions from $i-1$ to i is equal to the rate of transitions from i to $i-1$ for any i because between any two transitions from $i-1$ to i there must be one from i to $i-1$ and conversely.

Definition 15.5 *An irreducible Markov chain $\{X_n\}$ with finite state space I is said to be reversible if the equilibrium probabilities π_j satisfy the so-called detailed balance equations*

$$\pi_j p_{jk} = \pi_k p_{kj} \quad \textit{for all } j, k \in I.$$

Why the name "reversible Markov chain"? Suppose that at time 0 the Markov chain is started according to $P(X_0 = j) = \pi_j$ for $j \in I$. Then, for any $j, k \in I$,

$$P(X_{n-1} = j \mid X_n = k) = P(X_n = j \mid X_{n-1} = k) \quad \text{for all } n \geq 1.$$

In words, in equilibrium the evolution of the process backward in time is probabilistically identical to the evolution of the process forward in time. The proof is simple. Using the fact that for any k the state X_k is distributed according to $\{\pi_j\}$ (see below Definition 15.3), we have by $P(A \mid B) = P(B \mid A)\frac{P(A)}{P(B)}$ that

$$P(X_{n-1} = j \mid X_n = k) = P(X_n = k \mid X_{n-1} = j)\frac{P(X_{n-1} = j)}{P(X_n = k)} = p_{jk}\frac{\pi_j}{\pi_k}.$$

Since $p_{jk}\pi_j/\pi_k = p_{kj}$ by reversibility and $P(X_n = j \mid X_{n-1} = k) = p_{kj}$ by definition, the result follows. It is remarked that $\pi_j > 0$ for all j in a finite-state Markov chain that is irreducible.

Why the interest in the detailed balance equations? The answer is that a probability distribution satisfying these equations must be an equilibrium distribution of the Markov chain. This is easy to prove.

Rule 15.8 *Let $\{X_n\}$ be an irreducible Markov chain with finite state space I. If $\{a_j, j \in I\}$ is a probability distribution satisfying*

$$a_j p_{jk} = a_k p_{kj} \quad \text{for all } j, k \in I,$$

then $\{a_j, j \in I\}$ is the unique equilibrium distribution of the Markov chain.

To prove this result, sum both sides of the equation $a_j p_{jk} = a_k p_{kj}$ over $k \in I$. Together with $\sum_{k \in I} p_{jk} = 1$, this gives

$$a_j = \sum_{k \in I} a_k p_{kj} \quad \text{for all } j \in I.$$

By Definition 15.3, $\{a_j, j \in I\}$ is an equilibrium distribution of the Markov chain. The uniqueness of the equilibrium distribution follows from Rule 15.5. An interesting question is the following. Let $\{a_j, j \in I\}$ be a probability mass function on a finite set of states I with $a_j > 0$ for all j. Can we construct a Markov chain on I with $\{a_j, j \in I\}$ as unique equilibrium distribution? The answer is yes. The physical construction of the Markov chain proceeds as follows: if the current state of the process is j, then choose at random one of the other states as candidate state for the next state. Suppose the chosen candidate state is k. If $a_k > a_j$, then the next state of the Markov chain is state k; otherwise, the next state is either k with probability a_k/a_j or the current state j with probability $1 - a_k/a_j$. Thus, letting $N = |I|$ denote the number of states, we have constructed a Markov chain $\{X_n\}$ on I with one-step transition probabilities

$$p_{jk} = \begin{cases} \frac{1}{N-1} \min(\frac{a_k}{a_j}, 1) & \text{for } k \neq j \\ 1 - \sum_{l \neq j} p_{jl} & \text{for } k = j. \end{cases}$$

It is immediate that the Markov chain $\{X_n\}$ is irreducible (and aperiodic). Also, the Markov chain satisfies the reversibility condition: it holds for any $j, k \in I$ with $j \neq k$ that

$$a_j p_{jk} = \frac{1}{N-1} \min(a_k, a_j) = a_k \frac{1}{N-1} \min(\frac{a_j}{a_k}, 1) = a_k p_{kj}.$$

It now follows from Rule 15.8 that the Markov chain has $\{a_j\}$ as unique equilibrium distribution. A simple but useful extension of the above construction is as follows. For each $i \in I$ choose a set $N(i)$ of neighbors of i with $i \notin N(i)$, where the sets have the property that $k \in N(j)$ if $j \in N(k)$. Letting

$M = \max_i |N(i)|$, define a Markov chain by the one-step transition probabilities $p_{jk} = (1/M)\min(a_k/a_j,\ 1)$ for $k \in N(j)$, $p_{jj} = 1 - \sum_{k\in N(j)} p_{jk}$ and $p_{jk} = 0$ otherwise. Then this Markov chain satisfies the reversibility condition (verify!). Assuming sufficient connectivity between the sets $N(i)$ so that the Markov chain is irreducible, it follows that this Markov chain has also $\{a_j\}$ as unique equilibrium distribution. In the next section it will be seen that many other constructions are possible for Markov chains having $\{a_j\}$ as equilibrium distribution.

Problem 15.39 Consider an undirected graph with a finite number of nodes $i = 1, \ldots, N$, where some pairs of nodes are connected by arcs. It is assumed that the graph is connected, that is, there is a path of arcs between any two nodes. A positive weight w_{ij} is associated with each arc (i, j) in the graph $(w_{ij} = w_{ji})$. A particle is moving from node to node. If at any time the particle is in node i, it will next move to node j with probability $p_{ij} = w_{ij}/\sum_k w_{ik}$. Verify that the Markov chain describing the position of the particle is reversible and give the equilibrium probabilities. As an application, consider a mouse that is trapped in a closed maze. The maze has 15 rooms that are arranged in a three-by-five array, where doors connect neighboring rooms (room 1 is connected with the rooms 2 and 6, room 2 with the rooms 1, 3 and 7, room 7 with the rooms 2, 6, 8 and 12, and so on), see the figure.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15

The mouse moves randomly from room to room, where at each time one of the available doors is chosen with equal probability. What is the expected value of the total number of moves needed by the mouse to return to the starting room?

Problem 15.40 Let $c(i)$ be a given function on a finite but very large set I. For any $i \in I$, choose a local neighborhood $N(i)$ of i with $i \notin I$, where the sets $N(i)$ have the property 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. Prove that the unique equilibrium distribution of the Markov chain is given by $\pi_i = e^{-c(i)/T}/\sum_{k\in I} e^{-c(k)/T}$ for $i \in I$. Also, verify that $\pi_m \to 1$ as $T \to 0$ if

the function $c(i)$ assumes its absolute minimum at a unique point m. This fact can be used to find, by simulated annealing, the minimum of a function on a finite but very large set.

15.5.3 Metropolis–Hastings algorithm

The Metropolis–Hastings algorithm is an example of a Markov chain Monte Carlo method. It is widely used for high-dimensional problems. In contrast to the acceptance-rejection method which is only useful for one-dimensional problems and generates independent samples from the desired distribution, the Metropolis–Hastings algorithm generates a sequence of states from a Markov process. The algorithm is designed to draw samples from a given probability distribution that is known up to a multiplicative constant, where it is not feasible to compute directly this normalizing constant. The algorithm will be first explained for the case of a discrete probability distribution but the basic idea of the algorithm can be directly generalized to a continuous probability distribution.

Let S be a finite but very large set of states on which a probability mass function $\{\pi(s)\}$ is given, where $\pi(s) > 0$ for all $s \in S$ and the probability mass function is only known up to a multiplicative constant. The Metropolis–Hastings algorithm generates a sequence of states $(s_0, s_1, \ldots)$ from a Markov chain that has $\{\pi(s)\}$ as equilibrium distribution. To that end the algorithm uses a candidate-transition function $q(t \mid s)$ (for clarity, we use the notation $q(t \mid s)$ rather than p_{st}). This function is to be interpreted as saying that when the current state is s the candidate for the next state is t with probability $q(t \mid s)$. Thus one first chooses, for each $s \in S$, a probability mass function $\{q(t \mid s), t \in S\}$. These functions must be chosen in such a way that the Markov matrix with one-step transition probabilities $q(t \mid s)$ is irreducible. The idea is to next adjust these transition probabilities in such a way that a Markov chain results that has $\{\pi(s)\}$ as equilibrium distribution. The reversibility equations are the key to this idea. In the case that the candidate-transition function already satisfies the equations

$$\pi(s)q(t \mid s) = \pi(t)q(s \mid t) \quad \text{for all } s, t \in S,$$

we are done and can conclude from Rule 15.8 that the Markov chain with the probabilities $q(t \mid s)$ as one-step transition probabilities has $\{\pi(s)\}$ as unique equilibrium distribution. What should one do when the reversibility equations are not fully satisfied? The answer is to modify the transition probabilities by rejecting certain transitions. To work out this idea, fix two states s and t for which the reversibility equation is not satisfied. It is no restriction to assume

that

$$\pi(s)q(t \mid s) > \pi(t)q(s \mid t).$$

Otherwise, reverse the roles of s and t. If $\pi(s)q(t \mid s) > \pi(t)q(s \mid t)$, then, loosely speaking, the process moves from s to t too often. How could one restore this? A simple trick to reduce the number of transitions from s to t is to use an acceptance probability $\alpha(t \mid s)$: the process is allowed to make the transition from s to t with probability $\alpha(t \mid s)$ and otherwise the process stays in the current state s. The question remains how to choose $\alpha(t \mid s)$. Taking an acceptance probability of $\alpha(s \mid t) = 1$ for the desired transitions from t to s, the choice of $\alpha(t \mid s)$ is determined by the requirement

$$\pi(s)[q(t \mid s)\alpha(t \mid s)] = \pi(t)[q(s \mid t)\alpha(s \mid t)].$$

Together with $\alpha(s \mid t) = 1$, this gives

$$\alpha(t \mid s) = \frac{\pi(t)q(s \mid t)}{\pi(s)q(t \mid s)}.$$

Consequently, for any $s, t \in S$, we define the acceptance probability

$$\alpha(t \mid s) = \min\left[\frac{\pi(t)q(s \mid t)}{\pi(s)q(t \mid s)}, 1\right].$$

Next, the one-step transition probabilities of the Markov chain $\{X_n\}$ we are looking for are defined by

$$q_{\mathrm{MH}}(t \mid s) = \begin{cases} q(t \mid s)\alpha(t \mid s) & \text{for } t \neq s \\ 1 - \sum_{t \neq s} q(t \mid s)\alpha(t \mid s) & \text{for } t = s. \end{cases}$$

The Markov chain $\{X_n\}$ with these one-step transition probabilities satisfies the reversibility condition

$$\pi(s)q_{\mathrm{MH}}(t \mid s) = \pi(t)q_{\mathrm{MH}}(s \mid t) \quad \text{for all } s, t.$$

Using the assumption that the Markov matrix with the elements $q(t \mid s)$ is irreducible, it is easily verified that the Markov chain $\{X_n\}$ is also irreducible. It now follows from Rule 15.8 that the Markov chain $\{X_n\}$ has $\{\pi(s)\}$ as unique equilibrium distribution. It is important to note that for the construction of the Markov chain $\{X_n\}$ it suffices to know the π_s up to a multiplicative constant because the acceptance probabilities involve only the ratios $\pi(s)/\pi(t)$.

Summarizing, the Markov chain $\{X_n\}$ operates as follows. If the current state is s, a candidate state t is generated from the probability mass function $\{q(t \mid s), t \in S\}$. This state is accepted with probability $\alpha(s, t)$ as the next state

of the Markov chain. Otherwise, the next state of the Markov chain is the current state s.

We are now in a position to formulate the Metropolis–Hastings algorithm. The ingredients of the algorithm were derived for a probability mass function $\{\pi(s)\}$ on a discrete state space S. However, under suitable regularity conditions, this derivation generalizes to the case of a probability density $\{\pi(s)\}$ on a continuous state space S. In the literature one sometimes uses the terminology discrete probability density instead of probability mass function. In the formulation of the Metropolis–Hastings algorithm the terminology of density will be used to make clear that the algorithm applies both to the case of a discrete distribution and the case of a continuous distribution.

Metropolis–Hastings algorithm

Step 0. For each $s \in S$ choose a density $\{q(t \mid s), t \in S\}$. Let $X_0 := s_0$ for a starting state s_0 and let $n := 1$.
Step 1. Generate a candidate state t_n from the density $\{q(t \mid s_{n-1}), t \in S\}$. Calculate the acceptance probability

$$\alpha = \min\left[\frac{\pi(t_n)q(s_{n-1} \mid t_n)}{\pi(s_{n-1})q(t_n \mid s_{n-1})},\ 1\right].$$

Step 2. Generate a random number u from $(0, 1)$. If $u \leq \alpha$, accept t_n and let $s_n := t_n$; otherwise, $s_n := s_{n-1}$.
Step 3. $n := n + 1$. Repeat step 1 with s_{n-1} replaced by s_n.

Note that when the chosen densities $q(t \mid s)$ are symmetric, that is,

$$q(t \mid s) = q(s \mid t) \quad \text{for all } s, t,$$

the acceptance probability α in Step 2 reduces to

$$\alpha = \min\left(\frac{\pi(t_n)}{\pi(s_{n-1})},\ 1\right).$$

In many applications of the algorithm one wants to estimate $E[h(X)]$ for a given function $h(x)$, where X is a random variable with $\pi(s)$ as density. Then generate states $s_1, s_2, \ldots, s_m$ by applying the Metropolis–Hastings algorithm for a sufficiently large number m of steps. By the ergodic theorem for Markov chains (see Rule 15.7), one estimates $E[h(X)]$ by

$$\frac{1}{m}\sum_{k=1}^{m} h(s_k).$$

A practical approach to construct an approximate confidence interval for $E[h(X)]$ is dividing the single very long simulation run into 25 (say) long subruns and treating the averages of the data within the subruns as independent samples from an approximately normal distribution (this is the so-called batch-means method in simulation). How the starting state should be chosen and how many iterations should be done are empirical questions. A simple approach is to use multiple dispersed initial values to start several different chains. One suggestion for an initial value is to start the chain close to the mode of the target density if this density is unimodal. Diagnosing convergence to the target density is an art in itself. To diagnose the convergence of an average $\frac{1}{t}\sum_{k=1}^{t} h(s_k)$, one can look at a plot of this average as a function of t.

Implementation aspects

What are the best options for the proposal densities $q(t \mid s)$? There are two general approaches: independent chain sampling and random walk chain sampling. These approaches will be briefly discussed.[†]

(**a**) In independent chain sampling the candidate state t is drawn independently of the current state s of the Markov chain, that is, $q(t \mid s) = g(t)$ for some proposal density $g(x)$.

(**b**) In random walk chain sampling the candidate state t is the current state s plus a draw from a random variable Z that does not depend on the current state. In this case, $q(t \mid s) = g(t - s)$ with $g(z)$ the density of the random variable Z. If $g(z) = g(-z)$ for all z, the proposal density is symmetric and, as noted before, the acceptance probability reduces to $\alpha(t \mid s) = \min(\pi(t)/\pi(s), 1)$.

It is very important to have a well-mixing Markov chain, that is, a Markov chain that explores the state space S adequately and sufficiently fast. In other words, by the choice of the proposal densities one tries to avoid that the Markov chain stays in small regions of the state space for long periods of time. The variance of the proposal density can be thought of as a tuning parameter to get better mixing. It affects both the acceptance probability and the magnitude of the moves of the state. It is a matter of experimentation to find a tradeoff between these two features. In random walk chain sampling a rule of thumb is to choose the proposal density in such a way that on average about 50% of the candidate states are accepted. In independent chain sampling it is important that the tail of the proposal density $g(s)$ dominates the tail of the target density $\pi(s)$.

Let us illustrate the implementation aspects with the following example.

[†] For more details the interested reader is referred to S. Chib and E. Greenberg, "Understanding the Metropolis–Hastings algorithm," *The American Statistician* **49** (1995): 327–335.

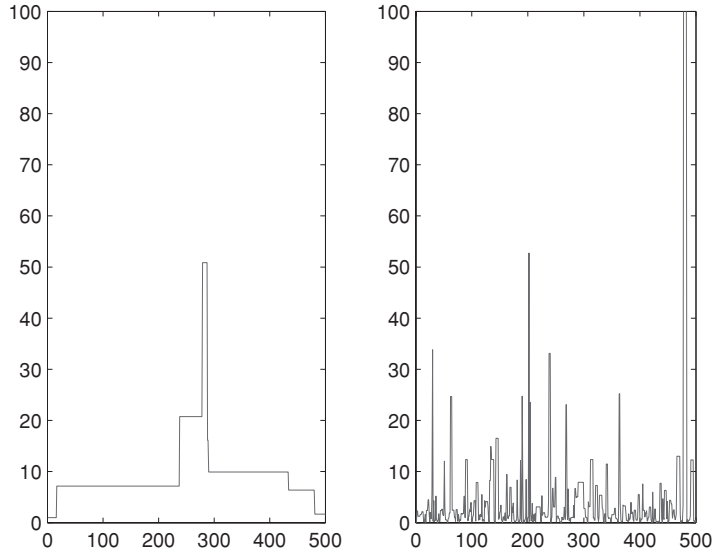


Fig. 15.4. Mixing of the state of the Markov chain.

Example 15.10 Let the density $\pi(x)$ of the positive random variable X be proportional to $x^{-2.5}e^{-2/x}$ for $x > 0$. Suppose that this target density is simulated by the Metropolis–Hastings algorithm with independent chain sampling, where both the uniform density on $(0, 1{,}000)$ and the Pareto density $1.5x^{-2.5}$ for $x > 1$ are taken as proposal density. How goes the mixing of the state of the Markov chain? What are the estimates for $E(X)$?

Solution. It is instructive to make a plot of the first 500 (say) states simulated by the Metropolis–Hastings algorithm. Such a plot is given in Figure 15.4, where the left figure corresponds to the uniform proposal density and the right figure to the Pareto proposal density. In each of the simulation studies the starting state was chosen as $s_0 = 1$. In the simulation with the Pareto density, we simulated the shifted density $\pi_{\text{shift}}(x) = \pi(x - 1)$ for $x > 1$. It is directly seen from the figure that the mixing of the state is very bad under the uniform density, though this proposal density covers nearly all the mass of the target density. A bad mixing means that the candidate state is often rejected. This happens for the uniform density because the variance of this proposal density is very large, which means that the candidate state is often far away from the current state and this in turn implies a very small acceptance probability. The Pareto density exhibits the same tail behavior as the target density $\pi(x)$ and gives an excellent mixing of the state of the Markov chain. The shortcoming of the uniform density as proposal density also appears from the fact that our

simulation study with one million runs gave the estimate of 3.75 for $E(X)$, where the exact value of $E(X)$ is 4. The simulation study with the Pareto density as proposal density gave the estimate of 3.98 for $E(X)$.

Problem 15.41 Simulate the probability distribution $\pi_1 = 0.2$ and $\pi_2 = 0.8$ with the proposal probabilities $q(t \mid s) = 0.5$ for $s, t = 1, 2$ and see how quickly the simulated Markov chain converges.

Problem 15.42 Let the joint density of the continuous random variables X_1 and X_2 be proportional to

$$e^{-\frac{1}{2}(x_1^2 x_2^2 + x_1^2 + x_2^2 - 7x_1 - 7x_2)} \quad \text{for } -\infty < x_1, x_2 < \infty.$$

What are the expected value, the standard deviation and the marginal density of the random variable X_1? Use the Metropolis–Hastings algorithm with random walk chain sampling, where the increments of the random walk are generated from (Z_1, Z_2) with Z_1 and Z_2 independent $N(0, a^2)$ random variables. Experiment with several values of a (say, $a = 0.02$, 0.2, 1, and 5) to see how the mixing in the Markov chain goes and what the average value of the acceptance probability is.

15.5.4 The Gibbs sampler

The Gibbs sampler is a special case of Metropolis–Hastings sampling and is used to simulate from a multivariate density whose univariate conditional densities are fully known. This sampler is frequently used in Bayesian statistics. To introduce the method, consider the discrete random vector $(X_1, \ldots, X_d)$ with the multivariate density function

$$\pi(x_1, \ldots, x_d) = P(X_1 = x_1, \ \ldots, \ X_d = x_d).$$

The univariate conditional densities of $(X_1, \ldots, X_d)$ are denoted by

$$\begin{aligned}&\pi_k(x \mid x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)\\ &= P(X_k = x \mid X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, X_{k+1} = x_{k+1}, \ldots, X_d = x_d).\end{aligned}$$

The key to the Gibbs sampler is the assumption that the univariate conditional densities are fully known. The Gibbs sampler generates random draws from the univariate densities and defines a valid Markov chain with $\pi(x_1, \ldots, x_d)$ as equilibrium distribution. The algorithm involves no adjustable parameters and has the feature that the acceptance probability is always equal to 1.

Gibbs sampling algorithm

Step 0. Choose a starting state $\mathbf{x} = (x_1, \ldots, x_d)$.
Step 1. Generate a random integer k from $\{1, \ldots, d\}$. Simulate a random draw y from the conditional density $\pi_k(x \mid x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)$. Define state $\mathbf{y}$ by $\mathbf{y} = (x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_d)$.
Step 2. The new state $\mathbf{x} := \mathbf{y}$. Return to step 1 with $\mathbf{x}$.

This algorithm is a special case of the Metropolis–Hastings algorithm with

$$q(\mathbf{y} \mid \mathbf{x}) = \frac{1}{d} P\left(X_k = y \mid X_j = x_j \text{ for } j = 1, \ldots, d \text{ with } j \neq k\right),$$

for any two states $\mathbf{x}$ and $\mathbf{y}$, where $\mathbf{x} = (x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_d)$ and $\mathbf{y} = (x_1, \ldots, x_{k-1}, y, x_{k+1}, \ldots, x_d)$. In the Gibbs sampler the acceptance probability $\alpha(\mathbf{y} \mid \mathbf{x})$ is always equal to 1. The proof is simple. It follows from

$$q(\mathbf{y} \mid \mathbf{x}) = \frac{1}{d} \frac{\pi(\mathbf{y})}{P(X_j = x_j,\ j \neq k)} \quad \text{and} \quad q(\mathbf{x} \mid \mathbf{y}) = \frac{1}{d} \frac{\pi(\mathbf{x})}{P(X_j = x_j,\ j \neq k)}$$

that $q(\mathbf{x} \mid \mathbf{y})/q(\mathbf{y} \mid \mathbf{x}) = \pi(\mathbf{x})/\pi(\mathbf{y})$ and so

$$\alpha(\mathbf{y} \mid \mathbf{x}) = \min\left[\frac{\pi(\mathbf{y})\pi(\mathbf{x})}{\pi(\mathbf{x})\pi(\mathbf{y})},\ 1\right] = 1.$$

Standard Gibbs sampler

In practice one usually uses the standard Gibbs sampler, where in each iteration all components of the state vector are adjusted rather than a single component. Letting $\mathbf{x}^{(n)} = (x_1^{(n)}, \ldots, x_d^{(n)})$ denote the state vector obtained in iteration n, the next iteration $n + 1$ proceeds as follows:

$$\begin{aligned}
&x_1^{(n+1)} \text{ is a random draw from } \pi_1(x \mid x_2^{(n)}, x_3^{(n)}, \ldots, x_d^{(n)}) \\
&x_2^{(n+1)} \text{ is a random draw from } \pi_2(x \mid x_1^{(n+1)}, x_3^{(n)}, \ldots, x_d^{(n)}) \\
&\qquad\qquad \vdots \\
&x_d^{(n+1)} \text{ is a random draw from } \pi_d(x \mid x_1^{(n+1)}, x_2^{(n+1)}, \ldots, x_{d-1}^{(n+1)}).
\end{aligned}$$

This gives the new state vector $\mathbf{x}^{(n+1)} = (x_1^{(n+1)}, \ldots, x_d^{(n+1)})$. The standard Gibbs sampler also generates a sequence of states $\{\mathbf{x}^{(k)}, k = 1, \ldots, m\}$ from a Markov chain having $\pi(x_1, \ldots, x_d)$ as equilibrium distribution. The technical proof is omitted

The expectation of any function h of the random vector $(X_1, \ldots, X_d)$ can be estimated by $(1/m) \sum_{k=1}^{m} h(\mathbf{x}^{(k)})$ for a Gibbs sequence $\{\mathbf{x}^{(k)}, k = 1, \ldots, m\}$ of

sufficient length m. In particular one might use the Gibbs sequence to estimate the marginal density of any component of the random vector, say the marginal density $\pi_1(x)$ of the random variable X_1. A naive estimate for $\pi_1(x)$ is based on the values $x_1^{(1)}, \ldots, x_1^{(m)}$ from the Gibbs sequence, but a better approach is to use the other values from the Gibbs sequence together with the explicit expression for the univariate conditional density of X_1. Formally, by the law of conditional expectation, $\pi_1(x) = E[\pi_1(x \mid X_2, \ldots, X_d)]$. Thus, a better estimate for $\pi_1(x)$ is given by

$$\hat{\pi_1}(x) = \frac{1}{m} \sum_{k=1}^{m} \pi_1(x \mid x_2^{(k)}, \ldots, x_d^{(k)}).$$

This estimate uses more information than the estimate based only on the individual values $x_1^{(1)}, \ldots, x_1^{(m)}$ and will typically be more accurate.

Example 15.11 The bivariate density $\pi(x, y)$ of the random vector (X, Y) is proportional to

$$\binom{r}{x} y^{x+\alpha-1}(1-y)^{r-x+\beta-1} \quad \text{for } x = 0,\ 1,\ \ldots, r, \quad 0 \le y \le 1,$$

where r, α and β are given positive integers. Assuming the data $r = 16, \alpha = 2$ and $\beta = 4$, we are interested in estimating the expected value, the variance and the marginal density of X. Since the univariate conditional densities of X and Y can be explicitly determined, we can use the Gibbs sampler. This method also works for the case of a random vector (X, Y) with a discrete component X and a continuous component Y. How do we find the univariate conditional densities? Applying basic results from Section 13.1, we have that $\pi_1(x \mid y)$ is the ratio of the joint density and the marginal density of Y. The marginal density of Y is given by $\sum_{u=0}^{r} \pi(u,\ y)$. Thus, for any fixed y, it follows from $\pi_1(x \mid y) = \pi(x, y) / \sum_{u=0}^{r} \pi(u, y)$ that

$$\pi_1(x \mid y) = \frac{\binom{r}{x} y^x (1-y)^{r-x}}{\sum_{u=0}^{r} \binom{r}{u} y^u (1-y)^{r-u}} = \binom{r}{x} y^x (1-y)^{r-x}$$

for $x = 0, 1, \ldots, r$. Hence $\pi_1(x \mid y)$ is the binomial(r, y) density for any fixed y. In the same way, we obtain from $\pi_2(y \mid x) = \pi(x,\ y) / \int_0^1 \pi(x, u)\, du$ that for any fixed x the conditional density $\pi_2(y \mid x)$ is proportional to

$$y^{x+\alpha-1}(1-y)^{r-x+\beta-1}$$

for $0 \le y \le 1$. Hence $\pi_2(y \mid x)$ is the beta$(x + \alpha,\ r - x + \beta)$ density.

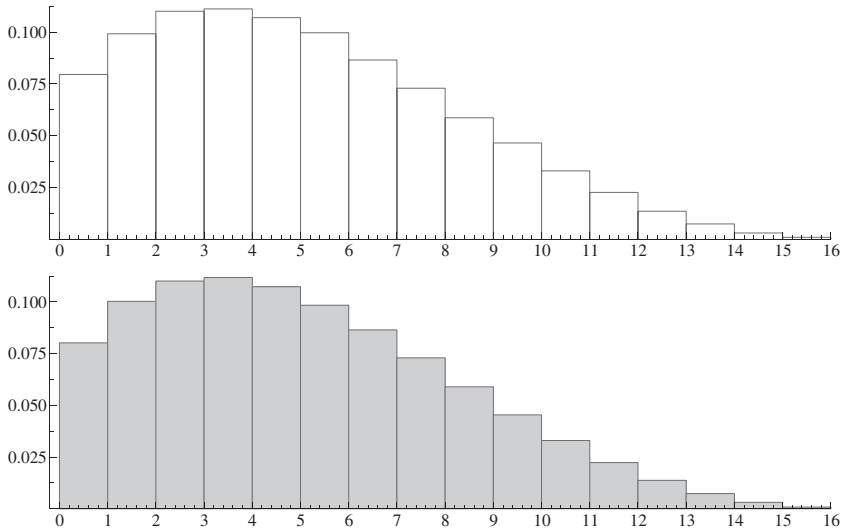


Fig. 15.5. Simulated and exact histogram for $\pi_1(x)$.

A Gibbs sequence $(x_0, y_0), (x_1, y_1), \ldots, (x_m, y_m)$ of length m is generated as follows by using the standard Gibbs sampler. Choose an integer x_0 between 0 and r. The other elements of the sequence are iteratively obtained by generating alternately a random draw y_j from the beta density $\pi_2(y \mid x_j)$ and a random draw x_{j+1} from the binomial density $\pi_1(x \mid y_j)$. Codes to simulate from the binomial density and the beta density are widely available. In our simulation we have generated $m = 250{,}000$ observations for the state vector (x, y). The first histogram in Figure 15.5 gives the simulated histogram for the marginal density $\pi_1(x)$ of the random variable X. By comparison, the second histogram gives the exact values of $\pi_1(x)$ (a direct computation of the proportionality constant for $\pi_1(x)$ is possible). The simulated histogram is based on the estimate

$$\hat{\pi_1}(x) = \frac{1}{m} \sum_{k=1}^{m} \binom{r}{x} y_k^x (1 - y_k)^{r-x}$$

for $\pi_1(x)$, using the explicit expression for $\pi_1(x \mid y)$. On the basis of $\hat{\pi_1}(x)$, the estimates 5.35 and 11.20 are found for $E(X)$ and $\text{var}(X)$, where the exact values of $E(X)$ and $\text{var}(X)$ are 5.33 and 11.17.

Problem 15.43 Consider Problem 15.42 again. Use the Gibbs sampler to simulate the marginal density of X_1 together with the expected value and the standard deviation.

Problem 15.44 In an actuarial model the random vector (X, Y, N) has a trivariate density $\pi(x, y, n)$ that is proportional to

$$\binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1} e^{-\lambda} \frac{\lambda^n}{n!}$$

for $x = 0, 1, \ldots, n$, $0 < y < 1$ and $n = 0, 1, \ldots$. The random variable N represents the number of policies in a portfolio, the random variable Y represents the claim probability for any policy and the random variable X represents the number of policies resulting in a claim. First verify that the univariate conditional density functions of X, Y and N are given by the binomial(n, y) density, the beta$(x + \alpha, n - x + \beta)$ density and the Poisson$(\lambda(1 - y))$ density shifted to the point x. Assuming the data $\alpha = 2$, $\beta = 8$ and $\lambda = 50$, use the Gibbs sampler to estimate the expected value, the standard deviation and the marginal density of the random variable X.

16

Continuous-time Markov chains

Many random phenomena happen in continuous time. Examples include occurrence of cell phone calls, spread of epidemic diseases, stock fluctuations, etc. A continuous-time Markov chain is a very useful stochastic process to model such phenomena. It is a process that goes from state to state according to a Markov chain, but the times between state transitions are continuous random variables having an exponential distribution.

The purpose of this chapter is to give an elementary introduction to continuous-time Markov chains. The basic concept of the continuous-time Markov chain model is the so-called transition rate function. Several examples will be given to illustrate this basic concept. Next we discuss the time-dependent behavior of the process and give Kolmogorov's differential equations to compute the time-dependent state probabilities. Finally, we present the flow-rate-equation method to compute the limiting state probabilities and illustrate this powerful method with several examples dealing with queueing systems.

16.1 Markov chain model

A continuous-time stochastic process $\{X(t),\ t \geq 0\}$ is a collection of random variables indexed by a continuous time parameter $t \in [0, \infty)$, where the random variable $X(t)$ is called the state of the process at time t. In an inventory problem $X(t)$ might be the stock on hand at time t and in a queueing problem $X(t)$ might be the number of customers present at time t. The formal definition of a continuous-time Markov chain is a natural extension of the definition of a discrete-time Markov chain.

Definition 16.1 *The stochastic process* $\{X(t),\ t \geq 0\}$ *with discrete state space* I *is said to be a continuous-time Markov chain if it possesses the Markovian*

property, that is, for all time points $s, t \geq 0$ *and states* $i,\ j, x(u)$ *with* $0 \leq u < s$,

$$P(X(t+s) = j \mid X(u) = x(u) \text{ for } 0 \leq u < s,\ X(s) = i) = P(X(t+s) = j \mid X(s) = i).$$

In words, the Markovian property says that if you know the present state at time s, then all additional information about the states at times prior to time s is irrelevant for the probabilistic development of the process in the future. All that matters for future states is what the present state is. A continuous-time Markov chain is said to be *time-homogeneous* if for any $s, t > 0$ and any states $i, j \in I$,

$$P(X(t+s) = j \mid X(s) = i) = P(X(t) = j \mid X(0) = i).$$

The transition functions $p_{ij}(t)$ are defined by

$$p_{ij}(t) = P(X(t) = j \mid X(0) = i) \quad \text{for } t \geq 0 \text{ and } i, j \in I.$$

In addition to the assumption of time-homogeneity, we now make the assumption that the state space I is *finite*. This assumption is made to avoid technical complications involved with a countably infinite state space. However, under some regularity conditions the results for the case of a finite state space carry over to the case of a countably infinite state space.

Transition rates

In continuous time there are no smallest time steps and hence we cannot speak about one-step transition probabilities as in discrete time. In a continuous-time Markov chain we would like to know, for very small time steps of length h, what the probability $p_{ij}(h)$ is of being in a *different* state j at time $t + h$ if the present state at time t is i. This probability is determined by the so-called transition rates.[†] These transition rates are to continuous-time Markov chains what the one-step transition probabilities are to discrete-time Markov chains. Formally, the transition rate q_{ij} can be introduced as the derivative of $p_{ij}(t)$ at $t = 0$. For the case of a finite state space, the $p_{ij}(t)$ are differentiable for any $t \geq 0$. The reader is asked to take for granted this deep result. In particular, the right-hand derivative of $p_{ij}(t)$ at $t = 0$ exists for all $i,\ j$. This limit is denoted

[†] Rate is a measure of how quickly something happens. It can be seen as the frequency at which a repeatable event happens per unit time. That is, an arrival rate λ means that the average frequency of arrivals per unit time is λ.

by q_{ij} for $j \neq i$ and by $-v_i$ for $j = i$. Using the fact that

$$p_{ij}(0) = \begin{cases} 0 & \text{for } j \neq i \\ 1 & \text{for } j = i, \end{cases}$$

we have $(p_{ij}(h) - p_{ij}(0))/h = p_{ij}(h)/h$ for $j \neq i$ and $(p_{ii}(h) - p_{ii}(0))/h = (p_{ii}(h) - 1)/h$ for $j = i$. The rates q_{ij} and the rates v_i are now defined by

$$q_{ij} = \lim_{h \to 0} \frac{p_{ij}(h)}{h} \quad \text{for } j \neq i \quad \text{and} \quad v_i = \lim_{h \to 0} \frac{1 - p_{ii}(h)}{h}.$$

Using the mathematical symbol $o(h)$ we can write this in a more convenient form. The symbol $o(h)$ is the generic notation for any function $f(h)$ with the property that

$$\lim_{h \to 0} \frac{f(h)}{h} = 0,$$

that is, $o(h)$ represents some unspecified function that is negligibly small compared to h itself as $h \to 0$. For example, any function $f(h) = h^a$ with $a > 1$ is an $o(h)$-function. A useful fact is that both the sum and the product of a finite number of $o(h)$-functions are again $o(h)$-functions. Summarizing,

Rule 16.1 *For any $t \geq 0$ and small h,*

$$P(X(t+h) = j \mid X(t) = i) = \begin{cases} q_{ij}h + o(h) & \text{for } j \neq i \\ 1 - v_i h + o(h) & \text{for } j = i. \end{cases}$$

The transition rate q_{ij} gives the rate at which the process tries to enter a different state j when the process is in state i. The exit rate v_i gives the rate at which the process tries to leave state i for another state when the process is in state i. Since $p_{ii}(h) + \sum_{j \neq i} p_{ij}(h) = 1$, it follows that

$$v_i = \sum_{j \neq i} q_{ij} \quad \text{for all } i \in I.$$

It is important to note that the q_{ij} are rates, not probabilities and, as such, while they must be nonnegative, they are not bounded by 1. However, for very small h, you can interpret $q_{ij}h$ as a probability, namely as the probability that in the next time interval h the process will jump to a different state j when the present state is i. Also, it is important to note from Rule 16.1 and the finiteness of the state space that the probability of two or more state changes in such a small time interval is $o(h)$. The transition rates q_{ij} are obtained from the transition functions $p_{ij}(t)$. Conversely, it can be proved that transition rates q_{ij} uniquely determine transition functions $p_{ij}(t)$ when the state space is finite.

How should you find the q_{ij} in practical applications? You will see that the exponential distribution is the building block for the transition rates. Before proceeding with the Markov chain model, we first discuss the most important properties of this distribution and the closely related Poisson process.

Intermezzo: the exponential distribution and the Poisson process

Recall from Chapter 10 that a continuous random variable T is said to have an exponential distribution with rate $\lambda > 0$ if its probability density function $f(t)$ is given by

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{for } t > 0 \\ 0 & \text{otherwise .} \end{cases}$$

The cumulative probability distribution function of T is $P(T \leq t) = 1 - e^{-\lambda t}$ for $t \geq 0$, and T has $1/\lambda$ as expected value. The key property of the exponential distribution is its memoryless property. That is, for any $s, t \geq 0$,

$$P(T > t + s \mid T > s) = P(T > t).$$

Imagining that T represents the lifetime of an item, this property says that the remaining lifetime of the item has the same exponential distribution as the original lifetime, regardless of how long the item has been already in use. This memoryless property has been proved in Section 10.4.3. The exponential distribution is the only continuous distribution possessing this property. For building and understanding continuous-time Markov chains, it is more convenient to express the memoryless property of the exponential distribution as

$$P(T \leq t + h \mid T > t) = \lambda h + o(h) \quad \text{as } h \to 0,$$

no matter what the value of t is. In other words, the exponential distribution has a constant failure rate. The proof is as follows. By the memoryless property, $P(T \leq t + h \mid T > t) = e^{-\lambda h}$. Expanding out $e^{-\lambda h}$ in a Taylor series, we find

$$\begin{aligned} P(T \leq t + h \mid T > t) &= 1 - \left(1 - \frac{\lambda h}{1!} + \frac{(\lambda h)^2}{2!} - \frac{(\lambda h)^3}{3!} + \cdots\right) \\ &= \lambda h + o(h) \quad \text{as } h \to 0. \end{aligned}$$

The Poisson process often appears in applications of continuous-time Markov chains. This process was already discussed in Section 4.2.3, see also Section 10.4.3. It is a continuous-time counting process $\{N(t),\ t \geq 0\}$, where $N(t)$ is the number of events (e.g. arrivals of customers or jobs) that have

occurred up to time t. Several equivalent definitions of the Poisson process can be given.

First definition A stochastic process $\{N(t),\ t \geq 0\}$ with $N(0) = 0$ is said to be a Poisson process with rate λ if

(a) the random variable $N(t)$ counts the number of events that have occurred up to time t;
(b) the times between events are independent random variables having a common exponential distribution with expected value $1/\lambda$.

From this definition the following memoryless property can be shown for the Poisson process: at each point of time the waiting time until the next occurrence of an event has the same exponential distribution as the original inter-occurrence times, regardless of how long ago the last event has occurred. For our purposes in continuous-time Markov chains, the following equivalent definition is more appropriate.

Second definition A stochastic process $\{N(t),\ t \geq 0\}$ with $N(0) = 0$ is said to be a Poisson process with rate λ if

(a) occurrences of events in any time interval $(t, t + h)$ are independent of what happened up to time t;
(b) for any $t \geq 0$, the probability $P(N(t + h) - N(t) = 1)$ of one occurrence of an event in the time interval $(t, t + h)$ is $\lambda h + o(h)$, the probability $P(N(t + h) - N(t) = 0)$ of no occurrence of an event is $1 - \lambda h + o(h)$, and the probability $P(N(t + h) - N(t) \geq 2)$ of more than one is $o(h)$ as $h \to 0$.

The proof of the equivalence of the two definitions will not be given. As pointed out in Chapter 4, the reason for the name of the Poisson process is the fact that the number of events occurring in any given time interval of length t has a Poisson distribution with expected value λt.

Alternative construction of a continuous-time Markov chain

In this paragraph we give a revealing way to think about a continuous-time Markov chain. The process can be characterized in another way that leads to an understanding of how to simulate the process. When the continuous-time Markov chain $\{X(t)\}$ enters state i at some time, say time 0, the process stays there a random amount of time before making a transition into a different state. Denote this amount of time by T_i. What does Definition 16.1 suggest about

this random variable T_i? Could we directly find $P(T_i > t)$? It is not obvious. But, instead suppose we ask about $P(T >_i t + s \mid T_i > s)$. Then, by Definition 16.1 and the assumption of time-homogeneity,

$$\begin{aligned} &P(T_i > t + s \mid T_i > s) \\ &= P(X(u) = i \text{ for } 0 \le u \le t + s \mid X(u) = i \text{ for } 0 \le u \le s) \\ &= P(X(u) = i \text{ for } s \le u \le t + s \mid X(s) = i) \\ &= P(X(u) = i \text{ for } 0 \le u \le t \mid X(0) = i). \end{aligned}$$

Hence, for any $s,\ t \ge 0$,

$$P(T_i > t + s \mid T_i > s) = P(T_i > t).$$

In other words, the random variable T_i has the memoryless property and thus has an exponential distribution. It is easily seen from Rule 16.1 that this exponential distribution must have rate ν_i. When the process leaves state i after the exponential time T_i, it will enter a different state j with some probability, call it p_{ij}. This probability is given by

$$p_{ij} = \frac{q_{ij}}{\nu_i} \quad \text{for } j \ne i.^{\dagger}$$

Summarizing, a continuous-time Markov chain specified by the transition rates q_{ij} can be simulated as follows.

(a) When the process enters state i, draw from the exponential distribution with rate $\nu_i = \sum_{j \ne i} q_{ij}$ to determine how long the process stays in i before moving to a different state.
(b) When the process leaves state i, draw the next state j from the probability mass function $p_{ij} = q_{ij}/\nu_i$ with $j \ne i$.

To conclude this section, we give several illustrative examples of continuous-time Markov chains.

Example 16.1 A source transmitting messages is alternately on and off. The off-times are independent random variables having a common exponential distribution with rate α and the on-times are independent random variables having a common exponential distribution with rate β. Also the off-times and on-times are independent of each other. What is an appropriate continuous-time Markov chain model?

† The probability of going to state j in the next time interval h when the process is in state i can be represented both by $\nu_i h \times p_{ij} + o(h)$ and by $q_{ij}h + o(h)$ as $h \to 0$, and so $\nu_i p_{ij} = q_{ij}$.

Solution. For any $t \geq 0$, let the random variable $X(t)$ be equal to 1 if the source is on at time t and be 0 otherwise. Because of the memoryless property of the exponentially distributed on-times and off-times, the process $\{X(t)\}$ satisfies the Markovian property and thus is a continuous-time Markov chain with state space $I = \{0, 1\}$. The off-time has a constant failure rate α and so $P(\text{an off-time will expire in } (t, t+h) \mid \text{the source is off at time } t) = \alpha h + o(h)$ for h small. It now follows that

$$P(X(t+h) = 1 \mid X(t) = 0) = \alpha h + o(h) \quad \text{for } h \text{ small}$$

and so the transition rate $q_{01} = \alpha$. In the same way, $q_{10} = \beta$. This completes the specification of the continuous-time Markov chain model.

Example 16.2 An inflammable product is stored in a special tank at a filling station. Customers asking for the product arrive according to a Poisson process with rate λ. Each customer asks for one unit. Any demand that occurs when the tank is out of stock is lost. Opportunities to replenish the stock in the tank occur according to a Poisson process having rate μ and being independent of the demand process. For reasons of security it is only allowed to replenish the stock when the tank is empty. The replenishment quantity is Q units of stock. Define an appropriate continuous-time Markov chain model for the stock on hand.

Solution. For any $t \geq 0$, let the random variable $X(t)$ denote how many units of stock are in the tank at time t. The process $\{X(t)\}$ is a continuous-time Markov chain with state space $I = \{0, 1, \ldots, Q\}$. The Markovian property is satisfied as a consequence of the fact that the Poisson process is memoryless, that is, at any time it is irrelevant for the occurrence of future events how long ago the last event occurred. What about the transition rates? If the process is in state i with $1 \leq i \leq Q$, a transition from i to $i-1$ occurs when a customer arrives. The probability of one customer arriving in $(t, t+h)$ is $\lambda h + o(h)$ for h small, while the probability of two or more customers arriving in $(t, t+h)$ is $o(h)$. Hence $P(X(t+h) = i-1 \mid X(t) = i) = \lambda h + o(h)$ for h small and so

$$q_{i,i-1} = \lambda \quad \text{for } 1 \leq i \leq Q.$$

Also, $q_{ij} = 0$ for $j \neq i-1$. If the process is in state 0, a transition from 0 to Q occurs when a replenishment opportunity arises. Replenishment opportunities occur at a rate μ and so $q_{0Q} = \mu$. The other q_{0j} are zero.

Example 16.3 Consider a population of m individuals, some of whom are infected and the others are susceptible. Initially one individual is infected. The times between contacts between any two individuals of the population have an

exponential distribution with rate μ. The times between the various contacts are independent of each other. If a susceptible individual comes in contact with an infected individual, the susceptible individual becomes infected. Once infected, an individual stays infected. What stochastic process describes the spread of the disease?

Solution. The key to the solution is the fact that the minimum of r independent random variables having exponential distributions with rates $\mu_1, \ldots, \mu_r$ is again exponentially distributed with rate $\mu_1 + \cdots + \mu_r$ (see Example 11.7). Denote by the random variable $X(t)$ the number of infected individuals at time t. Then, by the memoryless property of the exponential distribution, the process $\{X(t)\}$ is a continuous-time Markov chain with state space $I = \{1, 2, \ldots, m\}$. How do we find the transition rates? If there are i infected individuals, the time until one of the $m - i$ susceptible individuals becomes infected is the smallest of $(m - i) \times i$ independent exponentials each having rate μ and hence this time has an exponential distribution with rate $(m - i)i\mu$. It now follows that the transition rates $q_{i,i+1}$ are given by

$$q_{i,i+1} = (m - i)i\mu \quad \text{for } i = 1, 2, \ldots, m - 1.$$

The other transition rates q_{ij} are zero because of the assumption that an infected individual stays infected once it is infected. Note that state m is an absorbing state with leaving rate $v_m = 0$.

Example 16.4 An electronic system uses one operating unit but has built-in redundancy in the form of $s - 1$ standby units. The s units are identical. The standby units are not switched on (cold standby). The lifetime of the operating unit has an exponential distribution with rate λ. If the operating unit fails, it is immediately replaced by a standby unit if available. There are ample repair facilities so that each failed unit enters repair immediately. The repair time of a failed unit has an exponential distribution with rate μ. A repaired unit is as good as new. What process describes the number of units in repair?

Solution. Let the random variable $X(t)$ denote the number of units in repair at time t. The Markovian property is satisfied for the process $\{X(t)\}$ and so this process is a continuous-time Markov chain with state space $I = \{0, 1, \ldots, s\}$. For the determination of the transition rates, we will use the fact that the smallest of i independent random variables having an exponential distribution with rate μ is exponentially distributed with rate $i\mu$. If at time t there are i units in repair, then the probability of having exactly one repair completed in the time interval $(t, t + h)$ is $i\mu h + o(h)$ for h small. The probability of having of having two or more repair completions in this interval is $o(h)$. The same is true for the

probability that in this interval one or more repairs are completed and at the same time one or more units will fail. In general, it is true for a continuous-time Markov chain that the probability of two or more transitions in a small time interval $(t, t+h)$ is $o(h)$. Hence, for $1 \leq i \leq s$,

$$P(X(t+h) = i-1 \mid X(t) = i) = i\mu h + o(h) \quad \text{for } h \text{ small}$$

and so $q_{i,i-1} = i\mu$. Using the fact that a unit in operation at time t will fail in $(t, t+h)$ is $\lambda h + o(h)$ for h small, we obtain by similar arguments that

$$P(X(t+h) = i+1 \mid X(t) = i) = \lambda h + o(h) \quad \text{for } h \text{ small.}$$

This gives $q_{i,i+1} = \lambda$ for $0 \leq i \leq s-1$. The other transition rates q_{ij} are zero.

Problem 16.1 Consider a two-unit reliability system with one operating unit and one unit in warm standby. The operating unit has a constant failure rate of λ, while the unit in warm standby has a constant failure rate of η with $0 \leq \eta < \lambda$. Upon failure of the operating unit, the unit in warm standby is put into operation if available. The time it takes to replace a failed unit is exponentially distributed with rate μ. It is only possible to replace one failed unit at a time. Define an appropriate continuous-time Markov chain and specify its transition rates.

Problem 16.2 Cars arrive at a gasoline station according to a Poisson process with an average of λ cars per minute. A car enters the station only if fewer than four other cars are present. The gasoline station has only one pump. The amount of time required to serve a car has an exponential distribution with an expected value of $1/\mu$ minutes. Define an appropriate continuous-time Markov chain and specify its transition rates.

Problem 16.3 Consider the following modification of Example 16.3. In addition to infected and susceptible individuals, there are recovered individuals. Assume that an infected individual recovers after an exponentially distributed time with rate ρ. A recovered individual again becomes susceptible after an exponentially distributed time with rate η. Define an appropriate continuous-time Markov chain and specify its transition rates. *Hint*: use two-dimensional random vectors for the continuous-time Markov chain.

Problem 16.4 In a single-product inventory system the depletion of stock is due to demand and deterioration. Customers asking for one unit of the product arrive according to a Poisson process with rate λ. The lifetime of each unit product has an exponential distribution with rate μ. Each time the stock drops to zero, a replenishment order for Q units is placed. The lead time of each order is negligible. What process describes the stock on hand?

Problem 16.5 The so-called sheroot is a familiar sight in Middle East street scenes. It is a seven-seat taxi that drives from a fixed stand in a town to another town. A sheroot leaves the stand as soon as all seven seats are occupied by passengers. Consider a sheroot stand which has room for only one sheroot. Potential passengers arrive at the stand according to a Poisson process with rate λ. A customer who finds upon arrival no sheroot present and seven other passengers already waiting goes elsewhere for transport; otherwise, the customer waits until a sheroot departs. After departure of a sheroot, the time until a new sheroot becomes available has an exponential distribution with rate μ. Define an appropriate continuous-time Markov chain and specify its transition rates.

16.2 Time-dependent probabilities

How do we compute the time-dependent probabilities $p_{ij}(t) = P(X(t) = j \mid X(0) = i)$ for $t > 0$ and $i, j \in I$? The transition functions $p_{ij}(t)$ are the counterpart of the n-step transition probabilities in the discrete-time Markov chain. It will be no surprise that there is also a counterpart of the Chapman–Kolmogorov equations in Rule 15.1.

Rule 16.2 *For any $s, t \geq 0$ and $i, j \in I$,*

$$p_{ij}(t+s) = \sum_{k \in I} p_{ik}(t) p_{kj}(s).$$

The proof is essentially identical to the proof of Rule 15.1. By conditioning on $X(t)$ and using the Markovian property together with the assumption of time-homogeneity,

$$\begin{aligned} &P(X(t+s) = j \mid X(0) = i) \\ &= \sum_{k \in I} P(X(t+s) = j \mid X(0) = i, X(t) = k) P(X(t) = k \mid X(0) = i) \\ &= \sum_{k \in I} P(X(s) = j \mid X(0) = k) P(X(t) = k \mid X(0) = i), \end{aligned}$$

showing the desired result.

Unlike the discrete-time result in Rule 15.1, the continuous analog in Rule 16.2 is not directly amenable for numerical computations. How could we compute the $p_{ij}(t)$? To answer this question, let us differentiate both sides of $p_{ij}(t+s) = \sum_{k \in I} p_{ik}(t) p_{kj}(s)$ with respect to s and next set $s = 0$. Since

the state space I is finite, the sum can be differentiated term by term. This gives

$$p'_{ij}(t) = \sum_{k \in I} p_{ik}(t) p'_{kj}(0).$$

Next we use the result $p'_{kj}(0) = q_{kj}$ for $k \neq j$ and $p'_{kj}(0) = -\nu_j$ for $k = j$. Then we obtain the following so-called Kolmogorov forward equations.

Rule 16.3 *Assume a finite state space I. Then, for any given state i, the transition functions $p_{ij}(t)$, $j \in I$ satisfy the linear differential equations*

$$p'_{ij}(t) = -\nu_j p_{ij}(t) + \sum_{k \neq j} q_{kj} p_{ik}(t) \quad \textit{for } t > 0 \textit{ and } j \in I.$$

Another set of differential equations, known as the Kolmogorov backward equations, can also be obtained from Rule 16.2. Differentiating both sides of $p_{ij}(t+s) = \sum_{k \in I} p_{ik}(t) p_{kj}(s)$ with respect to t and next setting $t = 0$, we obtain for any $j \in I$ that (verify!)

$$p'_{ij}(s) = -\nu_i p_{ij}(s) + \sum_{k \neq i} q_{ik} p_{kj}(s) \quad \text{for } s > 0 \text{ and } i \in I.$$

Linear differential equations can be numerically solved by standard codes that are available in numerical software packages. It is only possible in special cases to give an explicit solution.

Example 16.1 (continued) For starting state $i = 1$, the Kolmogorov forward equations are given by

$$\begin{aligned} p'_{10}(t) &= -\alpha p_{10}(t) + \beta p_{11}(t) \quad \text{for } t > 0 \\ p'_{11}(t) &= -\beta p_{11}(t) + \alpha p_{10}(t) \quad \text{for } t > 0. \end{aligned}$$

Since $p_{11}(t) = 1 - p_{10}(t)$, the first equation can be written as

$$p'_{10}(t) = \beta - (\alpha + \beta) p_{10}(t) \quad \text{for } t > 0.$$

By the theory of linear differential equations, a solution of the form $p_{10}(t) = c_1 + c_2 e^{-(\alpha+\beta)t}$ is expected. Differentiation and substitution leads to

$$-c_2(\alpha+\beta)e^{-(\alpha+\beta)t} = \beta - c_1(\alpha+\beta) - c_2(\alpha+\beta)e^{-(\alpha+\beta)t}.$$

This gives $\beta - c_1(\alpha+\beta) = 0$ and so $c_1 = \beta/(\alpha+\beta)$. The constant c_2 is found by using the boundary condition $p_{10}(0) = 0$. This gives $c_1 + c_2 = 0$ and so $c_2 = -c_1$. We can now conclude that

$$p_{10}(t) = \frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta} e^{-(\alpha+\beta)t} \quad \text{and} \quad p_{11}(t) = \frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta} e^{-(\alpha+\beta)t}$$

for all $t \geq 0$. In the same way, for all $t \geq 0$,

$$p_{01}(t) = \frac{\alpha}{\alpha+\beta} - \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)t} \quad \text{and} \quad p_{00}(t) = \frac{\beta}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}e^{-(\alpha+\beta)t}.$$

In this example the $p_{ij}(t)$ have a limit as $t \to \infty$. Moreover, the limit is independent of the starting state i and is given by $\beta/(\alpha+\beta)$ when $j = 0$ and by $\alpha/(\alpha+\beta)$ when $j = 1$. What can we generally say about $p_{ij}(t)$ for t large? This question will be addressed in the next section.

First-passage time probabilities

First-passage time probabilities are important in reliability problems among others. For example, in Example 16.4 one might be interested in the probability distribution of the time until the system goes down for the first time when the down state corresponds to the situation that all units are in repair. For a given state r, let us define

$$Q_i(t) = P(X(u) \neq r \text{ for all } 0 \leq u \leq t \mid X(0) = i) \quad \text{for } t \geq 0 \text{ and } i \neq r,$$

where $Q_i(0) = 1$ for all $i \neq r$. In words, $Q_i(t)$ is the probability that the time until the first visit of the process to state r is more than t given that the starting state is i. How could we compute the first-passage time probabilities? The idea is simple. Modify the process $\{X(t)\}$ by making state r absorbing. That is, define a continuous-time Markov chain $\{\overline{X}(t)\}$ by the transition rates $\overline{q}_{ij}$ with $\overline{q}_{ij} = q_{ij}$ for $i \neq r$ and $\overline{q}_{rj}$=0 for all j. Let $\overline{p}_{ij}(t) = P(\overline{X}(t) = j \mid \overline{X}(0) = i)$. Using the fact that r is an absorbing state for the process $\{\overline{X}(t)\}$, we have (why?)

$$Q_i(t) = 1 - \overline{p}_{ir}(t) \quad \text{for } t > 0 \text{ and } i \neq r.$$

Applying the Kolmogorov backward equations for the $\overline{p}_{ir}(t)$ with $\overline{p}_{rr}(t) = 1$ for all t, we find that the probabilities $Q_i(t)$ can be obtained from the linear differential equations (verify!)

$$Q_i'(t) = -\nu_i Q_i(t) + \sum_{k \neq i,\, r} q_{ik} Q_k(t) \quad \text{for } t > 0 \text{ and } i \neq r.$$

These differential equations for the first-passage time probabilities $Q_i(t)$ can also be obtained from first principles, where $Q_i(t + \Delta t)$ is evaluated by conditioning on what may happen in $(0,\ \Delta t)$ for Δt small. This gives

$$Q_i(t + \Delta t) = (1 - \nu_i \Delta t) Q_i(t) + \sum_{k \neq i, r} q_{ik} \Delta t\ Q_k(t) + o(\Delta t).$$

Next, by letting $\Delta t \to 0$ in $[Q_i(t+\Delta t) - Q_i(t)]/\Delta t = \sum_{k \neq i,r} q_{ik} Q_k(t) + o(\Delta t)/\Delta t$ and noting that $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$, we get the linear differential equations for the $Q_i(t)$.

Example 16.4 (continued) For the special case of $s = 2$ units, let us derive an explicit expression for $Q_i(t)$ being the probability that the time until the first visit to state 2 (system is down) is more than t when the initial state is i for $i = 0, 1$. The linear differential equations for the $Q_i(t)$ are given by

$$\begin{aligned} Q_0'(t) &= -\lambda Q_0(t) + \lambda Q_1(t), \quad t > 0, \\ Q_1'(t) &= -(\lambda + \mu) Q_1(t) + \mu Q_0(t), \quad t > 0 \end{aligned}$$

with the boundary condition $Q_0(0) = Q_1(0) = 1$. Using standard theory for linear differential equations, it can be verified that

$$Q_i(t) = c_i e^{\eta_1 t} + (1 - c_i) e^{\eta_2 t}, \quad t > 0$$

for $i = 0, 1$, where $\eta_{1,2} = -\frac{1}{2}(2\lambda + \mu) \pm \frac{1}{2}\sqrt{\mu^2 + 4\lambda\mu}$, $c_0 = \eta_2/(\eta_2 - \eta_1)$ and $c_1 = (\eta_1 + \lambda)c_0/\lambda$. The first-passage time from state i to state 2 has the probability density $-Q_i'(t)$, which enables us to calculate the expected value of the first-passage time from state i to state 2. Solving a system of linear equations is in general a simpler method to calculate the expected values of the first-passage times, see Problem 16.7.

Problem 16.6 Let $\{J(t),\ t \geq 0\}$ be a stochastic process with finite state space I in which the state transitions occur at epochs generated by a Poisson process with rate ν. The state transitions are governed by a Markov matrix $\mathbf{R} = (r_{ij})$, that is, a state transition from i goes to j with probability r_{ij}.

(a) Let $p_{ij}(t) = P(J(t) = j \mid J(0) = i)$. Using the fact that the number of state transitions in $(0, t)$ has a Poisson distribution with expected value νt, verify that

$$p_{ij}(t) = \sum_{n=0}^{\infty} r_{ij}^{(n)} e^{-\nu t} \frac{(\nu t)^n}{n!} \quad \text{for } t > 0 \text{ and } i, j \in I,$$

where the $r_{ij}^{(n)}$ are the elements of the matrix product $\mathbf{R}^n$ with $\mathbf{R}^0$ the identity matrix $\mathbf{I}$.

(b) Let the numbers q_{ij} be the transition rates of a continuous-time Markov chain. Suppose the parameters ν and r_{ij} of the process $\{J(t)\}$ are chosen as $\nu = \max_i \nu_i$, $r_{ij} = q_{ij}/\nu$ for $j \neq i$ and $r_{ij} = 1 - \nu_i/\nu$ for $j = i$, where with $\nu_i = \sum_{j \neq i} q_{ij}$. Use part (a) to verify that the process $\{J(t)\}$ is a continuous-time Markov chain with transition rates q_{ij}. *Remark*: the

transition rates determine a unique continuous-time Markov chain and so the result of part (a) gives us an alternative method to compute the time-dependent probabilities. This method is known as the *uniformization method*.

Problem 16.7 Consider a continuous-time Markov chain with state space I and transition rates q_{ij}. For a given state r, define μ_i as the expected amount of time until the first visit to state r when the initial state is i with $i \neq r$. Give a system of linear equations for the μ_i.

Problem 16.8 Consider Problem 16.1 again. The system goes down if no standby unit is available upon failure of the operating unit. Let the random variable T denote the time until the system goes down for the first time given that each of the units is in a good condition at time 0. Take the numerical data $\lambda = 1$, $\eta = 0$ and $\mu = 50$. Calculate $E(T)$ and $P(T > t)$ for $t = 10$, 25, 50, 100, and 200. What are the numerical answers for the case of two standby units with the stipulation that it is only possible to replace one unit at a time?

Problem 16.9 The Hubble space telescope carries a variety of instruments including six gyroscopes to ensure stability of the telescope. The six gyroscopes are arranged in such a way that any three gyroscopes can keep the telescope operating with full accuracy. The operating times of the gyroscopes are independent of each other and have an exponential distribution with an expected value of $1/\lambda$ years. A fourth gyroscope failure triggers the sleep mode of the telescope. In the sleep mode further observations by the telescope are suspended. It requires an exponential time with an expected value of $1/\mu$ years to turn the telescope into sleep mode. Once the telescope is in the sleep mode, the base station on Earth receives a sleep signal. A shuttle mission to the telescope is then prepared. It takes an exponential time with an expected value of $1/\eta$ years before the repair crew arrives at the telescope and has repaired the stabilizing unit with the gyroscopes. In the mean time the other two gyroscopes may fail. If this happens, a crash destroying the telescope will be inevitable. What is the probability that the telescope will crash in the next T years? What is the probability that no shuttle mission will be prepared in the next T years? Compute these probabilities for the numerical data $T = 10, \lambda = 0.1, \mu = 100$ and $\eta = 5$.

16.3 Limiting probabilities

In this section the limiting behavior of the state probabilities $p_{ij}(t)$ will be studied under the assumption that the continuous-time Markov process $\{X(t)\}$

is irreducible. The process is said to be *irreducible* if for any two states i and j there is some $t > 0$ so that $p_{ij}(t) > 0$.

Rule 16.4 *Let $\{X(t)\}$ be an irreducible continuous-time Markov chain with a finite state space I. Then*

(a) $\lim_{t\to\infty} p_{ij}(t)$ *exists for all i, j and is independent of the starting state i;*
(b) *letting* $p_j = \lim_{t\to\infty} p_{ij}(t)$ *for $j \in I$, the limiting probabilities p_j are the unique solution to the equilibrium equations*

$$p_j \sum_{k\neq j} q_{jk} = \sum_{k\neq j} p_k q_{kj} \quad \text{for } j \in I \quad \text{and} \quad \sum_{j\in I} p_j = 1.$$

This rule is the continuous analog of Rule 15.5. Contrary to discrete-time Markov chains, we have for continuous-time Markov chains that possible periodicity in the state transitions is no issue, because the times between the state transitions are continuous random variables. A formal proof of Rule 16.4 will not be given. The linear equations in part (b) can be heuristically obtained from the Kolmogorov forward equations in Rule 16.3. This goes as follows. Since $p_{ij}(t)$ converges to a constant as $t \to \infty$, it will be intuitively clear that $\lim_{t\to\infty} p'_{ij}(t)$ must be equal to zero. Letting $t \to \infty$ on both sides of the differential equations in Rule 16.3, we obtain

$$0 = -\nu_j p_j + \sum_{k\neq j} q_{kj} p_k \quad \text{for } j \in I.$$

Noting that $\nu_j = \sum_{k\neq j} q_{jk}$, the balance equations in part (b) of Rule 16.4 follow. The normalization equation $\sum_{j\in I} p_j = 1$ is a direct consequence of $\sum_{j\in I} p_{ij}(t) = 1$. The normalization equation is needed to ensure the uniqueness of the solution; otherwise, the solution of the balance equations is uniquely determined up to a multiplicative constant.

The limiting probability p_j can be interpreted as the probability that an outside observer finds the system in state j when the observer enters the system after it has run for a very long time and the observer has no information about the past of the process. Another interpretation of the p_j is

$$\text{the long-run proportion of time the process will be in state } j = p_j$$

with probability 1, regardless of the starting state. Advanced probability theory is required to prove this deep result. Also, for any two states j and k with $k \neq j$, we can state that in the long run

$$\text{the average number of transitions per unit time from } j \text{ to } k = q_{jk} p_j$$

with probability 1, regardless of the starting state. These two results enable us to explain the balance equations in a physical way. To do so, note that in the

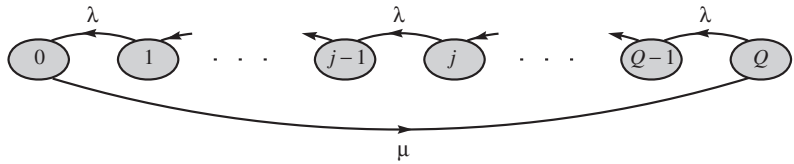


Fig. 16.1. The inventory process.

long run

$$\text{the average number of transitions per unit time out of } j = p_j \sum_{k \neq j} q_{jk}$$

and

$$\text{the average number of transitions per unit time into } j = \sum_{k \neq j} q_{kj}\, p_k.$$

Further, in any time interval the number of transitions out of j can differ by at most 1 from the number of transitions into j. Hence the long-run average number of transitions per unit time out of j must be equal to the long-run average number of transitions per unit time into j and thus $p_j \sum_{k \neq j} q_{jk} = \sum_{k \neq j} p_k q_{kj}$. This can be expressed as

$$\text{rate out of state } j = \text{rate into state } j.$$

The principle rate out of state j = rate into state j is the *flow-rate-equation method*. To write down the balance equations in specific applications, it is helpful to use the so-called *transition rate diagram* that visualizes the process. The nodes of the diagram represent the states and the arrows in the diagram give the possible state transitions. An arrow from node i to node j is only drawn when the transition rate q_{ij} is positive.

Example 16.2 (continued) How do we compute the long-run average stock on hand, the long-run average number of stock replenishments per unit time, the long-run fraction of time the system is out of stock, and the long-run fraction of demand that is lost?

Solution. The transition rates of the process $\{X(t)\}$ describing the number of units stock in the tank have been already determined in Section 16.1. The states and the transition rates of the process are displayed in the transition rate diagram of Figure 16.1. Using this diagram and the principle rate out of state j = rate into state j, it is easy to write down the balance equations. These equations are

given by

$$\mu p_0 = \lambda p_1, \quad \lambda p_j = \lambda p_{j+1} \text{ for } 1 \le j \le Q-1, \quad \lambda p_Q = \mu p_0.$$

These equations together with the equation $\sum_{j=0}^{Q} p_j = 1$ have a unique solution. In this special case, the solution can be explicitly given. It is readily verified that

$$p_0 = \left(1 + Q\,\frac{\mu}{\lambda}\right)^{-1}, \quad p_1 = p_2 = \cdots = p_Q = \frac{\mu}{\lambda} p_0.$$

The limiting probability p_j represents the long-run fraction of time that the number of units of stock in the tank is equal to j. Thus, with probability 1,

$$\text{the long-run average stock on hand} = \sum_{j=0}^{Q} j p_j = \frac{1}{2} Q(Q+1) \frac{\mu}{\lambda + Q\mu},$$

$$\text{the long-run fraction of time the system is out of stock} = p_0 = \frac{1}{1 + Q\mu/\lambda}.$$

A stock replenishment occurs each time the process makes a transition from state 0 to state Q. The long-run average number of transitions per unit time from 0 to Q is $q_{0Q} p_0 = \mu(1 + Q\mu/\lambda)^{-1}$ and so

$$\text{the long-run average frequency of stock replenishments} = \frac{\mu}{1 + Q\mu/\lambda}.$$

Inventory systems in which the demand process is a Poisson process have the remarkable property that the long-run fraction of demands finding the system out of stock is equal to the long-run fraction of time the system is out of stock. We can thus conclude that in the present model the long-run fraction of demand that is lost is equal to $p_0 = 1/(1 + Q\mu/\lambda)$.

Remark. A few words of explanation are in order for the last result in Example 16.2. Suppose you have a continuous-time Markov chain $\{X(t)\}$ in which customers/jobs arrive according to a Poisson process, where the arrival process can be seen as an exogenous factor to the system and is not affected by the system itself. Then, for each state j, the long-run fraction of customers/jobs finding the process in state j is equal to p_j, this being the long-run fraction of time the system is in state j. This property is known as *Poisson arrivals see time averages*. Intuitively, this property can be explained by the fact that Poisson arrivals occur completely random in time: each time interval $(t, t+h)$ has the same probability $\lambda h + o(h)$ of an arrival, no matter what happened up to time t. The property "Poisson arrivals see time averages" is also true for general stochastic processes under weak regularity conditions. This is an extremely useful result for queueing and inventory applications.

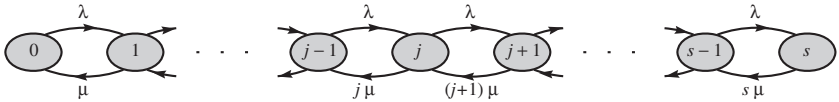


Fig. 16.2. The repair process.

Example 16.4 (continued) Suppose the system is down when all units are in repair. What are the long-run fraction of time the system is down and the long-run average number of units in repair?

Solution. The transition rates of the process $\{X(t)\}$ describing the number of units in repair have been already determined in Section 16.1. The states and the transition rates of the process are displayed in the transition rate diagram of Figure 16.2. Using this diagram and the principle rate out of state j = rate into state j, we find the balance equations

$$\lambda p_0 = \mu p_1, \quad (\lambda + j\mu)p_j = \lambda p_{j-1} + (j+1)\mu p_{j+1} \text{ for } 1 \le j \le s-1,$$
$$s\mu p_s = \lambda p_{s-1}.$$

These linear equations can be reduced to the recursive equation

$$j\mu p_j = \lambda p_{j-1} \quad \text{for } j = 1, 2, \ldots, s.$$

This recursion can be proved by induction. It is true for $j = 1$. Suppose it has been verified for $j = 1, \ldots, k$. Then, for $j = k+1$, it follows from the equation $(\lambda + k\mu)p_k = \lambda p_{k-1} + (k+1)\mu p_{k+1}$ and the induction assumption $k\mu p_k = \lambda p_{k-1}$ that $(k+1)\mu p_{k+1} = \lambda p_k$, which completes the induction proof. The recursive equations lead to an explicit solution for the p_j. Iterating $p_j = \frac{\lambda}{j\mu} p_{j-1}$ gives

$$p_j = \frac{\lambda}{j\mu} p_{j-1} = \frac{\lambda^2}{j\mu(j-1)\mu} p_{j-2} = \cdots = \frac{\lambda^j}{j\mu(j-1)\mu \cdots \mu} p_0.$$

Using the normalization equation $\sum_{j=0}^{s} p_j = 1$, it next follows that

$$p_j = \frac{(\lambda/\mu)^j / j!}{\sum_{k=0}^{s} (\lambda/\mu)^k / k!} \quad \text{for } j = 0, 1, \ldots, s.$$

By adding $e^{-\lambda/\mu}$ both to the denominator and to the numerator of the expression for the p_j, it follows that the p_j form a truncated Poisson distribution.

We can now answer the posed questions. Using the fact that p_j represents the long-run fraction of time that j units are in repair, it follows that

$$\text{the long-run average number of units in repair} = \sum_{j=0}^{s} jp_j,$$

$$\text{the long-run fraction of time the system is down} = p_s.$$

Remark. The reliability model for Example 16.4 is an application of the famous *Erlang loss model* from queueing theory. In the Erlang loss model customers arrive at a service facility with s identical servers according to a Poisson process with rate λ. A customer finding upon arrival all s servers busy is lost; otherwise, the customer gets assigned one of the free servers. The service times of the customers are independent random variables having a common exponential distribution with rate μ. To explain the equivalence of the reliability model with the Erlang loss model, let us formulate the reliability problem in other terms. Instead of assuming an exponentially distributed operating time with rate μ, we assume that there is an exogenous process in which catastrophes occur at epochs that are generated by a Poisson process with rate μ. If a catastrophe occurs, the unit in operation (if any) breaks down. This formulation of the reliability problem is equivalent to the original formulation (why?). By identifying units in repair with busy servers, it now follows that the reliability model is an application of the Erlang loss model. The formula for p_s is called the Erlang loss formula. The reason of this name is that, by the property of "Poisson arrivals see time averages," p_s also gives the long-run fraction of customers that upon arrival find all s servers busy and thus are lost. The Danish telephone engineer A.K. Erlang (1878–1929) was a genius. He was not only the first one who derived the formula named after him, but he also conjectured that the formula for the loss probability remains valid for generally distributed service times with expected value $1/\mu$. A proof of this important insensitivity result was only given many years after Erlang made his conjecture. More generally, for stochastic service systems in which customers arrive according to a Poisson process and customers never wait in a queue, steady-state performance measures are insensitive to the shape of the service-time distribution and require only the expected value of this distribution. Service systems in which customers never queue are loss systems and infinite-server systems.

Infinite state space and birth-and-death processes

To avoid technicalities, we have assumed so far a finite state space. What about a continuous-time Markov chain with a countably infinite number of states?

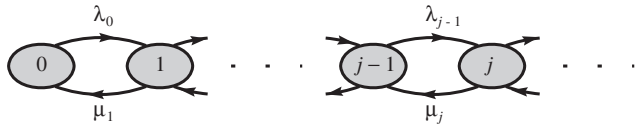


Fig. 16.3. The birth-and-death process.

One then has to rule out the possibility of an infinite number of state transitions in a finite time interval. For that purpose the assumption is made that the rates of the exponentially distributed holding times in the states are bounded. Under this assumption Rule 16.3 remains valid for the case of an infinite state space. To ensure that Rule 16.4 is also valid for the case of an infinite state space, a sufficient condition is the existence of some state r such that the expected time until the first transition into state r is finite for any starting state.

An important class of continuous-time Markov chains are the so-called birth-and-death processes. These processes have many applications. A continuous-time Markov chain with states $0, 1, \ldots$ is said to be a *birth-and-death process* if the transition rates satisfy $q_{ij} = 0$ for $|i - j| > 1$, that is, transitions are only possible to adjacent states. Queueing systems can often be modeled as a birth-and-death process in which the population consists of the number of waiting customers. Also, birth-and-death process are often used in applications in biology. In birth-and-death processes the limiting probabilities p_j can be given explicitly.

Rule 16.5 *Let* $\{X(t)\}$ *be a birth-and-death process with states* $0, 1, \ldots$ *and transition rates* $\lambda_i = q_{i,i+1}$ *and* $\mu_i = q_{i,i-1}$, *where* $\lambda_i > 0$ *for all* $i \geq 0$ *and* $\mu_i > 0$ *for all* $i \geq 1$. *Suppose the so-called ergodicity condition*

$$\sum_{j=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} < \infty$$

is satisfied. Then the limiting probabilities p_j *exist and satisfy the recursion*

$$\mu_j p_j = \lambda_{j-1} p_{j-1} \quad \textit{for all } j \geq 1.$$

An explicit expression for the p_j *is given by*

$$p_j = \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j} \, p_0 \quad \textit{for } j \geq 1,$$

where the probability $p_0 = \left(1 + \sum_{j=1}^{\infty} \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \cdots \mu_j}\right)^{-1}$.

It is instructive to derive the explicit formula for the p_j. It will appear from the derivation that the ergodicity condition is necessary for the existence of limiting probabilities. The reader has to take for granted that this condition is also sufficient for the existence of the p_j. The key idea for the derivation of the p_j is the observation that for any state $j \neq 0$ the long-run average number of transitions per unit time out of the set $\{j, j+1, \ldots\}$ must be equal to the long-run average number of transitions per unit time into the set $\{j, j+1, \ldots\}$ (explain!). This set can only be left by a transition from j to $j-1$ and can only be entered by a transition from $j-1$ to j. Thus

$$\mu_j p_j = \lambda_{j-1} p_{j-1} \quad \text{for all } j \geq 1.$$

Iterating this recursive equation gives

$$p_j = \frac{\lambda_{j-1}}{\mu_j} p_{j-1} = \frac{\lambda_{j-1}\lambda_{j-2}}{\mu_j \mu_{j-1}} p_{j-2} = \cdots = \frac{\lambda_{j-1}\lambda_{j-2}\cdots\lambda_0}{\mu_j \mu_{j-1}\cdots\mu_1} p_0.$$

Using the fact that $\sum_{j=0}^{\infty} p_j = 1$, the expression for p_0 follows and the derivation of the explicit formula for the p_j has been completed.

Example 16.5 Consider a multi-server station with s servers and infinite waiting room. Customers arrive according to a Poisson process with rate λ. If an arriving customer finds a server free, the customer immediately enters service; otherwise, the customer joins the queue. The service time of the customer has an exponential distribution with rate μ. This queuing model is often called the *Erlang delay model*. Letting $X(t)$ denote the number of customers in the system (including those in service), the continuous-time Markov chain $\{X(t)\}$ is a birth-and-death process with states $0, 1, \ldots$. A birth occurs when a new customer arrives and a death occurs when a service is completed. The parameters of this birth-and-death process are given by

$$\lambda_i = \lambda \quad \text{for all } i, \quad \mu_i = i\mu \quad \text{for } 1 \leq i \leq s-1 \quad \text{and} \quad \mu_i = s\mu \quad \text{for } i \geq s.$$

The ergodicity condition in Rule 16.5 boils down to the condition $\lambda < s\mu$. This is a natural condition which states that the arrival rate must be smaller than the total service rate. It is convenient to use the notation

$$\rho = \frac{\lambda}{s\mu}.$$

It is a matter of simple algebra to verify that under the condition $\rho < 1$ the expression for the probabilities p_j in Rule 16.5 reduces to

$$p_j = \frac{(s\rho)^j}{j!} p_0 \quad \text{for } 0 \leq j \leq s-1, \quad p_j = \frac{(s\rho)^s \rho^{j-s}}{s!} p_0 \quad \text{for } j \geq s,$$

where

$$p_0 = \left(\sum_{k=0}^{s-1} \frac{(s\rho)^k}{k!} + \frac{(s\rho)^s}{s!} \frac{1}{1-\rho} \right)^{-1}.$$

The limiting probability p_j represents the long-run fraction of time that j customers are present in the system. Of particular interest are the performance measures $P_{\text{delay}} = \sum_{j=s}^{\infty} p_j$ and $L_{\text{q}} = \sum_{j=s}^{\infty} (j-s)p_j$, where P_{delay} represents the long-run fraction of time that all s servers are occupied and L_{q} represents the long-run average number of customers waiting in queue (excluding those in service). It is a matter of simple algebra to verify that

$$P_{\text{delay}} = \frac{(s\rho)^s}{s!(1-\rho)} p_0 \quad \text{and} \quad L_{\text{q}} = \frac{\rho}{1-\rho} P_{\text{delay}}.$$

The formula for P_{delay} is often called the *Erlang delay formula*. By the property Poisson arrivals see time averages, the long-run fraction of customers finding upon arrival s or more other customers present is equal to the long-run fraction of time that s or more customers are present. Thus P_{delay} also gives the long-run fraction of customers who have to wait in the queue. Observe that the average queue size L_{q} increases in a nonlinear way as a function of the load factor ρ and explodes when ρ approaches 1. Another important quantity is the long-run average waiting time in the queue per customer (excluding service time). Denote this quantity by W_{q}. Observe that service completions occur at a rate of $s\mu$ as long as all s servers are busy. Using this observation together with the fact that the long-run fraction of customers finding upon arrival k other customers present is equal to the long-run fraction of time that k customers are present, it follows that $W_{\text{q}} = \sum_{k=s}^{\infty} \frac{k-s+1}{s\mu} p_k$. This leads to

$$W_{\text{q}} = \frac{1}{s\mu(1-\rho)} P_{\text{delay}}.$$

Comparing the formulas for L_{q} and W_{q}, you see that $L_{\text{q}} = \lambda W_{\text{q}}$. In words, the average number of customers waiting in the queue is equal to the average arrival rate of customers times the average waiting time in the queue per customer. This result is known as *Little's formula*. This famous formula is valid for almost any queueing system.

For the special case of a single server ($s = 1$), the above formulas simplify to $p_j = (1-\rho)\rho^j$ for all $j \geq 0$, $P_{\text{delay}} = \rho$, $L_{\text{q}} = \rho^2/(1-\rho)$ and $W_{\text{q}} = \rho/[\mu(1-\rho)]$.

To illustrate the foregoing, suppose that the management of a harbor is facing the decision to replace the current unloader. The choice is between one fast unloader and two slow unloaders. The fast unloader can handle on

average two ships per eight and a half hours and each of the slow unloaders can handle on average one ship per eight and a half hours. There is activity in the harbor day and night. Ships with cargo arrive according to a Poisson process with a rate of four ships per twenty-four hours. There is ample berth space for the ships. The unloading times of the ships are independent random variables with a same exponential distribution. What are the delay probability and the average sojourn time per ship for each of the two alternatives? To answer the question, we apply the above queueing model both for $s = 1$ and $s = 2$. Take the hour as unit of time. For the case of one fast unloader ($s = 1$), we have an arrival rate of $\lambda = 4/24$ and a service rate of $\mu = 2/8.5$. This gives a delay probability of $P_{\text{delay}} = 0.7083$ and an average sojourn time per ship of $W_{\text{q}} + 1/\mu = 10.32 + 4.25 = 14.57$ hours. For the case of two slow unloaders ($s = 2$), we have $\lambda = 4/24$ and $\mu = 1/8.5$. This gives a delay probability of $P_{\text{delay}} = 0.5874$ and an average sojourn time per ship of $W_{\text{q}} + 1/\mu = 8.56 + 8.5 = 17.06$ hours.

Problem 16.10 Consider Problem 16.1 again. The system is down when all units are in repair. What is the long-run fraction of time the system is down?

Problem 16.11 An assembly line for a certain product has two stations in series. Each station has only room for one unit of the product. If the assembly of a unit is completed at station 1, it is forwarded immediately to station 2 provided station 2 is idle; otherwise, the unit remains in station 1 and blocks this station until station 2 becomes free. Units for assembly arrive at station 1 according to a Poisson process with rate λ, but a newly arriving unit is only accepted by station 1 when no other unit is present at station 1. Each unit rejected is handled elsewhere. The assembly times at stations 1 and 2 have an exponential distribution with respective rates μ_1 and μ_2. Specify the transition rate diagram and give the balance equations. What is the long-run fraction of time that station 1 is blocked? What is the long-run fraction of time that station 2 is busy? What is the long-run fraction of items that are rejected?

Problem 16.12 Consider Problem 16.2 again. What is the average number of cars in the gasoline station? What is the long-run fraction of time the pump is occupied and what is the long-run fraction of cars that cannot enter the station?

Problem 16.13 Consider Problem 16.5 again. What is the long-run run average number of waiting passengers? What is the long-run fraction of potential passengers who go elsewhere for transport?

Problem 16.14 In an inventory system for a single product the depletion of stock is due to demand and deterioration. The demand process for the product is

a Poisson process with rate λ. The lifetime of each unit product is exponentially distributed with mean $1/\mu$. The stock control is exercised as follows. Each time the stock drops to zero a replenishment order for Q units is placed. The lead time of each order is negligible. Specify the transition rate diagram and give the balance equations. What is the average stock on hand and what is the average number of orders placed per time unit?

Problem 16.15 Suppose one wishes to determine the capacity of a stockyard at which containers arrive according to a Poisson process with a rate of $\lambda = 1$ per hour. A container finding a full yard upon arrival is taken elsewhere. The time that a container is stored in the yard is exponentially distributed with mean $1/\mu = 10$ hours. Determine the required capacity of the yard so that no more than 1% of the arriving containers find the yard full. How does the answer change when the time that a container is stored in the yard is uniformly distributed between 5 and 15 hours?

Problem 16.16 Long-term parkers and short-term parkers arrive at a parking place for cars according to independent Poisson processes with respective rates of $\lambda_1 = 4$ and $\lambda_2 = 6$ per hour. The parking place has room for $N = 10$ cars. Each arriving car which finds all places occupied goes elsewhere. The parking time of long-term parkers is uniformly distributed between 1 and 2 hours, while the parking time of short-term parkers has a uniform distribution between 20 and 60 minutes. Calculate the probability that a car finds all parking places occupied upon arrival.

Problem 16.17 Consider a continuous-review inventory system in which customers asking for a certain item arrive according to a Poisson process with rate λ. Each customer asks for one unit of the item. Customer demands occurring when the system is out of stock are lost. The $(S-1, S)$ control rule is used. Under this control rule the base stock is S and a replenishment order for exactly one unit is placed each time the stock on hand decreases by one unit. The lead time of each replenishment order is a constant $L > 0$. What is limiting distribution of the stock on hand? What is the long-run average stock on hand and what is the long-run fraction of demand that is lost?

Problem 16.18 Consider the multi-server queueing model from Example 16.5 again. Let W_n be the waiting time in the queue of the nth arriving customer. Under the assumption that customers enter service in order of arrival, verify that

$$\lim_{n\to\infty} P(W_n > t) = P_{\text{delay}}\, e^{-s\mu(1-\rho)t} \qquad \text{for all } t \geq 0.$$

Hint: note that the times between service completions have an exponential density with parameter $s\mu$ if all servers are busy and use the fact that the sum of k independent exponentials with parameter η is a gamma distributed random variable S_k with $P(S_k > x) = \sum_{i=0}^{k-1} e^{-\eta x}(\eta x)^i / i!$ for $x \geq 0$.

Problem 16.19 Phone-calls with service requests arrive at a call-center according to a Poisson process with rate λ. There are s agents to handle the service requests. The service times of the requests are independent random variables having an exponential density with rate μ. An arriving request gets immediately assigned a free server if one is available; otherwise, the request joins the queue. If the service of a request has not yet started, the caller becomes impatient after an exponentially distributed time with expected value $1/\theta$ and then leaves the system. Give a recursive scheme for the limiting distribution of the number of requests receiving service or waiting in the queue. What is the long-run fraction of balking callers?

Problem 16.20 Customers arrive at a service facility with infinite waiting room according to a Poisson process with rate λ. An arriving customer finding upon arrival j other customers present joins the queue with probability $1/(j+1)$ for $j = 0, 1, \ldots$ and goes elsewhere otherwise. There is single server who can handle only one customer at a time. The service time of each customer is exponentially distributed with rate μ. What are the limiting probabilities of the number in system? What is the long-run fraction of customers who go elsewhere for service?

Problem 16.21 Customers arrive at a service facility according to a Poisson process with rate λ. Each arriving customer gets immediately assigned a free server. There are ample servers. The service times of the customers are independent random variables having a common exponential distribution with mean $1/\mu$. What is the limiting distribution of the number of busy servers? Do you think this limiting distribution is insensitive to the shape of the service-time distribution? Imagine oil tankers leave Norway with destination Rotterdam according to a Poisson process with an average of two tankers per day, where the sailing time to Rotterdam is gamma distributed with an expected value of two days and a standard deviation of one day. Use the infinite-server model to estimate the probability that more than seven oil tankers are under way from Norway to the harbor of Rotterdam.

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 6. 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 6 different ways by reasoning as follows. For the first position, there are 3 available letters to choose from, for the second position there are 2 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 Prime Minister, 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 Prime Minister 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 Prime Minister. 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 Prime Minister are fixed, there remain 6! possible arrangements for the other six statesmen. The U.S. President and the British Prime Minister 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 Prime Minister stand next to each other, and so the sought probability equals $(6! \times 7 \times 2!)/8! = 1/4$.

To test your understanding of permutations, you are asked to verify that the total number of distinguishable permutations of the eleven letters in the word Mississippi is 34,650.

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 10: $\{A, B, C\}$, $\{A, B, D\}$, $\{A, B, E\}$, $\{A, C, D\}$, $\{A, C, E\}$, $\{A, D, E\}$, $\{B, C, D\}$, $\{B, C, E\}$, $\{B, D, E\}$, $\{C, D, E\}$. In this problem, the order in which the jury members are chosen is not

relevant. The answer of 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 $5\times4\times3$ 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

$$\begin{aligned}&\frac{n\times(n-1)\times\cdots\times(n-k+1)}{k!}\\&=\frac{n\times(n-1)\times\cdots\times(n-k+1)\times(n-k)\times\cdots\times1}{k!\times(n-k)!}\\&=\frac{n!}{k!\times(n-k)!}.\end{aligned}$$

For nonnegative integers n and k with $k\le n$, 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, respectively.

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 of 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 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.

[†] 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.

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

$$\begin{aligned}(1 - x)\sum_{k=0}^{n} x^k &= \sum_{k=0}^{n} x^k - \sum_{k=0}^{n} x^{k+1} \\ &= (1 + x + \cdots + x^n) - \left(x + x^2 + \cdots + x^n + x^{n+1}\right) \\ &= 1 - x^{n+1}.\end{aligned}$$

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 Yes, the same sample space with equally likely elements applies.

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 37.5% of your bankroll in the risky project each time. The effective rate of return is 6.6%.

2.11 The probability is 0.875.

2.13 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.

2.15 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.

2.17 The expected values in parts (a), (b), (c), and (d) are 0.333, 0.521, 0.905, and 0.365, respectively.

2.19 The house percentage is 6.1%.

2.21 The expected values of your loss and the total amount you bet are \$0.942 and \$34.86.

2.23 The probabilities are 0.200 and 0.045.

2.25 The probability of the bank winning is 0.809, and the average number of points collected by the bank is 9.4.

2.27 The optimal value of L is 50 and the maximal probability of candidate A winning is 0.482. *Remark*: if a tie is broken by drawing lots, then the optimal value of L remains the same but the probability of player A winning becomes 0.458 when both players act optimally.
2.29 The expected payoff is \$0.60 and the probability of getting 25 or more points is 0.693.
2.31 The probability is 0.257.
2.33 The probabilities are 0.705 and 0.002.
2.35 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.
2.37 The probabilities are 0.587, 0.312, 0.083, and 0.015.
2.39 The probabilities are 0.143, 0.858, 0.833, 0.800, 0.750, 0.667, and 0.500.
2.41 The probability is 0.60.
2.43 The probabilities are 0.329 and 0.536.
2.45 The simulated value of the expected number of turns is 17.133. The simulated values of the probability that more than r turns are needed are 0.9578, 0.9087, 0.8447, 0.7725, 0.6251, 0.4432, 0.2563, 0.1496, 0.0871, 0.0502, 0.0289, and 0.0097 for $r = 7, 8, 9, 10, 12, 15, 20, 25, 30, 35, 40$, and 50. The results are based on 100,000 simulation runs.

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 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 expected value is 61.217 rolls. The probability is 5.26×10^{-3} and the house edge is 7.99% (see also Example 4.8).
3.19 The probability of the player winning on the main point is 0.1910 and the probability of the player winning on a chance point is 0.3318. The house percentage is 5%.
3.21 36.
3.23 The house percentage is 2.88%.
3.25 The probabilities are 0.0515, 0.1084, 0.1592, and 0.1790.
3.27 The probability is 0.7853. The expected values are \$57.64 and \$2,133.

3.29 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 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%.

4.11 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.

4.13 The Poisson approximations are $1 - e^{-\lambda_0} = 0.0171$ 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.

4.15 Letting $\lambda = \binom{25}{2} \times \left(\frac{14}{365} \times \frac{1}{365}\right)$, a Poisson approximation to the desired probability is $1 - e^{-\lambda} = 0.0310$. The simulated value is 0.0299.

4.17 The Poisson approximations for the two probabilities are 0.487 and 0.059. The exact value of the first probability is $1 - \binom{40}{6}/\binom{45}{6} = 0.5287$. Simulation gives the value 0.056 for the other probability.

4.19 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 2.37).

4.21 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 to the desired probability is $1 - \sum_{k=0}^{6} e^{-\lambda}\lambda^k/k! = 0.499$. The simulated value is 0.516.

4.23 Let E_i be the expected number of times that a team scores i goals in a match for $0 \le i \le 3$, and let E_{4+} be the expected number of times that a team scores four or more goals in a match. Then, using a Poisson distribution with parameter $\lambda = 1.3125$, $E_0 = 25.8$, $E_1 = 33.9$, $E_2 = 22.3$, $E_3 = 9.7$, and $E_{4+} = 4.3$.

4.25 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.

4.27 The win probability is 0.026354 and the house edge is 7.76%.

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. 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 $\sigma = \frac{1.708}{\sqrt{1{,}000}} = 0.054$. The reported value of 3.25 lies 4.63 standard deviations below 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 $1 - \Phi(2.828) = 0.0023$.

5.27 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.

5.29 The probability can be approximated by $0.5 + 0.5 \times e^{-0.953419 \ln(2)} = 0.7582$. Simulation indicates that the exact value is about 0.713.

5.31 An approximation to the probability is 0.2582.

5.33 9.18 time units.

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 probabilities are $\frac{108}{216}$ and $\frac{27}{216}$.

7.3 Label the nine socks as $s_1, \ldots, s_9$. If the order in which the socks are chosen is considered important, then the sample space consists of $9 \times 8 = 72$ outcomes. There are $4 \times 5 = 20$ outcomes for which the first sock chosen is black and the second is white. Also, there $5 \times 4 = 20$ outcomes for which the first sock is white and the second is black. Hence the desired probability is $(20 + 20)/72 = 5/9$. Another probability model for this problem is one in which the order of selection of the socks is not considered relevant. Then the sample space consists of $\binom{9}{2} = 36$ outcomes. The number of outcomes in which the socks have different colors is $\binom{5}{1} \times \binom{4}{1} = 20$, giving the probability $20/36 = 5/9$.

7.5 The probability is $\frac{2}{3}$.

7.7 The probability is $\frac{1}{210}$.

7.9 The probability is $\binom{11}{4} \times \binom{7}{4} \times \binom{3}{2} \times 4^4 \times 4^4 \times 2^2 \times 1/11^{11} = 0.0318$.

7.11 The sample space is $\Omega = \{(x, y) : 0 \le x \le a,\ 0 \le y \le a\}$. The desired probability is $(a - d)^2/a^2$.

7.13 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 $q \ge 4$.

7.15 The first probability is $(h - d)^2/h^2$. The second probability is $\frac{1}{2} + \frac{\pi}{6\sqrt{3}} = 0.8023$.

7.17 The probability is $\pi(\frac{1}{2}r)^2/(\pi r^2) = \frac{1}{4}$.

7.19 The probabilities are $\ln(2) - 0.5 = 0.1931$ and $2\ln(2) - 1 = 0.3863$.

7.21 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 $\sum_{k=1}^{\infty} P(A_k) < \infty$.

7.23 The desired probability is equal to $\sum_{k=0}^{\infty} \left(\frac{7}{10}\right)^{2k} \frac{3}{10} = 0.5882$.

7.25 The desired probability is at least as large as $1 - P(\cap_{n=1}^{\infty} B_n)$. For any $n \ge 1$, $P(B_n) = (1 - (\frac{1}{2})^r - (\frac{1}{2})^r)^n$. By the continuity property of probabilities, $P(\cap_{n=1}^{\infty} B_n) = \lim_{n\to\infty} P(B_n)$ and so $1 - P(\cap_{n=1}^{\infty} B_n) = 0$.

7.27 The probabilities are 0.3412 and 0.00198.

7.29 The desired probability is 0.45.

7.31 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 $P(A \cup B) = 0.467$ and $P(A \cup B \cup C) = 0.543$.

7.33 The probabilities are 0.0154 and $1 - 0.3349 = 0.6651$.

7.35 The probability is $\frac{2\times 1}{5\times 4} + \frac{2\times 1}{5\times 4} = \frac{1}{5}$.

7.37 The probabilities are $\frac{1}{n}$, $\frac{1}{r}$, and $\frac{1}{2}$.

7.39 Let A_i be the event that the spinner wins on the ith toss. The win probability is $\sum_{i=3}^{15} P(A_i) = 0.11364$. The game is unfavorable to the spinner.

7.41 The probability is 0.1.

7.43 The probability is 0.8109.

7.45 The probabilities are 0.0511 and 0.7363.

7.47 The desired probability is $\binom{13}{1} \times \left[\binom{48}{9}/\binom{52}{13}\right] - \binom{13}{2} \times \left[\binom{44}{5}/\binom{52}{13}\right] + \binom{13}{3}\left[\binom{40}{1}/\binom{52}{13}\right] = 0.0342$.

7.49 The probability is $1 - \sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k}\frac{(n-k)^r}{n^r}$. The required number of spins is 152 (the corresponding probability is 0.5016).

7.51 The probability is $\binom{10}{2}(\frac{1}{9})^2 - \frac{1}{2!}\binom{10}{2}\binom{8}{2}(\frac{1}{9})^4 + \cdots - \frac{1}{5!}\binom{10}{2}\binom{8}{2}\binom{6}{2}\binom{4}{2}\binom{2}{2}(\frac{1}{9})^{10} = 0.4654$.

Chapter 8

8.1 The desired probability is $\frac{18/216}{120/216} = 0.15$.

8.3 The desired probability is $\frac{0.18}{0.26} = 0.692$.

8.5 The probabilities are $\frac{1}{5}$ and $\frac{1}{3}$.

8.7 The probabilities are 0.3696 and 0.5612.

8.9 Take the 20th position in the line. The probability of winning the free ticket is 0.0323.

8.11 The desired probability is $\frac{2}{7} \times \frac{2}{6} \times \frac{2}{5} \times \frac{2}{4} \times \frac{2}{3} = 0.0127$. The probability $2^6/7! = 0.0127$ can also be obtained by clever counting arguments.

8.13 The probability that the ith person gets the winning ball equals $\frac{1}{N}$ for $i = 1$ and equals $\frac{N-1}{N} \times \cdots \times \frac{N-i+1}{N-i+2} \times \frac{1}{N-i+1} = \frac{1}{N}$ for $2 \le i \le N$.

8.15 The probability is $1 - \frac{5}{18} = 0.7222$.

8.17 Yes: $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2}$.

8.19 The probabilities are $\frac{1}{2}$ and $\frac{2}{3}$.

8.21 The probability is $\frac{5}{6}$.

8.23 The probability is $\frac{6}{13}$ and is the solution of the equation $x = \frac{6}{36} \times \frac{6}{36} + \frac{6}{36} \times \frac{29}{36} \times x + \frac{29}{36} \times x + \frac{1}{36} \times 0$.

8.25 The recursion is $p(i,t) = \frac{1}{6-i+1}\sum_{u=0}^{6-i} p(i+1, t-u)$ with the convention $p(j,l) = 0$ for $l \le 0$ and the boundary condition $p(6,t) = 1$ for $t \ge 1$. The desired probability $p(1,6) = \frac{169}{720}$.

8.27 The probabilities are $a_5 = 0.4063$, $a_{10} = 0.1406$, $a_{25} = 5.85 \times 10^{-3}$, and $a_{50} = 2.93 \times 10^{-5}$.

8.29 The optimal strategy of player A is to stop after the first spin if this spin gives a score of more than 65 points. The win probabilities of the players A, B, and C are 0.3123, 0.3300, and 0.3577 (see also the solution of Problem 13.21).

8.31 It follows from $\frac{p}{1-p} = \frac{1/5}{4/5}\frac{(0.75)^3}{(0.5)^3} = 0.84375$ that $p = 0.4576$.

8.33 (a) $\frac{p}{1-p} = \frac{1/4}{3/4} \times \frac{1}{2/3}$ and so $p = \frac{1}{3}$; (b) $\frac{p}{1-p} = \frac{1/4}{3/4} \times \frac{13/49}{2/21}$ and so $p = \frac{13}{27}$; (c) $\frac{p}{1-p} = \frac{1/4}{3/4} \times \frac{1-(1-k/7)^2}{2k/21}$ and so $p = \frac{14-k}{28-k}$ for $k = 1, \ldots, 7$.

8.35 Let H denote the hypothesis that the suspect is guilty and let E denote the event that the suspect makes a confession. Then, by Bayes' rule in odds form,

$P(H \mid E) > P(H)$ only if $P(E \mid H) > P(E \mid \overline{H})$ (use the fact that $p/(1-p) > q/(1-q)$ for $0 < p, q < 1$ only if $p > q$).

8.37 The posterior probabilities are 0.2116, 0.3333, and 0.4550.

8.39 The posterior probabilities are 0.0035, 0.9181, and 0.0784.

Chapter 9

9.1 $P(X=1) = \frac{1}{36}$, $P(X=2) = \frac{3}{36}$, $P(X=3) = \frac{5}{36}$, $P(X=4) = \frac{7}{36}$, $P(X=5) = \frac{9}{36}$, and $P(X=6) = \frac{11}{36}$.

9.3 $P(X=0) = \frac{1}{8}$, $P(X=1) = \frac{2}{8}$, and $P(X=2) = \frac{5}{8}$.

9.5 (a) The expected payoff is equal to $100 \times \frac{6}{1{,}296} + 15 \times \frac{54}{1{,}296} = 0.9846$ dollars. (b) The expected payoff is $\frac{1}{4} \times 1 + \frac{1}{8} \times 2 + \frac{1}{16} \times 3 + \frac{1}{32} \times 4 + \frac{1}{32} \times (5+25) = 1.75$ dollars.

9.7 The expected number of points is 4.77.

9.9 The optimal strategy is to stop after the first spin if this spin gives a score of more than 414 points.

9.11 Denote by the random variable X the payoff of the game. Then, $E(X) = 2(\frac{1}{2} - \frac{1}{3}) \times 1 + 2(\frac{1}{3} - \frac{1}{4}) \times 2 + \cdots + 2(\frac{1}{m} - \frac{1}{m+1}) \times (m-1) + \frac{2}{m+1} \times m$.

9.13 $\sum_{k=0}^{\infty} P(X > k) = \sum_{k=0}^{\infty} \sum_{j=k+1}^{\infty} P(X = j)$. By interchanging the order of summation, $\sum_{k=0}^{\infty} P(X > k) = \sum_{j=0}^{\infty} \sum_{k=0}^{j-1} P(X = j) = \sum_{j=0}^{\infty} j P(X = j)$, which proves the desired result. The expected values are 24.617 (slightly larger than the median of 23 people) and 14.7.

9.15 (a) The expected number of distinct birthdays is 87.6. (b) The expected number of children sharing a birthday is $r \times [1 - (364/365)^s]$.

9.17 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, $E(X_1 + \cdots + X_{44}) = \frac{2}{3}$.

9.19 The expected value and the standard deviation of X are 4 and 2.449. The expected value and the standard deviation of the number of reimbursed treatments are 3.364 and 1.553.

9.21 $E(X) = 0.7087$, $\sigma(X) = 0.7440$, and $P(X \geq 1) = 0.5524$.

9.23 The expected value is \$281.00 and the standard deviation is \$555.85.

9.25 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 $7 \times \frac{161}{36}$.

9.27 The expected value and the variance are 22.5 and 71.25.

9.29 Let X_i denote the number of draws needed to go from i different integers to $i+1$ different integers for $i = 1, 2$. You need r draws 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 $j \geq 3$.

9.31 The probability is 0.5442.

9.33 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.

9.35 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 $r \to \infty$.

9.37 The probability is 0.8675.

9.39 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.
9.41 The probabilities are 0.4226 and 0.0475.
9.43 The probability is $\binom{2m-k}{m} p^{m+1}(1-p)^{m-k} + \binom{2m-k}{m}(1-p)^{m+1} p^{m-k}$.
9.45 The house percentage of the lottery is 36.9%.
9.47 The number of rounds is geometrically distributed with parameter $p = 146/1{,}296$. The probability of John paying is $38/73$.
9.49 The game is not advantageous for the player. The house percentage is 1.62%.

Chapter 10

10.1 The constant $c = 2$. The probabilities are 0.5556, 0.2844 and 0.6400.
10.3 The probability is 0.3906.
10.5 The probability density function of Y is $\frac{f(\sqrt{y})}{2\sqrt{y}}$ for $y > 0$ and is 0 otherwise. The density function of W is $\frac{1}{2a\sqrt{w}}$ for $0 < w < a^2$ and is 0 otherwise.
10.7 The random variable V satisfies $P(V \le v) = P(X \le v/(1+v)) = \frac{v}{1+v}$ for $v \ge 0$. Its density function is equal to $\frac{1}{(1+v)^2}$ for $v > 0$ and 0 otherwise. The random variable W satisfies $P(W \le w) = 1 - \sqrt{1-4w}$ for $0 \le w \le \frac{1}{4}$ and its density function is equal to $2(1-4w)^{-1/2}$ for $0 < w < \frac{1}{4}$ and 0 otherwise.
10.9 By $P(V \le v) = v^2$ for $0 \le v \le 1$, the random variable V has the density function $f(v) = 2v$ for $0 < v < 1$ and $f(v) = 0$ otherwise. By $P(W > w) = 1 - (1-w)^2$ for $0 \le w \le 1$, the random variable W has the density function $g(w) = 2(1-w)$ for $0 < w < 1$ and $g(w) = 0$ otherwise.
10.11 The expected value is $73\frac{1}{3}$ meters.
10.13 Let X be the distance from the point to the origin. Then $P(X \le a) = \frac{1}{4}\pi a^2$ for $0 \le a \le 1$ and $P(X \le a) = \frac{1}{4}\pi a^2 - 2\int_1^a \sqrt{a^2 - x^2}\,dx = \frac{1}{4}\pi a^2 - a^2 \arccos(\frac{1}{a}) + \sqrt{a^2-1}$ for $1 < a \le \sqrt{2}$. The density function $f(x)$ of X satisfies $f(x) = \frac{1}{2}\pi x$ for $0 < x \le 1$ and $f(x) = \frac{1}{2}\pi x - 2x \arccos(\frac{1}{x})$ for $1 < x < \sqrt{2}$. Numerical integration leads to $E(X) = \int_0^{\sqrt{2}} x f(x)\,dx = 0.765$.
10.15 The expected value is $\int_0^h x\,\frac{2(h-x)}{h^2}\,dx = \frac{1}{3}h$, see also Problem 7.15.
10.17 The expected value of the net profit is 194.10 dollars. The probability of running out of stock is 0.0404.
10.19 The expected value is $\frac{3}{4}r$ and the standard deviation is $0.194r$.
10.21 (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) $E(|X-c|) = \int_{-\infty}^{c}(c-x)f(x)dx + \int_c^{\infty}(x-c)f(x)dx$. Putting the derivative of this function of c equal to 0 gives $F(c) = \frac{1}{2}$, where $F(c) = P(X \le c)$.
10.23 The expected value and standard deviation of the shortage are given by $\frac{1}{\lambda}e^{-\lambda s}$ and $\frac{1}{\lambda}\left[e^{-\lambda s}(2 - e^{-\lambda s})\right]^{1/2}$.
10.25 The median is 3.
10.27 The expected value and the standard deviation are 6.8 and 2.613.

10.29 $E(Y) = \frac{1}{2}\sqrt{\pi/\lambda}$ and $\text{var}(Y) = (1 - \pi/4)/\lambda$ (use the substitution rule and the gamma function).

10.31 The expected value and the median are 50 minutes and 34.7 minutes. Each of the desired probabilities is $e^{-1.2\times 1/3} - e^{-1.2\times 3/4} = 0.2638$.

10.33 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, $E(X^2) = \frac{\Gamma(\alpha+2)}{\lambda^2\Gamma(\alpha)} = \frac{(\alpha+1)\alpha}{\lambda^2}$.

10.35 The probability is $\Phi(-k) + 1 - \Phi(k) = 2 - 2\Phi(k)$.

10.37 The claim is highly implausible. The number of heads in 10,000 tosses of a fair coin is approximately normally distributed with expected value 5,000 and standard deviation 50. The probability that a normally distributed random variable takes on a value 5 or more standard deviations above the expected value is $1 - \Phi(5) = 2.87 \times 10^{-7}$.

10.39 Approximate $P(-\frac{1}{2} \le S_n < \frac{1}{2})$ by $\Phi((6+n)/\sqrt{11n}) - \Phi((-6+n)/\sqrt{11n})$. The approximate values are 0.288, 0.095, 0.022 and 0.0022 for $n = 10, 25, 50$ and 100. The values obtained by one million simulation runs are 0.273, 0.094, 0.023, and 0.0021.

10.41 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 log-normal distribution with parameters $\mu = \ln(s_0)$ and $\sigma = 0.22314\sqrt{n}$.

10.43 The density is $\frac{1}{y\ln(10)}$ for $1 < y < 10$ and zero otherwise. The expected value and the standard deviation are 3.909 and 2.494.

10.45 A random observation is $-\frac{1}{\lambda}\ln(u_1 \times \cdots \times u_n)$, where $u_1, \ldots, u_n$ are independent random numbers from (0, 1).

10.47 Generate two random numbers u_1 and u_2 from (0, 1). The random observation is $-(1/\lambda_1)\ln(u_2)$ if $u_1 \le p$ and is $-(1/\lambda_2)\ln(u_2)$ otherwise.

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 all i and $P(X = i,\ Y = j) = \frac{2}{36}$ for $i < j$.

11.3 $P(X = i,\ Y = j) = (j - i - 1)/\binom{10}{3}$ for $1 \le i \le 8$ and $i + 2 \le j \le 10$. Further, $P(X = i) = (10 - i)(9 - i)/240$ for $1 \le i \le 8$, $P(Y = j) = (j - 1)(j - 2)/240$ for $3 \le j \le 10$, and $P(Y - X = k) = (10 - k)(k - 1)/120$ for $2 \le k \le 9$.

11.5 $P(X = i,\ Y = j) = \binom{i+j-2}{i-1}(\frac{1}{3})^{i+j-1}$ for $i,\ j \ge 1$, $P(X = i) = (\frac{1}{2})^i$ for $i \ge 1$, and $P(X = Y) = \frac{1}{3}\sum_{k=0}^\infty \binom{2k}{k}(\frac{1}{9})^k = 0.4472$, where the latter result uses the identity $\sum_{k=0}^\infty \binom{2k}{k}x^k = 1/\sqrt{1-4x}$ for $|x| < \frac{1}{4}$.

11.7 The constant $c = (15/4)(4\sqrt{2} - 2)^{-1}$. The density of $Z = X + Y$ is given by $f(z) = cz\sqrt{z}$ for $0 < z < 1$ and $f(z) = c(2 - z)\sqrt{z}$ for $1 \le z < 2$.

11.9 Let X and Y denote the smallest and the largest of the two random numbers. Then, $P(x \le X \le x + \Delta x,\ y \le Y \le y + \Delta y) = 2\Delta x \Delta y$ 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 $P(V \le v) = \int_0^1 2\,dx \int_x^{\min(1,x+v)} dy$, we have $P(V \le v) = 2v - v^2$ for $0 \le v \le 1$, 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 $0 \le a \le 1/3$.

11.11 The joint density function $f(x, y) = 2$ equals 2 for (x, y) inside the triangle and $f(x, y) = 0$ otherwise. Using Rule 11.1, it follows that $P(V \le v) = 2v^2$ for $0 \le v \le \frac{1}{2}$ and $P(v \le v) = 4v - 2v^2 - 1$ for $\frac{1}{2} \le v \le 1$. The probability distribution functions of W and Z are given by $P(W \le w) = 1 - (1 - 2w)^2$ for $0 \le w \le 0.5$ and $P(Z \le z) = 1 - (1 - z)^2$ for $0 \le z \le 1$. The derivation of $P(Z \le z)$ uses the relation $P(|X - Y| \le z) = P(X - Y \le z) - P(X - Y \le -z)$ and a little geometry in a triangle. Differentiation of the probability distribution functions gives the desired probability densities.

11.13 The marginal densities are $f_X(x) = 4xe^{-2x}$ for $x > 0$ and $f_Y(y) = 2e^{-2y}$ for $y > 0$.

11.15 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.

11.17 The density is $-\ln(z)$ for $0 < z < 1$ and the expected value is $\frac{1}{4}$.

11.19 The range of Z is $(2c, 2d)$. The random variable 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.

11.21 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 $\frac{1}{2}e^{-\frac{1}{2}(v^2+w^2)}$.

11.23 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 $\rho(X, Y) = 0.3273$.

11.25 $E(\text{annual rainfall}) = 779.5$ millimeters and $\sigma(\text{annual rainfall}) = 121.4$ millimeters. The desired probability is $1 - \Phi(\frac{1{,}000-799.5}{121.4}) = 0.035$.

11.27 This follows from 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 $j \ne i$.

11.29 $\rho(X, Y) = \frac{441/36-(91/36)(161/36)}{(1.40408)^2} = 0.479$.

11.31 $\rho(X, Y) = (1717 - 50.5 \times 25.75)/\sqrt{833.25 \times 490.1875} = 0.652$.

11.33 $\rho(X, Y) = (\frac{1}{5} - \frac{1}{4} \times \frac{3}{4})/\sqrt{\frac{3}{80} \times \frac{3}{80}} = \frac{1}{3}$.

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

$$P(W \le w) = \frac{1/b}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} dx \left[\int_{-\infty}^{(w-ax)/b} e^{-\frac{1}{2}(x^2-2\rho xy+y^2)/(1-\rho^2)}\, dy \right]$$

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 $P(Z \le z) = \int_0^{\infty} dy \int_{-\infty}^{yz} f(x, y)\, dx + \int_{-\infty}^{0} dy \int_{yz}^{\infty} f(x, y)\, dx$. 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 10.13, 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.

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 parameter of the hypothesized Poisson distribution is estimated as $\lambda = 2.25$. The games with five or more goals are aggregated and so six data groups are considered. The test statistic has approximately a chi-square distribution with $6 - 1 - 1 = 4$ degrees of freedom and its value is 1.521. The probability $P(\chi_4^2 > 1.521) = 0.8229$. The Poisson distribution gives an excellent fit.

12.17 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 For any y, $P(X = x \mid Y = y) = (\frac{8}{9})^{x-1}\,\frac{1}{9}$ for $1 \le x < y$ and $P(X = x \mid Y = y) = (\frac{8}{9})^{y-1}(\frac{9}{10})^{x-y-1}\,\frac{1}{10}$ for $x > y$.

13.3 $P(X = x \mid Y = 0) = \frac{7}{5}(\frac{5}{12})^x$ and $P(X = x \mid Y = y) = \binom{x}{y}\frac{7}{12}\frac{(5/12)^x}{(5/7)^y}$ for $y \ge 1$.

13.5 $P(X = x \mid Y = y) = \binom{24}{x}\binom{x}{y}(\frac{1}{6})^x \big/ \sum_{s=y}^{24}\binom{24}{s}\binom{s}{y}(\frac{1}{6})^s$ for $y \le x \le 24$.

13.7 $f_X(x \mid y) = (y+1)^2 x e^{-x(y+1)}$ for $x > 0$ and $f_Y(y \mid x) = x e^{-xy}$ for $y > 0$. The desired probability is $e^{-1} = 0.3679$.

13.9 For any fixed y, the conditional probability mass function $p_X(x \mid y)$ is computed from $p_X(x \mid y) = p_X(x) f_Y(y \mid x) / \sum_x p_X(x) f_Y(y \mid x)$.

13.11 The joint density $f(x, y) = \frac{1}{x}$ for $0 < x < 1$ and $1 - x < y < 1$. The probabilities are 0.1534 and 0.8466.

13.13 Condition on the unloading time. The probability of no breakdown is given by $\int_{-\infty}^{\infty} e^{-\lambda y} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2/\sigma^2}\, dy = e^{-\mu\lambda + \frac{1}{2}\sigma^2\lambda^2}$.

13.15 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.

13.17 The desired probability $P(B^2 \ge 4AC)$ can be calculated as

$$\int_0^1 P\left(AC \le \frac{b^2}{4}\right) db = \int_0^1 db \left[\int_0^1 P\left(C \le \frac{b^2}{4a}\right) da\right]$$

$$= \int_0^1 db \left[\int_0^{\frac{b^2}{4}} 1\, da + \int_{\frac{b^2}{4}}^1 \frac{b^2}{4a}\, da\right] = \int_0^1 db \left[\frac{b^2}{4} - \frac{b^2}{4}\ln\left(\frac{b^2}{4}\right)\right].$$

This leads to $P\left(B^2 \ge 4AC\right) = \frac{5}{36} + \frac{1}{6}\ln(2) = 0.2544$.

13.19 (a) A good choice is the exponential density $g(x) = e^{-x}$. For this choice, $c = \sqrt{2e/\pi} \approx 1.3155$. (b) A good choice is $g(x) = \frac{1}{2}$ for $-1 < x < 1$. For this choice, $c = \frac{4}{\pi} \approx 1.27$.

13.21 The optimal stopping level for player A is the solution to the equation $a^4 = \frac{1}{5}(1 - a^5)$ and is given by $a_3 = 0.64865$. The overall win probability of player A is 0.3052. Let $P(a)$ denote the probability that the final score of player A is no more than a. Then, $P(0) = \frac{1}{2}a_3^2$, $P(a) = P(0) + \frac{1}{2}a^2$ for $0 \le a \le a_3$ and $P(a) = (1 + a_3)a - a_3$ for $a \ge a_3$. The overall win probabilities of the players B and C are 0.3295 and 0.3653.

13.23 $E(X \mid Y = y) = \frac{1}{3}y$ and $E(Y \mid X = x) = \frac{2}{3} + \frac{1}{3}x$, using the conditional densities $f_X(x \mid y) = 2(y - x)/y^2$ and $f_Y(y \mid x) = 2(y - x)/(1 - x)^2$.

13.25 $E(X \mid X + Y = v) = \frac{1}{2}v + \frac{1}{2}(\mu_1 - \mu_2)$.

13.27 $E(X \mid Y = y) = y + 0.5$ and $E(Y \mid X = x) = 0.5x$.

13.29 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 $\frac{2}{3}\mu_1 + \frac{1}{3}\mu_2 - Q$.

13.31 By conditioning 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 $0 \le i \le 11$. Define the constants $\gamma_1 = 5, \gamma_2 = 4, \gamma_3 = 2$ and $\gamma_i = 1$ for $i \ge 4$. Then, the expected value of the stake is equal to $10 + \sum_{i=7}^{11}\alpha_i \times 10 = \$11.81$ and the expected value of the payoff is

equal to $\sum_{i=1}^{6} \alpha_i \times \frac{4i}{50} \times \gamma_i \times 10 + \sum_{i=7}^{11} \alpha_i \times \frac{4i}{50} \times \gamma_i \times 20 + \alpha_0 \times 10 + 13 \times \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times 120 = \$10.93$. The house percentage is 7.45%.

13.33 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.

13.35 The expected value is 42.

13.37 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.

13.39 For fixed n, let $F(i, k)$ be the expected value of the maximal payoff you can still achieve if k tosses are still possible and heads turned up i times so far. The recursion is $F(i, k) = \max\{\frac{1}{2}F(i+1, k-1) + \frac{1}{2}F(i-1, k-1), \frac{i}{n-k}\}$ for $k = 1, \ldots, n$ with $F(i, 0) = \frac{i}{n}$. The maximal expected payoff $F(0, n)$ has the values 0.7679, 0.7780, 0.7834 and 0.7912 for $n = 25$, 50, 100 and 1,000.

13.41 The 95% Bayesian confidence interval is (72.617, 74.095). Letting $n = 10, \sigma^2 = 2$, $\mu_0 = 73$ and $\sigma_0^2 = 0.49$, the posterior density is maximal at $\theta = [n\bar{x}/\sigma^2 + \mu_0/\sigma_0^2]/[n/\sigma^2 + 1/\sigma_0^2] = 73.356$.

13.43 The posterior density is a normal density with an expected value of 143.90 light years and a standard deviation of 15.62 light years. The posterior density is maximal for $\theta = 143.90$ and a 95% Bayesian confidence interval is (113.3, 174.5).

13.45 A gamma density with shape parameter $\alpha + r$ and scale parameter $\lambda + \sum_{i=1}^{r} t_i + (m-r)T$.

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, $E(z^S) = z^0 P(N = 0) + \sum_{n=1}^{\infty} E(z^{X_1 + \cdots + X_n}) P(N = n)$. This gives $E(z^S) = \sum_{n=0}^{\infty} [A(z)]^n e^{-\mu} \frac{\mu^n}{n!} = e^{-\mu[1-A(z)]}$. Differentiating this expression gives the formulas for $E(S)$ and 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 Use the fact that the moment-generating function is $[\lambda/(\lambda - t)]^{\alpha_1+\cdots+\alpha_n}$.

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. By the decomposition formula for the standard bivariate normal density $f(x, y)$ in Section 12.1, $E(e^{vZ_1+wZ_2})$ can be written as

$$\int_{-\infty}^{\infty} e^{vx}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\,dx\int_{-\infty}^{\infty} e^{wy}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{1-\rho^2}}e^{-\frac{1}{2}(y-\rho x)^2/(1-\rho^2)}\,dy.$$

This implies the result $E(e^{vZ_1+wZ_2}) = e^{\frac{1}{2}(v^2+2\rho vw+w^2)}$, using twice the fact that $E(e^{tU}) = e^{\mu t+\frac{1}{2}\sigma^2t^2}$ for an $N(\mu, \sigma^2)$ 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

$$E[e^{vX+wY}] = E[e^{v\sigma_1Z_1+v\mu_1+w\sigma_2Z_2+w\mu_2} = e^{v\mu_1+w\mu_2}E[e^{v\sigma_1Z_1+w\sigma_2Z_2}],$$

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 $E(e^{tU}) = e^{\mu t+\frac{1}{2}\sigma^2t^2}$ 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 $P(X \le c) = P(tX \ge tc) = P(e^{tX} \ge e^{tc})$. Next apply Markov's inequality.

14.19 $[c_1 + (c_2 - c_1)F(T)]/\int_0^T (1 - F(x))\,dx$, where $F(x)$ is the probability distribution function of the lifetime of a bulb.

14.21 $\mu_{on}/(\mu_{on} + \mu_{off}) = 0.673$.

Chapter 15

15.1 Let X_n be 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 $p_{i,i-1} = i^2/r^2$, $p_{ii} = 2i(r - i)/r^2$, $p_{i,i+1} = (r - i)^2/r^2$, 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 $1 \leq i \leq K$, $p_{K+1,K} = 1 - f$, $p_{ij} = (1 - f)a_{j-i+1} + f a_{j-i}$ 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 $1 - \Phi\left(\frac{240.5-217.294}{12.101}\right) = 0.0276$ (the simulated value is 0.0267).

15.15 The probabilities are 0.4584 and 0.2499. The expected number of tosses until a run of three heads appears is 14.

15.17 The probabilities are 0.9288, 0.3395, 0.0648, 0.0105, and 0.0016 for $n = 50$, 75, 100, 125 and 150. The expected value is 71.4 weeks.

15.19 (a) 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 n spins of a European roulette 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 0.0368 for $n = 5{,}000{,}000$ (and 0.0723 for $n = 10{,}000{,}000$). (b) The probability that in 5 million spins of a American roulette wheel any of the numbers $1, \ldots, 36$ will come up six or more times in a row is 0.0565 (the probability is 0.1099 for 10 million spins).

15.21 (a) The expected value is 83.20 and the probability is 0.7054. (b) The expected value is 46,659 and the probability is 0.00203.

15.23 The probabilities are 0.2857, 0.2381, 0.1905, 0.1429, 0.0952, and 0.0476 for the final scores 101, 102, 103, 104, 105, and 106.

15.25 The probability is 0.1123.

15.27 The probabilities are 0.7574 and 0.1436.

15.29 The probabilities are 0.2692 and 0.3836.

15.31 The answers are $\frac{1}{3}$ and 0.4286.

15.33 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.

15.35 (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.

15.37 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 $0 \le i \le j \le 6$ 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 $0 \le i < j \le 5$, $p_{(i,j),(i+1,j+1)} = (1 - r_i)(1 - r_j)$, $p_{(i,j),(0,i+1)} = (1 - r_i)r_j$, $p_{(i,j),(0,j+1)} = r_i(1 - r_j)$, $p_{(i,j),(0,0)} = r_i r_j$, 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.

15.39 Since $\pi_i p_{ij} = \pi_j p_{ji}$ boils down to $\pi_i / \sum_k w_{ik} = \pi_j / \sum_k w_{jk}$, it follows that $\pi_i = \sum_k w_{ik} / \sum_j \sum_k w_{jk}$ satisfies the reversibility equations. By this result, the equilibrium probabilities for the mouse problem are $\pi_1 = \pi_5 = \pi_{11} = \pi_{15} = \frac{2}{44}$, $\pi_2 = \pi_3 = \pi_4 = \pi_6 = \pi_{10} = \pi_{12} = \pi_{13} = \pi_{14} = \frac{3}{44}$, and $\pi_7 = \pi_8 = \pi_9 = \frac{4}{44}$ (take $w_{ij} = 1$ if the rooms i and j are connected by a door). The mean recurrence time from state i to itself is $\mu_{ii} = 1/\pi_i$, see Remark 15.1.

15.43 The univariate conditional densities $\pi_1(x_1 \mid x_2)$ and $\pi_1(x_2 \mid x_1)$ are given by the $N(7(1 + x_2^2)^{-1}/2,\ (1 + x_2^2)^{-1})$ density and the $N(7(1 + x_1^2)^{-1}/2,\ (1 + x_1^2)^{-1})$ density. The estimates 1.6495 and 1.4285 are found for $E(X_1)$ and $\sigma(X_1)$ after one million runs (the exact values are 1.6488 and 1.4294).

Chapter 16

16.1 Let state i mean that i units are in working condition, where $i = 0, 1, 2$. The state-process $\{X(t)\}$ is a continuous-time Markov chain with transition rates $q_{01} = 2\mu$, $q_{10} = \lambda$, $q_{12} = \mu$, and $q_{21} = \lambda + \eta$.

16.3 Let $X_1(t)$ denote the number of infected individuals at time t and let $X_2(t)$ denote the number of susceptible individuals at time t. The continuous-time Markov chain $\{(X_1(t),\ X_2(t))\}$ has transition rates $q_{(i,j),(i+1,j-1)} = ij\mu$, $q_{(i,j),(i,j+1)} = (m - i - j)\eta$ and $q_{(i,j),(i-1,j)} = i\rho$.

16.5 Let state $(i, 0)$ mean that i passengers are waiting at the stand and no sheroot is present $(0 \le i \le 7)$, and let state $(i, 1)$ mean that i passengers are waiting at the stand and a sheroot is present $(0 \le i \le 6)$. The state-process $\{X(t)\}$ is a continuous-time Markov chain with transition rates $q_{(i,0),(i+1,0)} = \lambda$ and $q_{(i,0),(i,1)} = \mu$ for $0 \le i \le 6$, $q_{(7,0),(0,0)} = \mu$, $q_{(i,1),(i+1,1)} = \lambda$ for $0 \le i \le 5$ and $q_{(6,1),(0,0)} = \lambda$.

16.7 $\mu_i = 1/\nu_i + \sum_{j \neq i,r}(q_{ij}/\nu_i)\mu_j$ for $i \neq r$.

16.9 A continuous-time Markov chain with nine states applies. Using a numerical code for linear differential equations, the values 0.000504 and 0.3901 are found for the desired probabilities.

16.11 There are five states. Let state (0, 0) mean that both stations are free, state (0, 1) that station 1 is free and station 2 is busy, state (1, 0) that station 1 is busy and station 2 is free, state (1, 1) that both stations are busy and state $(b, 1)$ that station 1 is blocked and station 2 is busy. The balance equations are given by $\lambda p(0,0) = \mu_2 p(0,1)$, $(\mu_2 + \lambda)p(0,1) = \mu_1 p(1,0) + \mu_2 p(b,1)$, $\mu_1 p(1,0) = \lambda p(0,0) + \mu_2 p(1,1)$, $(\mu_1 + \mu_2)p(1,1) = \lambda p(0,1)$ and $\mu_2 p(b,1) = \mu_1 p(1,1)$. These equations together with the normalization equation are easily solved (one of the balance equations can be omitted). The long-run fraction of time station 1 is blocked is equal to $p(b, 1)$ and the long-run fraction of items that are rejected is equal to $p(1,0) + p(1,1) + p(b,1)$.

16.13 The balance equations are given by $(\lambda + \mu)p(0,0) = \mu p(7,0) + \lambda p(6,1)$, $(\lambda + \mu)p(i,0) = \lambda p(i-1,0)$ for $1 \leq i \leq 6$, $\mu p(7,0) = \lambda p(6,0)$, $\lambda p(0,1) = \mu p(0,0)$, $\lambda p(0,1) = \mu p(0,0)$ and $\lambda p(i,1) = \mu p(i,0) + \lambda p(i-1,1)$ for $1 \leq i \leq 5$. The long-run average number of waiting passengers is $\sum_{i=1}^{6} i[p(i,0) + p(i,1)] + 7p(7,0)$ and the long-run fraction of potential passengers who go elsewhere is $p(7, 0)$.

16.15 A capacity for 18 containers, where the loss probability is 0.0071. By the insensitivity property of the Erlang loss model, the answer is the same when the holding time of a customer has a uniform distribution with the same expected value of 10 hours.

16.17 This inventory model is a special case of the Erlang loss model. Identify the number of outstanding orders with the number of busy servers. The limiting distribution of the stock on hand is given by $r_j = \gamma(\lambda L)^{S-j}/(S-j)!$ for $0 \leq j \leq S$, where $\gamma = 1/\sum_{k=0}^{S}(\lambda L)^k/k!$. The average stock on hand is $\sum_{j=0}^{S} j r_j$ and the fraction of lost demand is r_S.

16.19 The recursion is $j\mu p_j = \lambda p_{j-1}$ for $1 \leq j \leq s$ and $(s\mu + (j-s)\theta)p_j = \lambda p_{j-1}$ for $j > s$. The fraction of balking callers is $\sum_{j=s+1}^{\infty}(j-s)\theta p_j/\lambda$.

16.21 A Poisson distribution with an expected value of λ/μ. The limiting distribution is insensitive to the shape of the service-time distribution and requires only the expected value of the service time. The probability that more than seven oil tankers are under way to Rotterdam is 0.0511.

Bibliography

J. Albert. "Teaching Bayes' rule: a data-oriented approach." *The American Statistician* **51** (1997): 247–253.

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.

J.O. Berger and D.A. Berry. "Statistical analysis and the illusion of objectivity." *American Scientist* **76** (1988): 159–165.

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. "Bayesian clinical trials." *Nature Reviews Drug Discovery* **5** (2006): 27–36.

S. Chib and E. Greenberg. "Understanding the Metropolis-Hastings algorithm." *The American Statistician* **49** (1995): 327–335.

P.R. Coe and W. Butterworth, "Optimal stopping in the showcase showdown." *The American Statistician* **49** (1995): 271–275.

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." *The 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.

T.P. Hill. "Knowing when to stop." *American Scientist* **97** (2007): 126–133.

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** (July 29, 2000): 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.
T. Neller and C.G.M. Presser. "Optimal play of the dice game Pig." *The UMAP Journal* **25** (2004): 25–47.
R.B. Nelsen and J.E. Schultz. "The probability that the sum of rounds equals the round of the sum," *The College Mathematics Journal* **18** (1987): 390–396.
M. Orkin and R. Kakigi. "What is the worth of free casino credit?" *The American Mathematical Monthly* **102** (1995): 3–8.
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.
I. Peterson. "Puzzling names in boxes." *Science News* (August 16, 2006).
S. Savage. "The flaw of averages." *San Jose Mercury News* (October 8, 2000).
M.F. Schilling. "The longest run of heads." *The College Mathematics Journal* **21** (1990): 196–207.
G. Székely and D. Richards. "The St. Petersburg paradox and the crash of high-tech stocks in 2000." *The American Statistician* **58** (2004): 225–231.
L.A. Stefanski. "The North Carolina lottery coincidence." *The American Statistician* **62** (2008): 130–134.
R.H. Thaler. *The Winner's Curse, Paradoxes and Anomalies in Economic Life*. Princeton NJ: Princeton University Press, 1992.
E.O. Thorp. "The Kelly criterion in blackjack, sports betting, and the stock market," revised version 1998, www.bjmath.com.
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): 65–70.
W.A. Whitworth. *Choice and Chance*. Cambridge: Deighton Bell, 3rd edition, 1886.

Index

Abbott, D., 73
absorbing state, 472
acceptance-rejection method, 416, 492
accessibility, 472
addition rule, 33, 246
Albert, J., 281
Aldous, D.J., 57
aperiodicity, 484
arc-sine law, 27
array method, 61
axioms of probability, 33, 231

Bachelier, L., 187
Banach match problem, 316
Bayes' rule, 220, 270, 271, 428
Bayes, T., 7, 220
Bayesian statistics, 7, 204, 278, 428
Benford, F., 200
Benford's law, 200
Bennett, D.J., 53
Berger, J.O., 205
Bernoulli, D., 41, 47
Bernoulli, J., 2, 22, 108
Bernoulli distribution, 303
Bernoulli experiment, 108, 304
Berry, D.A., 205, 279
best-choice problem, 14, 37, 62, 69, 95, 105
beta density, 342
Big–Martingale system, 45
binomial coefficient, 534
binomial distribution, 110, 304, 438
birth-and-death process, 526
birthday problem, 11, 16, 76, 103, 116, 248, 439
bivariate normal distribution, 382, 448
Black, F., 196
Black–Scholes formula, 194
Boole's inequality, 238
bootstrap method, 101, 398
Borel–Cantelli lemma, 239
Box–Muller method, 377
boy–girl problem, 277
branching process, 441
Brown, R., 186
Brownian motion process, 186
  gambler's ruin formula, 190
Buffon, G., 209
Buffon's needle problem, 209, 234
Butterworth, W., 70

Cantor, G., 231
card shuffle, 56
Cardano, G., 2
Carson, J., 78
Cauchy density, 385, 443
Caulkins, J.P., 273
central limit theorem, 163, 166, 345, 454
  multidimensional, 393
chance tree, 215
Chapman–Kolmogorov equations, 468
Chebyshev's inequality, 155, 449
Chernoff bound, 448
chi-square density, 351, 372, 447
chi-square test, 399
Chib, S., 500
Chu, S., 121
Chuck-a-Luck, 304
closed set, 481
Coe, P.R., 70
coefficient of variation, 159
coin-tossing, 22, 70, 164, 197, 269, 426

coincidences, 13, 81, 136
common random numbers, 54
complement rule, 77, 245
compound experiment, 31, 239
compound Poisson distribution, 438
conditional expectation, 418
  law of, 418
conditional probability, 87, 257
  law of, 87, 266, 412
conditional probability density, 407
conditional probability mass function, 405
confidence interval, 173, 353
  Bayesian, 432
continuity property, 238
continuous random variable, 146, 320
continuous-time Markov chain, 507
  equilibrium equations, 521
  first-passage time probabilities, 518
  irreducible, 521
  limiting probabilities, 521
  Markovian property, 508
  time-dependent probabilities, 516
  transition rates, 508
convergence with probability one, 21, 242, 450
convolution formula, 302, 365, 437
correlation coefficient, 157, 379
countably infinite, 231
counting principle, 245
coupon collector's problem, 84, 104, 120, 128, 255, 439, 472, 479
covariance, 156, 378
craps, 83, 86, 104, 128, 476

D'Alembert system, 45
De Méré, C., 2, 315
De Moivre, A., 6, 145, 164, 346, 435
Diaconis, P., 57, 81
discrete random variable, 30, 284
discrete-time Markov chain, 461
  equilibrium equations, 485
  irreducible, 494
  limiting probabilities, 481
  Markovian property, 461
  time-dependent probabilities, 468
disjoint sets, 33, 232
doubling strategy, 90
Doyle, P., 57
draft lottery, 55, 98
drunkard's walk, 38, 71, 181, 301, 386, 460
  in dimension three, 40, 390
  in dimension two, 40, 386
dynamic programming, 424

Efron, B., 101, 277
Efron's dice game, 29
Ehrenfest model, 463, 486
equilibrium distribution, 483
equilibrium equations, 485, 521
equilibrium probabilities, 485
Erlang delay formula, 528
Erlang delay model, 527
Erlang density, 338, 341, 447
Erlang loss formula, 525
Erlang loss model, 525
Euler, L., 435
Euler's constant, 85
event, 28, 32, 232
expectation, *see* expected value
expected value, 34, 286
  continuous random variable, 148, 326
  discrete random variable, 34, 286
  of sum, 290
exponential density, 335, 370, 510
exponential function, 536
extinction probability, 442

failure rate, 510
failure rate function, 357
Fermat, P., 2, 102
first moment, *see* expected value
first-step analysis, 441
flow-rate-equation method, 522
franc-carreau, 236

Galilei, G., 2
gambler's fallacy, 18, 24
gambler's ruin problem, 93, 190, 262
game of Pig, 73, 427, 481
gamma density, 341, 446
gamma function, 341
Gardner, M., 38, 214
Gauss, C.F., 145
Gaussian distribution, 344
generating function, 436
geometric Brownian motion, 193
geometric distribution, 85, 312
geometric probability, 234
geometric series, 537
Gibbs sampler, 502
Gigerenzer, G., 222

Gosset, W., 352
Greenberg, E., 500

Hanley, J.A., 114
Harmer, G.P., 73
hat-check problem, 118, 139
hazard, the game of, 104
Henze, N., 135
Hill, T.P., 97, 201
histogram, 144
hit-or-miss method, 64, 491
house percentage, 90
hundred prisoners, 74, 250
Huygens, C., 2, 22, 34, 93
hypergeometric distribution, 129, 309, 380

importance sampling, 493
inclusion-exclusion rule, 251
independent events, 157, 265
independent random variables, 157, 299
indicator random variable, 291
insensitivity, 525
intersection of events, 238, 243
inverse-transformation method, 354
irreducible, 494

Jacobian, 375
Jeu de Treize, 107, 118
joint probability density, 363
joint probability mass function, 360

Kadell, D., 402
Kahneman, D., 17, 224
Kakigi, R., 94
Kelly, J.L., Jr., 47
Kelly betting, 47, 182, 194
Kelly betting fraction, 49, 66
Kolmogorov, A.N., 22
Kolmogorov differential equations, 517

Laplace, P.S., 2, 146, 164
Laplace density, 447
Laplace model, 32
law of large numbers, 20, 21, 35, 242, 450
law of the iterated logarithm, 456
Lazzarini, M., 209
Leibniz, G., 4
linear predictor, 380
Little's formula, 528
lognormal distribution, 349, 446
lost boarding pass problem, 59, 72, 269
Lotto, 114, 132, 202, 398
Lyapunov, A., 164

Makholm, H., 72
Maor, E., 536
marginal density, 368
Markov chain, 459
Markov chain Monte Carlo, 490
Markov's inequality, 448
Markovian property, 461
Marsaglia's polar method, 377
matching problems, 118, 139, 251
Matthews, R., 209
McKean, K., 17
mean, *see* expected value
median, 313, 332, 349
memoryless property, 123, 336
Mendel, G., 401
Merton, R., 196
Merz, J.F., 273
Metropolis, N., 64
Metropolis–Hastings algorithm, 497
Miltersen, P.B., 74
mixed random variable, 292
moment-generating function, 444
Monte Carlo simulation, 75
Montmort, M., 118
Monty Hall dilemma, 15, 213, 223
Morrison, K.E., 202
Mosteller, F., 81
multinomial distribution, 317
multiplication game, 105, 202
multiplication rule, 260
multivariate normal distribution, 391
  simulating from, 392
mutually exclusive events, 232

Nauss, J.I., 80
negative binomial distribution, 313, 438
Neller, T., 427
Nelsen, R.B., 347
New Age Solitaire, 57
Newcomb, S., 199
Newton, I., 137
normal curve, 145
normal distribution, 148, 344, 370, 445
  bivariate, 382
  multivariate, 391
  simulating from, 375

$o(h)$, 509
odds, 270
odds form of Bayes' rule, 270
optimal stopping, 95, 424, 458
Orkin, M., 94

p-value, 278, 429
Pareto density, 332
Parrondo's paradox, 73
Pascal, B., 2, 44, 93, 102
Paulos, J.A., 78, 183
payoff odds, 47
Pearson, K., 400
Pepy's problem, 137
percentile, 150, 349
permutation, 532
Peterson, I., 74
Pickover, C., 218
Poisson, S.D., 6, 111
Poisson distribution, 112, 306, 438
Poisson process, 120, 339, 511
  simulation of, 126
polar coordinates, 371
Pollaczek–Khintchine formula, 160
positive semi-definite, 380
posterior probability, 221, 270, 278, 428
Powerball Lottery, 132
Presser, C.G.M., 427
prior probability, 221, 270, 278, 428
prisoner's problem, 216
probability density function, 146, 320
probability distribution function, 320
probability mass function, 30, 284
probability measure, 28, 32, 231
probability space, 28, 33, 232
product rule, 31, 239

Quetelet, A., 146
queues, 160, 525, 527, 530

random number, 52
random permutation, 62, 250, 293
  cycle, 250, 293
random sum, 438
random variable, 30, 283
random walk, 24, 38, 106, 178, 386, 390
random-number generator, 52
rare event, 337
rate, 508
rate of return, 152
Rayleigh density, 340
Rayleigh distribution, 140, 390
record value, 438
recurrent state, 487
regression curve, 420
regression to the mean, 381, 422
reliability, 337, 515
renewal-reward theorem, 452
reversible Markov chain, 494
Richards, D., 42
Riedwyl, H., 135
roulette, 44, 90, 180, 480
runs, 28, 71, 197, 440, 474

Salk vaccine, 168
Samford, M., 220
sample mean, 171
sample space, 20, 28, 230
sample variance, 172
Savage, S., 152
Schilling, M.F., 71
Scholes, M., 196
Schultz, J.E., 347
scratch-and-win lottery, 12, 117
Shackleford, M., 92
Simpson, O.J., 273
simulation, 57, 75, 171, 490
  accuracy, 63, 176
Slovic, P., 17
square-root law, 161, 301
St. Petersburg paradox, 41
standard Brownian motion, 188
standard deviation, 155, 294
standard normal distribution, 149, 196, 344
statistical equilibrium, 483
Stefanski, L.A., 79
Steutel, F., 437
stochastic process, 452
Student-t density, 351
subjective probability, 205
substitution rule, 293, 372
Székely, G, 42

tabu probability, 478
test paradox, 218
Thaler, R.H., 27
Thorp, E.O., 51, 186
transformation, 353, 374

transient state, 487
transition probabilities, 463, 468
transition rate diagram, 522
triangular density, 334
Tversky, A., 17
two-up, game of, 250

Ulam, S., 64
uncountable, 231
unders and overs, game of, 36
uniform density, 333
uniform distribution, 53, 312
uniformization method, 520
union of events, 33, 232, 243

variance, 154, 294, 328
Von Bortkiewicz, L., 113, 402
vos Savant, M., 118, 214

waiting-time paradox, 159
Weibull density, 342
Weiss, G.H., 39
Whitworth, W.A., 47
Wiener, N., 188

Ylvisaker, D., 402

z-value, 168
Zarin case, 93