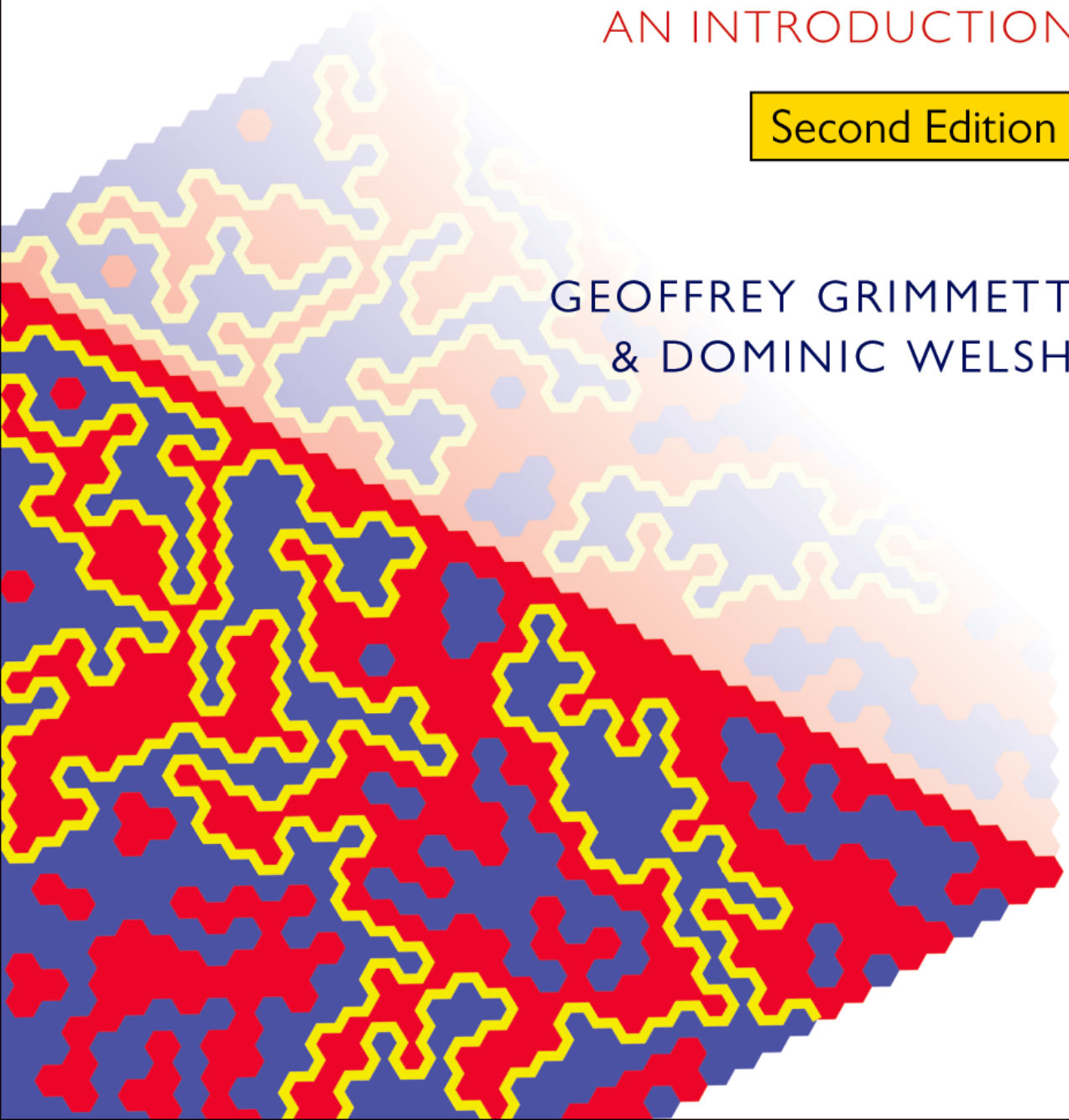
OXFORD

Probability

AN INTRODUCTION

Second Edition



GEOFFREY GRIMMETT
& DOMINIC WELSH

Probability

An Introduction

Probability

An Introduction

Second Edition

Geoffrey Grimmett
University of Cambridge

Dominic Welsh
University of Oxford

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 1986
Second Edition published in 2014

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2014938231

ISBN 978–0–19–870996–1 (hbk.)
ISBN 978–0–19–870997–8 (pbk.)

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Preface to the second edition

Probability theory is now fully established as a crossroads discipline in mathematical science. Its connections across pure and applied mathematics are in excellent health. In its role as the 'science of risk', it is a lynchpin of political, economic, and social science. Never before have so many students in schools and universities been exposed to the basics of probability. This introductory text on probability is designed for first and second year mathematics students. It is based upon courses given at the Universities of Bristol, Cambridge, and Oxford.

Broadly speaking, we cover the usual material, but we hope that our account will have certain special attractions for the reader and we shall say what these may be in a moment. The first eight chapters form a course in basic probability, being an account of events, random variables, and distributions—we treat discrete and continuous random variables separately—together with simple versions of the law of large numbers and the central limit theorem. There is an account of moment generating functions and their applications. The next three chapters are about branching processes, random walks, and continuous-time random processes such as the Poisson process. We hope that these chapters are adequate at this level and are suitable appetizers for further courses in applied probability and random processes. The final chapter is devoted to Markov chains in discrete time.

As in the first edition, this text is divided into three sections: (A) Probability, (B) Further Probability, and (C) Random Processes. We hope thus to indicate two things. First, the probability in Part A seems to us to be core material for first-year students, whereas the material in Part B is somewhat more difficult. Secondly, although random processes are collected together in the final four chapters, they may well be introduced much earlier in the course. The chapters on branching processes and random walks might be studied after Chapter 5, and the chapter on continuous-time processes after Chapter 6. The chapter on Markov chains can in addition be used as the basis for a free-standing 12–16 lecture course.

The major difference of substance between the first and second editions of this text is the new Chapter 12 on Markov chains. This chapter is a self-contained account of discrete-time chains, culminating in a proper account of the convergence theorem. Numerous lesser changes and additions have been made to the text in response to the evolution of course syllabuses and of our perceptions of the needs of readers. These include more explicit accounts of geometrical probability, indicator functions, the Markov and Jensen inequalities, the multivariate normal distribution, and Cramér's large deviation theorem, together with further exercises and problems often taken from recent examination papers at our home universities.

We have two major aims: to be concise, and to be honest about mathematical rigour. Some will say that this book reads like a set of lecture notes. We would not regard this as entirely unfair; indeed a principal reason for writing it was that we believe that most students benefit more from possessing a compact account of the subject in 250 printed pages or so (at a suitable price) than a diffuse account of 400 or more pages. Most undergraduates learn probability

theory by attending lectures, at which they may take copious and occasionally incorrect notes; they may also attend tutorials and classes. Few are they who learn probability in private by relying on a textbook as the sole or principal source of inspiration and learning. Although some will say that this book is quite difficult, it is the case that first-year students at many universities learn some quite difficult things, such as axiomatic systems in algebra and ϵ/δ analysis, and we doubt if the material covered here is inherently more challenging than these. Also, lecturers and tutors have certain advantages over authors—they have the power to hear and speak to their audiences—and these advantages should help them explain the harder things to their students.

Here are a few words about our approach to rigour. It is impossible to prove everything with complete rigour at this level. On the other hand, we believe it is important that students should understand why rigour is necessary. We try to be rigorous where possible, and elsewhere we go to some lengths to point out how and where we skate on thin ice.

Most sections finish with a few exercises; these are usually completely routine, and students should do them as a matter of course. Each chapter finishes with a collection of problems; these are often much harder than the exercises, and include numerous questions taken from examination papers set in Cambridge and Oxford. We acknowledge permission from Oxford University Press in this regard, and also from the Faculties of Mathematics at Cambridge and Oxford for further questions added in this second edition. There are two useful appendices, followed by a final section containing some hints for solving the problems. Problems marked with an asterisk may be more difficult.

We hope that the remaining mistakes and misprints are not held against us too much, and that they do not pose overmuch of a hazard to the reader. Only with the kind help of our students have we reduced them to the present level.

Finally, we extend our appreciation to our many students in Bristol, Oxford, and Cambridge for their attention, intelligence, and criticisms during our many years of teaching probability to undergraduates.

GG, DW
Cambridge, Oxford
February 2014

Contents

PART C RANDOM PROCESSES

Part A

Basic Probability

1
Events and probabilities

Summary. The very basic principles and tools of probability theory are set out. An event involving randomness may be described in mathematical terms as a probability space. Following an account of the properties of probability spaces, the concept of conditional probability is explained, and also that of the independence of events. There are many worked examples of calculations of probabilities.

1.1 Experiments with chance

Many actions have outcomes which are largely unpredictable in advance—tossing a coin and throwing a dart are simple examples. Probability theory is about such actions and their consequences. The mathematical theory starts with the idea of an *experiment* (or *trial*), being a course of action whose consequence is not predetermined. This experiment is reformulated as a mathematical object called a *probability space*. In broad terms, the probability space corresponding to a given experiment comprises three items:

(i) the set of all possible outcomes of the experiment,
(ii) a list of all the events which may possibly occur as consequences of the experiment,
(iii) an assessment of the likelihoods of these events.

For example, if the experiment is the throwing of a fair six-sided die, then the probability space amounts to the following:

(i) the set $\{1, 2, 3, 4, 5, 6\}$ of possible outcomes,
(ii) a list of events such as
- the result is 3,
- the result is at least 4,
- the result is a prime number,

(iii) each number 1, 2, 3, 4, 5, 6 is equally likely to be the result of the throw.

Given any experiment involving chance, there is a corresponding probability space, and the study of such spaces is called *probability theory*. Next, we shall see how to construct such spaces more explicitly.

1.2 Outcomes and events

We use the letter $\mathcal{E}$ to denote a particular experiment whose outcome is not completely predetermined. The first thing which we do is to make a list of all the possible outcomes of $\mathcal{E}$.

The set of all such possible outcomes is called the *sample space* of $\mathcal{E}$ and we usually denote it by Ω. The Greek letter ω denotes a typical member of Ω, and we call each member ω an *elementary event*.

If, for example, $\mathcal{E}$ is the experiment of throwing a fair die once, then

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

There are many questions which we may wish to ask about the actual outcome of this experiment (questions such as 'is the outcome a prime number?'), and all such questions may be rewritten in terms of subsets of Ω (the previous question becomes 'does the outcome lie in the subset $\{2, 3, 5\}$ of Ω?'). The second thing which we do is to make a list of all the events which are interesting to us. This list takes the form of a collection of subsets of Ω, each such subset A representing the event 'the outcome of $\mathcal{E}$ lies in A'. Thus we ask 'which possible events are interesting to us', and then we make a list of the corresponding subsets of Ω. This relationship between *events* and *subsets* is very natural, especially because two or more events combine with each other in just the same way as the corresponding subsets combine. For example, if A and B are subsets of Ω, then

- the set $A \cup B$ corresponds to the event 'either A or B occurs',
- the set $A \cap B$ corresponds to the event 'both A and B occur',
- the complement $A^{\mathrm{c}} := \Omega \setminus A$ corresponds to the event 'A does not occur',[1]

where we say that a subset of C of Ω 'occurs' whenever the outcome of $\mathcal{E}$ lies in C. Thus all set-theoretic statements and combinations may be interpreted in terms of events. For example, the formula

$$\Omega \setminus (A \cap B) = (\Omega \setminus A) \cup (\Omega \setminus B)$$

may be read as 'if A and B do not both occur, then either A does not occur or B does not occur'. In a similar way, if $A_1, A_2, \ldots$ are events, then the sets $\bigcup_{i=1}^{\infty} A_i$ and $\bigcap_{i=1}^{\infty} A_i$ represent the events 'A_i occurs, for some i' and 'A_i occurs, for every i', respectively.

Thus we write down a collection $\mathcal{F} = \{A_i : i \in I\}$ of subsets of Ω which are interesting to us; each $A \in \mathcal{F}$ is called an *event*. In simple cases, such as the die-throwing example above, we usually take $\mathcal{F}$ to be the set of *all* subsets of Ω (called the *power set* of Ω), but for reasons which may be appreciated later there are many circumstances in which we take $\mathcal{F}$ to be a very much smaller collection than the entire power set.[2] In all cases we demand a certain consistency of $\mathcal{F}$, in the following sense. If $A, B, C, \ldots \in \mathcal{F}$, we may reasonably be interested also in the events 'A does *not* occur' and 'at least one of $A, B, C, \ldots$ occurs'. With this in mind, we require that $\mathcal{F}$ satisfy the following definition.

[1] For any subset A of Ω, the *complement* of A is the set of all members of Ω which are not members of A. We denote the complement of A by either $\Omega \setminus A$ or A^{c}, depending on the context.

[2] This is explained in Footnote 3 on p. 6.

Definition 1.1 *The collection $\mathcal{F}$ of subsets of the sample space Ω is called an* ***event space*** *if*

$$\mathcal{F} \text{ is non-empty}, \tag{1.2}$$

$$\text{if } A \in \mathcal{F} \text{ then } \Omega \setminus A \in \mathcal{F}, \tag{1.3}$$

$$\text{if } A_1, A_2, \ldots \in \mathcal{F} \text{ then } \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}. \tag{1.4}$$

We speak of an event space $\mathcal{F}$ as being 'closed under the operations of taking complements and countable unions'. Here are some elementary consequences of axioms (1.2)–(1.4).

(a) An event space $\mathcal{F}$ must contain the empty set $\varnothing$ and the whole set Ω. This holds as follows. By (1.2), there exists some $A \in \mathcal{F}$. By (1.3), $A^{\mathrm{c}} \in \mathcal{F}$. We set $A_1 = A$, $A_i = A^{\mathrm{c}}$ for $i \geq 2$ in (1.4), and deduce that $\mathcal{F}$ contains the union $\Omega = A \cup A^{\mathrm{c}}$. By (1.3) again, the complement $\Omega \setminus \Omega = \varnothing$ lies in $\mathcal{F}$ also.

(b) An event space is closed under the operation of *finite* unions, as follows. Let $A_1, A_2, \ldots, A_m \in \mathcal{F}$, and set $A_i = \varnothing$ for $i > m$. Then $A := \bigcup_{i=1}^{m} A_i$ satisfies $A = \bigcup_{i=1}^{\infty} A_i$, so that $A \in \mathcal{F}$ by (1.4).

(c) The third condition (1.4) is written in terms of *unions*. An event space is also closed under the operations of taking finite or countable *intersections*. This follows by the elementary formula $(A \cap B)^{\mathrm{c}} = A^{\mathrm{c}} \cup B^{\mathrm{c}}$, and its extension to finite and countable families.

Here are some examples of pairs $(\Omega, \mathcal{F})$ of sample spaces and event spaces.

Example 1.5 Ω is any non-empty set and $\mathcal{F}$ is the power set of Ω. △

Example 1.6 Ω is any non-empty set and $\mathcal{F} = \{\varnothing, A, \Omega \setminus A, \Omega\}$, where A is a given non-trivial subset of Ω. △

Example 1.7 $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\mathcal{F}$ is the collection

$$\varnothing,\ \{1, 2\},\ \{3, 4\},\ \{5, 6\},\ \{1, 2, 3, 4\},\ \{3, 4, 5, 6\},\ \{1, 2, 5, 6\},\ \Omega$$

of subsets of Ω. This event space is unlikely to arise naturally in practice. △

In the following exercises, Ω is a set and $\mathcal{F}$ is an event space of subsets of Ω.

Exercise 1.8 If $A, B \in \mathcal{F}$, show that $A \cap B \in \mathcal{F}$.

Exercise 1.9 The *difference* $A \setminus B$ of two subsets A and B of Ω is the set $A \cap (\Omega \setminus B)$ of all points of Ω which are in A but not in B. Show that if $A, B \in \mathcal{F}$, then $A \setminus B \in \mathcal{F}$.

Exercise 1.10 The *symmetric difference* $A \triangle B$ of two subsets A and B of Ω is defined to be the set of points of Ω which are in either A or B but not both. If $A, B \in \mathcal{F}$, show that $A \triangle B \in \mathcal{F}$.

Exercise 1.11 If $A_1, A_2, \ldots, A_m \in \mathcal{F}$ and k is positive integer, show that the set of points in Ω which belong to exactly k of the A_i belongs to $\mathcal{F}$ (the previous exercise is the case when $m = 2$ and $k = 1$).

Exercise 1.12 Show that, if Ω is a finite set, then $\mathcal{F}$ contains an even number of subsets of Ω.

1.3 Probabilities

From our experiment $\mathcal{E}$, we have so far constructed a sample space Ω and an event space $\mathcal{F}$ associated with $\mathcal{E}$, but there has been no mention yet of probabilities. The third thing which we do is to allocate probabilities to each event in $\mathcal{F}$, writing $\mathbb{P}(A)$ for the probability of the event A. We shall assume that this can be done in such a way that the probability function $\mathbb{P}$ satisfies certain intuitively attractive conditions:

(a) each event A in the event space has a probability $\mathbb{P}(A)$ satisfying $0 \leq \mathbb{P}(A) \leq 1$,
(b) the event Ω, that 'something happens', has probability 1, and the event $\varnothing$, that 'nothing happens', has probability 0,
(c) if A and B are disjoint events (in that $A \cap B = \varnothing$), then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

We collect these conditions into a formal definition as follows.[3]

Definition 1.13 *A mapping $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called a **probability measure** on $(\Omega, \mathcal{F})$ if*

(a) $\mathbb{P}(A) \geq 0$ *for* $A \in \mathcal{F}$,
(b) $\mathbb{P}(\Omega) = 1$ *and* $\mathbb{P}(\varnothing) = 0$,
(c) *if $A_1, A_2, \ldots$ are disjoint events in $\mathcal{F}$ (in that $A_i \cap A_j = \varnothing$ whenever $i \neq j$) then*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{1.14}$$

We emphasize that a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is defined only on those subsets of Ω which lie in $\mathcal{F}$. Here are two notes about probability measures.

(i) The second part of condition (b) is superfluous in the above definition. To see this, define the disjoint events $A_1 = \Omega$, $A_i = \varnothing$ for $i \geq 2$. By condition (c),

$$\mathbb{P}(\Omega) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}(\Omega) + \sum_{i=2}^{\infty} \mathbb{P}(\varnothing).$$

(ii) Condition (c) above is expressed as saying that $\mathbb{P}$ is countably additive. The probability measure $\mathbb{P}$ is also *finitely additive* in that

$$\mathbb{P}\left(\bigcup_{i=1}^{m} A_i\right) = \sum_{i=1}^{m} \mathbb{P}(A_i)$$

for disjoint events A_i. This is deduced from condition (c) by setting $A_i = \varnothing$ for $i > m$.

Condition (1.14) requires that the probability of the union of a countable collection of disjoint sets is the sum of the individual probabilities.[4]

[3] This is where the assumptions of an event space come to the fore. Banach and Kuratowski proved in 1929 that there exists no probability measure $\mathbb{P}$ defined on *all* subsets of the interval $[0, 1]$ satisfying $\mathbb{P}(\{x\}) = 0$ for every singleton $x \in [0, 1]$.

[4] A set S is called *countable* if it may be put in one–one correspondence with a subset of the natural numbers $\{1, 2, 3, \ldots\}$.

Example 1.15 Let Ω be a non-empty set and let A be a proper, non-empty subset of Ω (so that $A \neq \varnothing, \Omega$). If $\mathcal{F}$ is the event space $\{\varnothing, A, \Omega \setminus A, \Omega\}$, then all probability measures on $(\Omega, \mathcal{F})$ have the form

$$\begin{aligned} \mathbb{P}(\varnothing) &= 0, & \mathbb{P}(A) &= p, \\ \mathbb{P}(\Omega \setminus A) &= 1 - p, & \mathbb{P}(\Omega) &= 1, \end{aligned}$$

for some p satisfying $0 \leq p \leq 1$. △

Example 1.16 Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ be a finite set of exactly N points, and let $\mathcal{F}$ be the power set of Ω. It is easy to check that the function $\mathbb{P}$ defined by

$$\mathbb{P}(A) = \frac{1}{N}|A| \qquad \text{for } A \in \mathcal{F}$$

is a probability measure on $(\Omega, \mathcal{F})$.[5] △

Exercise 1.17 Let $p_1, p_2, \ldots, p_N$ be non-negative numbers such that $p_1 + p_2 + \cdots + p_N = 1$, and let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$, with $\mathcal{F}$ the power set of Ω, as in Example 1.16. Show that the function $\mathbb{Q}$ given by

$$\mathbb{Q}(A) = \sum_{i:\, \omega_i \in A} p_i \qquad \text{for } A \in \mathcal{F}$$

is a probability measure on $(\Omega, \mathcal{F})$. Is $\mathbb{Q}$ a probability measure on $(\Omega, \mathcal{F})$ if $\mathcal{F}$ is not the power set of Ω but merely some event space of subsets of Ω?

1.4 Probability spaces

We combine the previous ideas in a formal definition.

Definition 1.18 *A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ of objects such that*

(a) *Ω is a non-empty set,*
(b) *$\mathcal{F}$ is an event space of subsets of Ω,*
(c) *$\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$.*

There are many elementary consequences of the axioms underlying this definition, and we describe some of these. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Property If $A, B \in \mathcal{F}$, then[6] $A \setminus B \in \mathcal{F}$.

Proof The complement of $A \setminus B$ equals $(\Omega \setminus A) \cup B$, which is the union of events and is therefore an event. Hence $A \setminus B$ is an event, by (1.3). □

Property If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

[5] The *cardinality* $|A|$ of a set A is the number of points in A.
[6] $A \setminus B = A \cup (\Omega \setminus B)$ is the set of points in A which are not in B.

Proof The complement of $\bigcap_{i=1}^{\infty} A_i$ equals $\bigcup_{i=1}^{\infty}(\Omega \setminus A_i)$, which is the union of the complements of events and is therefore an event. Hence the intersection of the A_i is an event also, as before. □

Property If $A \in \mathcal{F}$ then $\mathbb{P}(A) + \mathbb{P}(\Omega \setminus A) = 1$.

Proof The events A and $\Omega \setminus A$ are disjoint with union Ω, and so

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(\Omega \setminus A).$$ □

Property If $A, B \in \mathcal{F}$ then $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Proof The set A is the union of the disjoint sets $A \setminus B$ and $A \cap B$, and hence

$$\mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) \qquad \text{by (1.14).}$$

A similar remark holds for the set B, giving that

$$\begin{aligned} \mathbb{P}(A) + \mathbb{P}(B) &= \mathbb{P}(A \setminus B) + 2\mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A) \\ &= \mathbb{P}\big((A \setminus B) \cup (A \cap B) \cup (B \setminus A)\big) + \mathbb{P}(A \cap B) \qquad \text{by (1.14)} \\ &= \mathbb{P}(A \cup B) + \mathbb{P}(A \cap B). \end{aligned}$$ □

Property If $A, B \in \mathcal{F}$ and $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Proof We have that $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$. □

It is often useful to draw a Venn diagram when working with probabilities . For example, to illustrate the formula $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$, we might draw the diagram in Figure 1.1, and note that the probability of $A \cup B$ is the sum of $\mathbb{P}(A)$ and $\mathbb{P}(B)$ minus $\mathbb{P}(A \cap B)$, since the last probability is counted twice in the simple sum $\mathbb{P}(A) + \mathbb{P}(B)$.

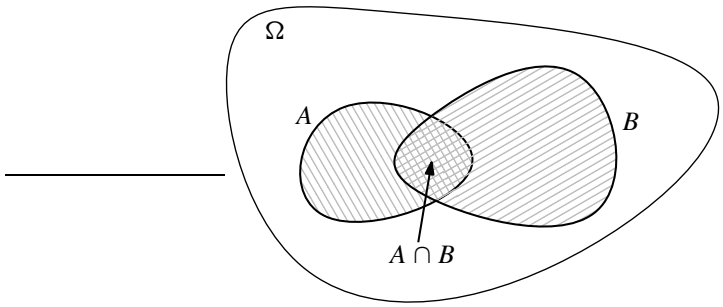


Fig. 1.1 A Venn diagram illustrating the fact that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

In the following exercises, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Exercise 1.19 If $A, B \in \mathcal{F}$, show that $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$.

Exercise 1.20 If $A, B, C \in \mathcal{F}$, show that

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(B \cap C) - \mathbb{P}(A \cap C) + \mathbb{P}(A \cap B \cap C).$$

Exercise 1.21 Let A, B, C be three events such that

$$\begin{aligned} \mathbb{P}(A) &= \tfrac{5}{10}, & \mathbb{P}(B) &= \tfrac{7}{10}, & \mathbb{P}(C) &= \tfrac{6}{10}, \\ \mathbb{P}(A \cap B) &= \tfrac{3}{10}, & \mathbb{P}(B \cap C) &= \tfrac{4}{10}, & \mathbb{P}(A \cap C) &= \tfrac{2}{10}, \\ & & \mathbb{P}(A \cap B \cap C) &= \tfrac{1}{10}. \end{aligned}$$

By drawing a Venn diagram or otherwise, find the probability that exactly two of the events A, B, C occur.

Exercise 1.22 A fair coin is tossed 10 times (so that heads appears with probability $\frac{1}{2}$ at each toss). Describe the appropriate probability space in detail for the two cases when

(a) the outcome of every toss is of interest,

(b) only the total number of tails is of interest.

In the first case your event space should have $2^{2^{10}}$ events, but in the second case it need have only 2^{11} events.

1.5 Discrete sample spaces

Let $\mathcal{E}$ be an experiment with probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The structure of this space depends greatly on whether Ω is a countable set (that is, a finite or countably infinite set) or an uncountable set. If Ω is a countable set, we normally take $\mathcal{F}$ to be the set of *all* subsets of Ω, for the following reason. Suppose that $\Omega = \{\omega_1, \omega_2, \dots\}$ and, for each $\omega \in \Omega$, we are interested in whether or not this given ω is the actual outcome of $\mathcal{E}$; then we require that each singleton set $\{\omega\}$ belongs to $\mathcal{F}$. Let $A \subseteq \Omega$. Then A is countable (since Ω is countable), and so A may be expressed as the union of the countably many ω_i which belong to A, giving that $A = \bigcup_{\omega \in A} \{\omega\} \in \mathcal{F}$ by (1.4). The probability $\mathbb{P}(A)$ of the event A is determined by the collection of probabilities $\mathbb{P}(\{\omega\})$ as ω ranges over Ω, since, by (1.14),

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

We usually write $\mathbb{P}(\omega)$ for the probability $\mathbb{P}(\{\omega\})$ of an event containing only one point in Ω.

Example 1.23 (Equiprobable outcomes) If $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$ and $\mathbb{P}(\omega_i) = \mathbb{P}(\omega_j)$ for all i and j, then $\mathbb{P}(\omega) = 1/N$ for $\omega \in \Omega$, and $\mathbb{P}(A) = |A|/N$ for $A \subseteq \Omega$. △

Example 1.24 (Random integers) There are 'intuitively clear' statements which are without meaning in probability theory, and here is an example: *if we pick a positive integer at random, then it is an even integer with probability* $\frac{1}{2}$. Interpreting 'at random' to mean that each positive integer is equally likely to be picked, then this experiment would have probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where

(a) $\Omega = \{1, 2, \dots\}$,
(b) $\mathcal{F}$ is the set of all subsets of Ω,
(c) if $A \subseteq \Omega$, then $\mathbb{P}(A) = \sum_{i \in A} \mathbb{P}(i) = \pi |A|$, where π is the probability that any given integer, i say, is picked.

However,

$$\text{if } \pi = 0 \text{ then } \mathbb{P}(\Omega) = \sum_{i=1}^{\infty} 0 = 0,$$

$$\text{if } \pi > 0 \text{ then } \mathbb{P}(\Omega) = \sum_{i=1}^{\infty} \pi = \infty,$$

neither of which is in agreement with the rule that $\mathbb{P}(\Omega) = 1$. One possible way of interpreting the italicized statement above is as follows. Let N be a large positive integer, and let $\mathcal{E}_N$ be the experiment of picking an integer from the finite set $\Omega_N = \{1, 2, \dots, N\}$ at random. The probability that the outcome of $\mathcal{E}_N$ is even is

$$\frac{1}{2} \text{ if } N \text{ is even, and } \frac{1}{2}\left(1 - \frac{1}{N}\right) \text{ if } N \text{ is odd,}$$

so that, as $N \to \infty$, the required probability tends to $\frac{1}{2}$. Despite this sensible interpretation of the italicized statement, we emphasize that this statement is without meaning in its present form and should be shunned by serious probabilists. △

The most elementary problems in probability theory are those which involve experiments such as the shuffling of cards or the throwing of dice, and these usually give rise to situations in which every possible outcome is equally likely to occur. This is the case of Example 1.23 above. Such problems usually reduce to the problem of counting the number of ways in which some event may occur, and the following exercises are of this type.

Exercise 1.25 Show that if a coin is tossed n times, then there are exactly

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

sequences of possible outcomes in which exactly k heads are obtained. If the coin is fair (so heads and tails are equally likely on each toss), show that the probability of getting at least k heads is

$$\frac{1}{2^n} \sum_{r=k}^{n} \binom{n}{r}.$$

Exercise 1.26 We distribute r distinguishable balls into n cells at random, multiple occupancy being permitted. Show that

(a) there are n^r possible arrangements,

(b) there are $\binom{r}{k}(n-1)^{r-k}$ arrangements in which the first cell contains exactly k balls,

(c) the probability that the first cell contains exactly k balls is

$$\binom{r}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{r-k}.$$

Exercise 1.27 In a game of bridge, the 52 cards of a conventional pack are distributed at random between the four players in such a way that each player receives 13 cards. Show that the probability that each player receives one ace is

$$\frac{24\cdot 48!\cdot 13^4}{52!} = 0.105\ldots.$$

Exercise 1.28 Show that the probability that two given hands in bridge contain k aces between them is

$$\binom{4}{k}\binom{48}{26-k}\Big/\binom{52}{26}.$$

Exercise 1.29 Show that the probability that a hand in bridge contains 6 spades, 3 hearts, 2 diamonds and 2 clubs is

$$\binom{13}{6}\binom{13}{3}\binom{13}{2}^2\Big/\binom{52}{13}.$$

Exercise 1.30 Which of the following is more probable:

(a) getting at least one six with 4 throws of a die,

(b) getting at least one double six with 24 throws of two dice?

This is sometimes called 'de Méré's paradox', after the professional gambler Chevalier de Méré, who believed these two events to have equal probability.

1.6 Conditional probabilities

Let $\mathcal{E}$ be an experiment with probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We may sometimes possess some incomplete information about the actual outcome of $\mathcal{E}$ without knowing this outcome entirely. For example, if we throw a fair die and a friend tells us that an even number is showing, then this information affects our calculations of probabilities. In general, if A and B are events (that is, $A, B \in \mathcal{F}$) and we are given that B occurs, then, in the light of this information, the new probability of A may no longer be $\mathbb{P}(A)$. Clearly, in this new circumstance, A occurs if and only if $A \cap B$ occurs, suggesting that the new probability of A should be proportional to $\mathbb{P}(A \cap B)$. We make this chat more formal in a definition.[7]

Definition 1.31 *If $A, B \in \mathcal{F}$ and $\mathbb{P}(B) > 0$, the **(conditional) probability of** A **given** B is denoted by $\mathbb{P}(A \mid B)$ and defined by*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1.32}$$

[7] We emphasize that this is a definition rather than a theorem.

Note that the constant of proportionality in (1.32) has been chosen so that the probability $\mathbb{P}(B \mid B)$, that B occurs given that B occurs, satisfies $\mathbb{P}(B \mid B) = 1$. We must next check that this definition gives rise to a probability space.

Theorem 1.33 *If $B \in \mathcal{F}$ and $\mathbb{P}(B) > 0$ then $(\Omega, \mathcal{F}, \mathbb{Q})$ is a probability space where $\mathbb{Q} : \mathcal{F} \to \mathbb{R}$ is defined by $\mathbb{Q}(A) = \mathbb{P}(A \mid B)$.*

Proof We need only check that $\mathbb{Q}$ is a probability measure on $(\Omega, \mathcal{F})$. Certainly $\mathbb{Q}(A) \geq 0$ for $A \in \mathcal{F}$ and

$$\mathbb{Q}(\Omega) = \mathbb{P}(\Omega \mid B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = 1,$$

and it remains to check that $\mathbb{Q}$ satisfies (1.14). Suppose that $A_1, A_2, \ldots$ are disjoint events in $\mathcal{F}$. Then

$$\begin{aligned}
\mathbb{Q}\left(\bigcup_i A_i\right) &= \frac{1}{\mathbb{P}(B)}\mathbb{P}\left(\left(\bigcup_i A_i\right) \cap B\right) \\
&= \frac{1}{\mathbb{P}(B)}\mathbb{P}\left(\bigcup_i (A_i \cap B)\right) \\
&= \frac{1}{\mathbb{P}(B)} \sum_i \mathbb{P}(A_i \cup B) \qquad \text{since } \mathbb{P} \text{ satisfies (1.14)} \\
&= \sum_i \mathbb{Q}(A_i).
\end{aligned}$$

□

Exercise 1.34 If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and A, B, C are events, show that

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A \mid B \cap C)\mathbb{P}(B \mid C)\mathbb{P}(C)$$

so long as $\mathbb{P}(B \cap C) > 0$.

Exercise 1.35 Show that

$$\mathbb{P}(B \mid A) = \mathbb{P}(A \mid B)\frac{\mathbb{P}(B)}{\mathbb{P}(A)}$$

if $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$.

Exercise 1.36 Consider the experiment of tossing a fair coin 7 times. Find the probability of getting a prime number of heads given that heads occurs on at least 6 of the tosses.

1.7 Independent events

We call two events A and B 'independent' if the occurrence of one of them does not affect the probability that the other occurs. More formally, this requires that, if $\mathbb{P}(A), \mathbb{P}(B) > 0$, then

$$\mathbb{P}(A \mid B) = \mathbb{P}(A) \quad \text{and} \quad \mathbb{P}(B \mid A) = \mathbb{P}(B). \tag{1.37}$$

Writing $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$, we see that the following definition is appropriate.

Definition 1.38 *Events A and B of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are called* ***independent*** *if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \tag{1.39}$$

and ***dependent*** *otherwise.*

This definition is slightly more general than (1.37) since it allows the events A and B to have zero probability. It is easily generalized as follows to more than two events. A family $\mathcal{A} = (A_i : i \in I)$ of events is called *independent* if, for all *finite* subsets J of I,

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i). \tag{1.40}$$

The family $\mathcal{A}$ is called *pairwise independent* if (1.40) holds whenever $|J| = 2$.

Thus, three events, A, B, C, are independent if and only if the following equalities hold:

$$\begin{aligned} \mathbb{P}(A \cap B \cap C) &= \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C), & \mathbb{P}(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B), \\ \mathbb{P}(A \cap C) &= \mathbb{P}(A)\mathbb{P}(C), & \mathbb{P}(B \cap C) &= \mathbb{P}(B)\mathbb{P}(C). \end{aligned}$$

There are families of events which are pairwise independent but not independent.

Example 1.41 Suppose that we throw a fair four-sided die (you may think of this as a square die thrown in a two-dimensional universe). We may take $\Omega = \{1, 2, 3, 4\}$, where each $\omega \in \Omega$ is equally likely to occur. The events $A = \{1, 2\}$, $B = \{1, 3\}$, $C = \{1, 4\}$ are pairwise independent but not independent. △

Exercise 1.42 Let A and B be events satisfying $\mathbb{P}(A)$, $\mathbb{P}(B) > 0$, and such that $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Show that $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.

Exercise 1.43 If A and B are events which are disjoint and independent, what can be said about the probabilities of A and B?

Exercise 1.44 Show that events A and B are independent if and only if A and $\Omega \setminus B$ are independent.

Exercise 1.45 Show that events $A_1, A_2, \ldots, A_m$ are independent if and only if $\Omega \setminus A_1$, $\Omega \setminus A_2, \ldots,$ $\Omega \setminus A_m$ are independent.

Exercise 1.46 If $A_1, A_2, \ldots, A_m$ are independent and $\mathbb{P}(A_i) = p$ for $i = 1, 2, \ldots, m$, find the probability that

(a) none of the A_i occur,

(b) an even number of the A_i occur.

Exercise 1.47 On your desk, there is a very special die with a prime number p of faces, and you throw this die once. Show that no two events A and B can be independent unless either A or B is the whole sample space or the empty set.

1.8 The partition theorem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *partition* of Ω is a collection $\{B_i : i \in I\}$ of disjoint events (in that $B_i \in \mathcal{F}$ for each i, and $B_i \cap B_j = \varnothing$ if $i \neq j$) with union $\bigcup_i B_i = \Omega$. The following *partition theorem* is extremely useful.

Theorem 1.48 (Partition theorem) *If $\{B_1, B_2, \dots\}$ is a partition of Ω with $\mathbb{P}(B_i) > 0$ for each i, then*

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i) \qquad \text{for } A \in \mathcal{F}.$$

This theorem has several other fancy names such as 'the theorem of total probability', and it is closely related to 'Bayes' theorem', Theorem 1.50.

Proof We have that

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}\left(A \cap \left(\bigcup_i B_i\right)\right) \\
&= \mathbb{P}\left(\bigcup_i (A \cap B_i)\right) \\
&= \sum_i \mathbb{P}(A \cap B_i) && \text{by (1.14)} \\
&= \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i) && \text{by (1.32).}
\end{aligned}$$

□

Here is an example of this theorem in action in a two-stage calculation.

Example 1.49 Tomorrow there will be either rain or snow but not both; the probability of rain is $\frac{2}{5}$ and the probability of snow is $\frac{3}{5}$. If it rains, the probability that I will be late for my lecture is $\frac{1}{5}$, while the corresponding probability in the event of snow is $\frac{3}{5}$. What is the probability that I will be late?

Solution Let A be the event that I am late and B be the event that it rains. The pair B, B^{c} is a partition of the sample space (since exactly one of them must occur). By Theorem 1.48,

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^{\mathrm{c}})\mathbb{P}(B^{\mathrm{c}}) \\
&= \tfrac{1}{5} \cdot \tfrac{2}{5} + \tfrac{3}{5} \cdot \tfrac{3}{5} = \tfrac{11}{25}.
\end{aligned}$$

△

There are many practical situations in which you wish to deduce something from a piece of evidence. We write A for the evidence, and $B_1, B_2, \dots$ for the possible 'states of nature'. Suppose there are good estimates for the conditional probabilities $\mathbb{P}(A \mid B_i)$, but we seek instead a probability of the form $\mathbb{P}(B_j \mid A)$.

Theorem 1.50 (Bayes' theorem) *Let $\{B_1, B_2, \dots\}$ be a partition of the sample space Ω such that $\mathbb{P}(B_i) > 0$ for each i. For any event A with $\mathbb{P}(A) > 0$,*

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.$$

Proof By the definition of conditional probability (see Exercise 1.35),

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\mathbb{P}(A)},$$

and the claim follows by the partition theorem, Theorem 1.48. □

Example 1.51 (False positives) A rare but potentially fatal disease has an incidence of 1 in 10^5 in the population at large. There is a diagnostic test, but it is imperfect. If you have the disease, the test is positive with probability $\frac{9}{10}$; if you do not, the test is positive with probability $\frac{1}{20}$. Your test result is positive. What is the probability that you have the disease?

Solution Write D for the event that you have the disease, and P for the event that the test is positive. By Bayes' theorem, Theorem 1.50,

$$\begin{aligned}\mathbb{P}(D \mid P) &= \frac{\mathbb{P}(P \mid D)\mathbb{P}(D)}{\mathbb{P}(P \mid D)\mathbb{P}(D) + \mathbb{P}(P \mid D^{\mathrm{c}})\mathbb{P}(D^{\mathrm{c}})} \\ &= \frac{\frac{9}{10} \cdot \frac{1}{10^5}}{\frac{9}{10} \cdot \frac{1}{10^5} + \frac{1}{20} \cdot \frac{10^5 - 1}{10^5}} \approx 0.0002.\end{aligned}$$

It is more likely that the result of the test is incorrect than that you have the disease. △

Exercise 1.52 Here are two routine problems about balls in urns. You are presented with two urns. Urn I contains 3 white and 4 black balls, and Urn II contains 2 white and 6 black balls.

(a) You pick a ball randomly from Urn I and place it in Urn II. Next you pick a ball randomly from Urn II. What is the probability that the ball is black?

(b) This time, you pick an urn at random, each of the two urns being picked with probability $\frac{1}{2}$, and you pick a ball at random from the chosen urn. Given the ball is black, what is the probability you picked Urn I?

Exercise 1.53 A biased coin shows heads with probability $p = 1 - q$ whenever it is tossed. Let u_n be the probability that, in n tosses, no two heads occur successively. Show that, for $n \geq 1$,

$$u_{n+2} = qu_{n+1} + pqu_n,$$

and find u_n by the usual method (described in Appendix B) when $p = \frac{2}{3}$.

1.9 Probability measures are continuous

There is a certain property of probability measures which will be very useful later, and we describe this next. Too great an emphasis should not be placed on the property at this stage, and we recommend to the reader that he or she omit this section at the first reading.

A sequence $A_1, A_2, \ldots$ of events in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called *increasing* if

$$A_n \subseteq A_{n+1} \qquad \text{for } n = 1, 2, \ldots.$$

The union

$$A = \bigcup_{i=1}^{\infty} A_i$$

of such a sequence is called the *limit* of the sequence, and it is an elementary consequence of the axioms for an event space that A is an event. It is perhaps not surprising that the probability $\mathbb{P}(A)$ of A may be expressed as the limit $\lim_{n\to\infty} \mathbb{P}(A_n)$ of the probabilities of the A_n.

Theorem 1.54 (Continuity of probability measures) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $A_1, A_2, \ldots$ is an increasing sequence of events in $\mathcal{F}$ with limit A, then*

$$\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

We precede the proof of the theorem with an application.

Example 1.55 It is intuitively clear that the chance of obtaining no heads in an infinite set of tosses of a fair coin is 0. A rigorous proof goes as follows. Let A_n be the event that the first n tosses of the coin yield at least one head. Then

$$A_n \subseteq A_{n+1} \qquad \text{for } n = 1, 2, \ldots,$$

so that the A_n form an increasing sequence. The limit set A is the event that heads occurs sooner or later. By the continuity of $\mathbb{P}$, Theorem 1.54,

$$\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

However, $\mathbb{P}(A_n) = 1 - (\frac{1}{2})^n$, and so $\mathbb{P}(A_n) \to 1$ as $n \to \infty$. Thus $\mathbb{P}(A) = 1$, giving that the probability $\mathbb{P}(\Omega \setminus A)$, that no head ever appears, equals 0. △

Proof of Theorem 1.54 Let $B_i = A_i \setminus A_{i-1}$. Then

$$A = A_1 \cup B_2 \cup B_3 \cup \cdots$$

is the union of disjoint events in $\mathcal{F}$ (draw a Venn diagram to make this clear). By (1.14),

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A_1) + \mathbb{P}(B_2) + \mathbb{P}(B_3) + \cdots \\ &= \mathbb{P}(A_1) + \lim_{n\to\infty} \sum_{k=2}^{n} \mathbb{P}(B_k).\end{aligned}$$

However,

$$\mathbb{P}(B_i) = \mathbb{P}(A_i) - \mathbb{P}(A_{i-1}) \qquad \text{for } i = 2, 3, \ldots,$$

and so

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A_1) + \lim_{n\to\infty} \sum_{k=2}^{n} \big[\mathbb{P}(A_k) - \mathbb{P}(A_{k-1})\big] \\ &= \lim_{n\to\infty} \mathbb{P}(A_n)\end{aligned}$$

as required, since the last sum collapses. □

The conclusion of Theorem 1.54 is expressed in terms of an *increasing* sequence of events, but the corresponding statement for a *decreasing* sequence is valid too: if $B_1, B_2, \ldots$ is a sequence of events in $\mathcal{F}$ such that $B_i \supseteq B_{i+1}$ for $i = 1, 2, \ldots$, then $\mathbb{P}(B_n) \to \mathbb{P}(B)$ as $n \to \infty$, where $B = \bigcap_{i=1}^{\infty} B_i$ is the limit of the B_i as $i \to \infty$. The shortest way to show this is to set $A_i = \Omega \setminus B_i$ in the theorem.

1.10 Worked problems

Example 1.56 A fair six-sided die is thrown twice (when applied to such objects as dice or coins, the adjectives 'fair' and 'unbiased' imply that each possible outcome has equal probability of occurring).

(a) Write down the probability space of this experiment.
(b) Let B be the event that the first number thrown is no larger than 3, and let C be the event that the sum of the two numbers thrown equals 6. Find the probabilities of B and C, and the conditional probabilities of C given B, and of B given C.

Solution The probability space of this experiment is the triple $(\Omega, \mathcal{F}, \mathbb{P})$, where

(i) $\Omega = \big\{(i, j) : i, j = 1, 2, \ldots, 6\big\}$, the set of all ordered pairs of integers between 1 and 6,
(ii) $\mathcal{F}$ is the set of all subsets of Ω,
(iii) each point in Ω has equal probability, so that

$$\mathbb{P}\big((i, j)\big) = \tfrac{1}{36} \qquad \text{for } i, j = 1, 2, \ldots, 6,$$

and, more generally,

$$\mathbb{P}(A) = \tfrac{1}{36}|A| \qquad \text{for each } A \subseteq \Omega.$$

The events B and C are subsets of Ω given by

$$\begin{aligned}B &= \big\{(i, j) : i = 1, 2, 3 \text{ and } j = 1, 2, \ldots, 6\big\}, \\ C &= \big\{(i, j) : i + j = 6 \text{ and } i, j = 1, 2, \ldots, 6\big\}.\end{aligned}$$

The event B contains $3 \times 6 = 18$ ordered pairs, and C contains 5 ordered pairs, giving that

$$\mathbb{P}(B) = \tfrac{18}{36} = \tfrac{1}{2}, \qquad \mathbb{P}(C) = \tfrac{5}{36}.$$

Finally, $B \cap C$ is given by

$$B \cap C = \big\{(1,5),(2,4),(3,3)\big\}$$

containing just 3 ordered pairs, so that

$$\mathbb{P}(C \mid B) = \frac{\mathbb{P}(C \cap B)}{\mathbb{P}(B)} = \tfrac{3}{36}\big/\tfrac{18}{36} = \tfrac{1}{6},$$

and

$$\mathbb{P}(B \mid C) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)} = \tfrac{3}{36}\big/\tfrac{5}{36} = \tfrac{3}{5}. \qquad \triangle$$

Example 1.57 You are travelling on a train with your sister. Neither of you has a valid ticket, and the inspector has caught you both. He is authorized to administer a special punishment for this offence. He holds a box containing nine apparently identical chocolates, three of which are contaminated with a deadly poison. He makes each of you, in turn, choose and immediately eat a single chocolate.

(a) If you choose before your sister, what is the probability that you will survive?
(b) If you choose first and survive, what is the probability that your sister survives?
(c) If you choose first and die, what is the probability that your sister survives?
(d) Is it in your best interests to persuade your sister to choose first?
(e) If you choose first, what is the probability that you survive, given that your sister survives?

Solution Let A be the event that the first chocolate picked is not poisoned, and let B be the event that the second chocolate picked is not poisoned. Elementary calculations, if you are allowed the time to perform them, would show that

$$\mathbb{P}(A) = \tfrac{6}{9}, \qquad \mathbb{P}(B \mid A) = \tfrac{5}{8}, \qquad \mathbb{P}(B \mid A^{\mathrm{c}}) = \tfrac{6}{8},$$

giving by the partition theorem, Theorem 1.48, that

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B \mid A)\mathbb{P}(A) + \mathbb{P}(B \mid A^{\mathrm{c}})\mathbb{P}(A^{\mathrm{c}}) \\ &= \tfrac{5}{8} \cdot \tfrac{6}{9} + \tfrac{6}{8} \cdot (1 - \tfrac{6}{9}) = \tfrac{2}{3}.\end{aligned}$$

Hence $\mathbb{P}(A) = \mathbb{P}(B)$, so that the only reward of choosing second is to increase your life expectancy by a few seconds.

The final question (e) seems to be the wrong way round in time, since your sister chooses her chocolate *after* you. The way to answer such a question is to reverse the conditioning as follows:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \mathbb{P}(B \mid A)\frac{\mathbb{P}(A)}{\mathbb{P}(B)}, \tag{1.58}$$

and hence

$$\mathbb{P}(A \mid B) = \tfrac{5}{8} \cdot \tfrac{6}{9}\big/\tfrac{2}{3} = \tfrac{5}{8}.$$

We note that $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)$, in agreement with our earlier observation that the order in which you and your sister pick from the box is irrelevant to your chances of survival. $\triangle$

Example 1.59 A coin is tossed $2n$ times. What is the probability of exactly n heads? How does your answer behave for large n?

Solution The sample space is the set of possible outcomes. It has 2^{2n} elements, each of which is equally likely. There are $\binom{2n}{n}$ ways to throw exactly n heads. Therefore, the answer is

$$\frac{1}{2^{2n}}\binom{2n}{n}. \tag{1.60}$$

To understand how this behaves for large n, we need to expand the binomial coefficient in terms of polynomials and exponentials. The relevant asymptotic formula is called *Stirling's formula*,

$$n! \sim (n/e)^n \sqrt{2\pi n} \qquad \text{as } n \to \infty, \tag{1.61}$$

where $a_n \sim b_n$ means $a_n/b_n \to 1$ as $n \to \infty$. See Theorem A.4 for a partial proof of this.

Applying Stirling's formula to (1.60), we obtain

$$\begin{aligned}\frac{1}{2^{2n}}\binom{2n}{n} &= 2^{-2n}\frac{(2n)!}{(n!)^2} \\ &\sim 2^{-2n}\frac{(2n/e)^{2n}\sqrt{2\pi 2n}}{(n/e)^{2n}(2\pi n)} = \frac{1}{\sqrt{\pi n}}.\end{aligned}$$

The factorials and exponentials are gigantic but they cancel out. △

Example 1.62 (Simpson's paradox) The following comparison of surgical procedures is taken from Charig *et al.* (1986). Two treatments are considered for kidney stones, namely open surgery (abbreviated to OS) and percutaneous nephrolithotomy (PN). It is reported that OS has a success rate of 78% (= 273/350) and PN a success rate of 83% (= 289/350). This looks like a marginal advantage to PN. On looking more closely, the patients are divided into two groups depending on whether or not their stones are smaller than 2 cm, with the following success rates.

	stone < 2 cm	stone > 2 cm	Total
OS	93% (= 81/87)	73% (= 192/263)	78% (= 273/350)
PN	87% (= 234/270)	68% (= 55/80)	83% (= 289/350)

Open surgery wins in both cases! Discuss. △

1.11 Problems

1. A fair die is thrown n times. Show that the probability that there are an even number of sixes is $\frac{1}{2}\left[1 + (\frac{2}{3})^n\right]$. For the purpose of this question, 0 is an even number.
2. Does there exist an event space containing just six events?
3. Prove *Boole's inequality*:

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

4. Prove that
$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - n + \sum_{i=1}^{n} \mathbb{P}(A_i).$$
This is sometimes called *Bonferroni's inequality*, but the term is not recommended since it has multiple uses.
5. Two fair dice are thrown. Let A be the event that the first shows an odd number, B be the event that the second shows an even number, and C be the event that either both are odd or both are even. Show that A, B, C are pairwise independent but not independent.
6. Urn I contains 4 white and 3 black balls, and Urn II contains 3 white and 7 black balls. An urn is selected at random, and a ball is picked from it. What is the probability that this ball is black? If this ball is white, what is the probability that Urn I was selected?
7. A single card is removed at random from a deck of 52 cards. From the remainder we draw two cards at random and find that they are both spades. What is the probability that the first card removed was also a spade?
8. A fair coin is tossed $3n$ times. Find the probability that the number of heads is twice the number of tails. Expand your answer using Stirling's formula.
9. Two people toss a fair coin n times each. Show that the probability they throw equal numbers of heads is
$$\binom{2n}{n}\left(\frac{1}{2}\right)^{2n}.$$

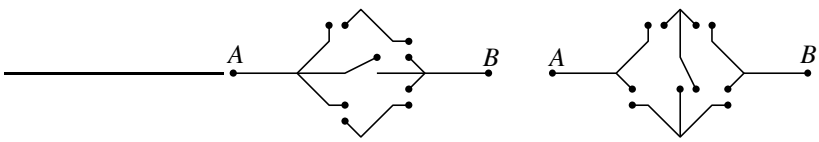


Fig. 1.2 Two electrical circuits incorporating switches.

10. In the circuits in Figure 1.2, each switch is closed with probability p, independently of all other switches. For each circuit, find the probability that a flow of current is possible between A and B.
11. Show that if u_n is the probability that n tosses of a fair coin contain no run of 4 heads, then for $n \geq 4$
$$u_n = \tfrac{1}{2}u_{n-1} + \tfrac{1}{4}u_{n-2} + \tfrac{1}{8}u_{n-3} + \tfrac{1}{16}u_{n-4}.$$
Use this difference equation to show that $u_8 = \frac{208}{256}$.

* 12. Any number $\omega \in [0, 1]$ has a decimal expansion
$$\omega = 0.x_1x_2\ldots,$$
and we write $f_k(\omega, n)$ for the proportion of times that the integer k appears in the first n digits in this expansion. We call ω a *normal number* if
$$f_k(\omega, n) \to \tfrac{1}{10} \qquad \text{as } n \to \infty$$
for $k = 0, 1, 2, \ldots, 9$. On intuitive grounds we may expect that most numbers $\omega \in [0, 1]$ are normal numbers, and Borel proved that this is indeed true. It is quite another matter to exhibit specific normal numbers. Prove the number

$$0.1234567891011121314\ldots$$

is normal. It is an unsolved problem of mathematics to show that $e-2$ and $\pi-3$ are normal numbers also.

13. A square board is divided into 16 equal squares by lines drawn parallel to its sides. A counter is placed at random on one of these squares and is then moved n times. At each of these moves, it can be transferred to any neighbouring square, horizontally, vertically, or diagonally, all such moves being equally likely.

Let c_n be the probability that a particular corner site is occupied after n such independent moves, and let the corresponding probabilities for an intermediate site at the side of the board and for a site in the middle of the board be s_n and m_n, respectively. Show that

$$4c_n + 8s_n + 4m_n = 1, \qquad n = 0, 1, 2, \ldots,$$

and that

$$c_n = \tfrac{2}{5}s_{n-1} + \tfrac{1}{8}m_{n-1}, \qquad n = 1, 2, \ldots.$$

Find two other relations for s_n and m_n in terms of c_{n-1}, s_{n-1}, and m_{n-1}, and hence find c_n, s_n, and m_n. (Oxford 1974M)

14. (a) Let $\mathbb{P}(A)$ denote the probability of the occurrence of an event A. Prove carefully, for events $A_1, A_2, \ldots, A_n$, that

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1}\mathbb{P}\left(\bigcap_i A_i\right).$$

(b) One evening, a bemused lodge-porter tried to hang n keys on their n hooks, but only managed to hang them independently and at random. There was no limit to the number of keys which could be hung on any hook. Otherwise, or by using (a), find an expression for the probability that at least one key was hung on its own hook.

The following morning, the porter was rebuked by the Bursar, so that in the evening she was careful to hang only one key on each hook. But she still only managed to hang them independently and at random. Find an expression for the probability that no key was then hung on its own hook.

Find the limits of both expressions as n tends to infinity.

You may assume that, for real x,

$$e^x = \sum_{r=0}^{\infty} \frac{x^r}{r!} = \lim_{N\to\infty}\left(1 + \frac{x}{N}\right)^N.$$

(Oxford 1978M)

15. Two identical decks of cards, each containing N cards, are shuffled randomly. We say that a *k-matching* occurs if the two decks agree in exactly k places. Show that the probability that there is a k-matching is

$$\pi_k = \frac{1}{k!}\left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^{N-k}}{(N-k)!}\right)$$

for $k = 0, 1, 2, \ldots, N$. We note that $\pi_k \simeq 1/(k!\,e)$ for large N and fixed k. Such matching probabilities are used in testing departures from randomness in circumstances such as psychological tests and wine-tasting competitions. (The convention is that $0! = 1$.)

16. The buses which stop at the end of my road do not keep to the timetable. They should run every quarter hour, at 08.30, 08.45, 09.00, ..., but in fact each bus is either five minutes early or five minutes late, the two possibilities being equally probable and different buses being independent. Other people arrive at the stop in such a way that, t minutes after the departure of one bus, the probability that no one is waiting for the next one is $e^{-t/5}$. What is the probability that no one is waiting at 09.00? One day, I come to the stop at 09.00 and find no one there; show that the chances are more than four to one that I have missed the nine o'clock bus.

    You may use an approximation $e^3 \approx 20$. (Oxford 1977M)

17. A coin is tossed repeatedly; on each toss a head is shown with probability p, or a tail with probability $1 - p$. The outcomes of the tosses are independent. Let E denote the event that the first run of r successive heads occurs earlier that the first run of s successive tails. Let A denote the outcome of the first toss. Show that

$$\mathbb{P}(E \mid A = \text{head}) = p^{r-1} + (1 - p^{r-1})\mathbb{P}(E \mid A = \text{tail}).$$

    Find a similar expression for $\mathbb{P}(E \mid A = \text{tail})$, and hence find $\mathbb{P}(E)$. (Oxford 1981M)

* 18. Show that the axiom that $\mathbb{P}$ is countably additive is equivalent to the axiom that $\mathbb{P}$ is finitely additive and continuous. That is to say, let Ω be a set and $\mathcal{F}$ an event space of subsets of Ω. If $\mathbb{P}$ is a mapping from $\mathcal{F}$ into $[0, 1]$ satisfying

    (i) $\mathbb{P}(\Omega) = 1, \mathbb{P}(\varnothing) = 0$,
    (ii) if $A, B \in \mathcal{F}$ and $A \cap B = \varnothing$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$,
    (iii) if $A_1, A_2, \ldots \in \mathcal{F}$ and $A_i \subseteq A_{i+1}$ for $i = 1, 2, \ldots$, then

$$\mathbb{P}(A) = \lim_{i \to \infty} \mathbb{P}(A_i),$$

    where $A = \bigcup_{i=1}^{\infty} A_i$,

    then $\mathbb{P}$ satisfies $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$ for all sequences $A_1, A_2, \ldots$ of disjoint events.

19. There are n socks in a drawer, three of which are red and the rest black. John chooses his socks by selecting two at random from the drawer, and puts them on. He is three times more likely to wear socks of different colours than to wear matching red socks. Find n.

    For this value of n, what is the probability that John wears matching black socks? (Cambridge 2008)

2

Discrete random variables

Summary. Discrete random variables are studied via their probability mass functions. This leads to the definition of the 'mean value' or 'expectation' of a random variable. There are discussions of variance, and of functions of random variables. Methods are presented for calculating expectations, including the use of conditional expectation.

2.1 Probability mass functions

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we are often interested in situations involving some real-valued function X acting on Ω. For example, let $\mathcal{E}$ be the experiment of throwing a fair die once, so that $\Omega = \{1, 2, 3, 4, 5, 6\}$, and suppose that we gamble on the outcome of $\mathcal{E}$ in such a way that the profit is

$$\begin{aligned} -1 &\quad \text{if the outcome is 1, 2, or 3,} \\ 0 &\quad \text{if the outcome is 4,} \\ 2 &\quad \text{if the outcome is 5 or 6,} \end{aligned}$$

where negative profits are positive losses. If the outcome is ω, then our profit is $X(\omega)$, where $X : \Omega \to \mathbb{R}$ is defined by

$$X(1) = X(2) = X(3) = -1, \quad X(4) = 0, \quad X(5) = X(6) = 2.$$

The mapping X is an example of a 'discrete random variable'.

More formally, a *discrete random variable* X on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is defined to be a mapping $X : \Omega \to \mathbb{R}$ such that

$$\text{the image } X(\Omega) \text{ is a countable subset of } \mathbb{R},^{1} \text{ and} \tag{2.1}$$

$$\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F} \quad \text{for } x \in \mathbb{R}. \tag{2.2}$$

The word 'discrete' here refers to the condition that X takes only countably many values in $\mathbb{R}$.[2] Condition (2.2) is obscure at first sight, and the point here is as follows. A discrete random variable X takes values in $\mathbb{R}$, but we cannot predict the actual value of X with certainty

[1] If $X : \Omega \to \mathbb{R}$ and $A \subseteq \Omega$, the *image* of A is the set $X(A) = \{X(\omega) : \omega \in A\}$ of values taken by X on A.

[2] A slightly different but morally equivalent definition of a discrete random variable is a function $X : \Omega \to \mathbb{R}$ such that there exists a countable subset $S \subseteq \mathbb{R}$ with $\mathbb{P}(X \in S) = 1$.

since the underlying experiment $\mathcal{E}$ involves chance. Instead, we would like to measure the probability that X takes a given value, x say. To this end, we note that X takes the value x if and only if the result of $\mathcal{E}$ lies in that subset of Ω which is mapped into x, namely the subset $X^{-1}(x) = \{\omega \in \Omega : X(\omega) = x\}$. Condition (2.2) postulates that all such subsets are events, in that they belong to $\mathcal{F}$, and are therefore assigned probabilities by $\mathbb{P}$.

The most interesting things about a discrete random variable are the values which it may take and the probabilities associated with these values. If X is a discrete random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then its *image* $\operatorname{Im} X$ is the image of Ω under X, that is, the set of values taken by X.

Henceforth, we abbreviate events of the form $\{\omega \in \Omega : X(\omega) = x\}$ to the more convenient form $\{X = x\}$.

Definition 2.3 *The **(probability) mass function** (or **pmf**) of the discrete random variable X is the function $p_X : \mathbb{R} \to [0, 1]$ defined by*

$$p_X(x) = \mathbb{P}(X = x). \tag{2.4}$$

Thus, $p_X(x)$ is the probability that the mapping X takes the value x. Note that $\operatorname{Im} X$ is countable for any discrete random variable X, and

$$p_X(x) = 0 \qquad \text{if } x \notin \operatorname{Im} X, \tag{2.5}$$

$$\begin{aligned} \sum_{x \in \operatorname{Im} X} p_X(x) &= \mathbb{P}\left(\bigcup_{x \in \operatorname{Im} X} \{\omega \in \Omega : X(\omega) = x\} \right) \quad \text{by (1.14)} \\ &= \mathbb{P}(\Omega) = 1. \end{aligned} \tag{2.6}$$

Equation (2.6) is sometimes written as

$$\sum_{x \in \mathbb{R}} p_X(x) = 1,$$

in the light of the fact that only countably many values of x make non-zero contributions to this sum. Condition (2.6) essentially characterizes mass functions of discrete random variables in the sense of the following theorem.

Theorem 2.7 *Let $S = \{s_i : i \in I\}$ be a countable set of distinct real numbers, and let $\{\pi_i : i \in I\}$ be a collection of real numbers satisfying*

$$\pi_i \geq 0 \quad \text{for } i \in I, \text{ and} \quad \sum_{i \in I} \pi_i = 1.$$

There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a discrete random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ such that the probability mass function of X is given by

$$\begin{aligned} p_X(s_i) &= \pi_i \qquad \text{for } i \in I, \\ p_X(s) &= 0 \qquad \text{if } s \notin S. \end{aligned}$$

Proof Take $\Omega = S$, $\mathcal{F}$ to be the set of all subsets of Ω, and

$$\mathbb{P}(A) = \sum_{i:\, s_i \in A} \pi_i \qquad \text{for } A \in \mathcal{F}.$$

Finally, define $X : \Omega \to \mathbb{R}$ by $X(\omega) = \omega$ for $\omega \in \Omega$. □

This theorem is very useful, since for many purposes it allows us to forget about sample spaces, event spaces, and probability measures; we need only say 'let X be a random variable taking the value s_i with probability π_i, for $i \in I$' and we can be sure that such a random variable exists without having to construct it explicitly.

In the next section, we present a list of some of the most common types of discrete random variables.

Exercise 2.8 If X and Y are discrete random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, show that U and V are discrete random variables on this space also, where

$$U(\omega) = X(\omega) + Y(\omega), \quad V(\omega) = X(\omega)Y(\omega), \qquad \text{for } \omega \in \Omega.$$

Exercise 2.9 Show that if $\mathcal{F}$ is the power set of Ω, then all functions which map Ω into a countable subset of $\mathbb{R}$ are discrete random variables.

Exercise 2.10 If E is an event of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ show that the *indicator function* of E, defined to be the function 1_E on Ω given by

$$1_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E, \\ 0 & \text{if } \omega \notin E, \end{cases}$$

is a discrete random variable.

Exercise 2.11 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space in which

$$\Omega = \{1, 2, 3, 4, 5, 6\}, \qquad \mathcal{F} = \big\{\varnothing, \{2, 4, 6\}, \{1, 3, 5\}, \Omega\big\},$$

and let U, V, W be functions on Ω defined by

$$U(\omega) = \omega, \qquad V(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is even}, \\ 0 & \text{if } \omega \text{ is odd}, \end{cases} \qquad W(\omega) = \omega^2,$$

for $\omega \in \Omega$. Determine which of U, V, W are discrete random variables on the probability space.

Exercise 2.12 For what value of c is the function p, defined by

$$p(k) = \begin{cases} \dfrac{c}{k(k+1)} & \text{if } k = 1, 2, \ldots, \\ 0 & \text{otherwise}, \end{cases}$$

a mass function?

2.2 Examples

Certain types of discrete random variables occur frequently, and we list some of these. Throughout this section, n is a positive integer, p is a number in $[0, 1]$, and $q = 1 - p$. We never describe the underlying probability space.

Bernoulli distribution. This is the simplest non-trivial distribution. We say that the discrete random variable X has the Bernoulli distribution with parameter p if the image of X is $\{0, 1\}$, so that X takes the values 0 and 1 only.

Such a random variable X is often called simply a *coin toss*. There exists $p \in [0, 1]$ such that

$$\mathbb{P}(X = 0) = q, \qquad \mathbb{P}(X = 1) = p, \tag{2.13}$$

and the mass function of X is given by

$$p_X(0) = q, \quad p_X(1) = p, \quad p_X(x) = 0 \text{ if } x \neq 0, 1.$$

Coin tosses are the building blocks of probability theory. There is a sense in which the entire theory can be constructed from an infinite sequence of coin tosses.

Binomial distribution. We say that X has the binomial distribution with parameters n and p if X takes values in $\{0, 1, \ldots, n\}$ and

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k} \qquad \text{for } k = 0, 1, 2, \ldots, n. \tag{2.14}$$

Note that (2.14) gives rise to a mass function satisfying (2.6) since, by the binomial theorem,

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p + q)^n = 1.$$

Poisson distribution. We say that X has the Poisson distribution with parameter λ (> 0) if X takes values in $\{0, 1, 2, \ldots\}$ and

$$\mathbb{P}(X = k) = \frac{1}{k!} \lambda^k e^{-\lambda} \qquad \text{for } k = 0, 1, 2, \ldots. \tag{2.15}$$

Again, this gives rise to a mass function since

$$\sum_{k=0}^{\infty} \frac{1}{k!} \lambda^k e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{k!} \lambda^k = e^{-\lambda} e^{\lambda} = 1.$$

Geometric distribution. We say that X has the geometric distribution with parameter $p \in (0, 1)$ if X takes values in $\{1, 2, 3, \ldots\}$ and

$$\mathbb{P}(X = k) = pq^{k-1} \qquad \text{for } k = 1, 2, 3, \ldots. \tag{2.16}$$

As before, note that

$$\sum_{k=1}^{\infty} pq^{k-1} = \frac{p}{1-q} = 1.$$

Negative binomial distribution. We say that X has the negative binomial distribution with parameters n and $p \in (0, 1)$ if X takes values in $\{n, n+1, n+2, \dots\}$ and

$$\mathbb{P}(X = k) = \binom{k-1}{n-1} p^n q^{k-n} \qquad \text{for } k = n, n+1, n+2, \dots. \tag{2.17}$$

As before, note that

$$\begin{aligned}\sum_{k=n}^{\infty} \binom{k-1}{n-1} p^n q^{k-n} &= p^n \sum_{l=0}^{\infty} \binom{n+l-1}{l} q^l \qquad \text{where } l = k-n \\ &= p^n \sum_{l=0}^{\infty} \binom{-n}{l} (-q)^l \\ &= p^n (1-q)^{-n} = 1,\end{aligned}$$

using the binomial expansion of $(1-q)^{-n}$, see Theorem A.3.

Example 2.18 Here is an example of some of the above distributions in action. Suppose that a coin is tossed n times and there is probability p that heads appears on each toss. Representing heads by H and tails by T, the sample space is the set Ω of all ordered sequences of length n containing the letters H and T, where the kth entry of such a sequence represents the result of the kth toss. The set Ω is finite, and we take $\mathcal{F}$ to be the set of all subsets of Ω. For each $\omega \in \Omega$, we define the probability that ω is the actual outcome by

$$\mathbb{P}(\omega) = p^{h(\omega)} q^{t(\omega)},$$

where $h(\omega)$ is the number of heads in ω and $t(\omega) = n - h(\omega)$ is the number of tails. Similarly, for any $A \in \mathcal{F}$,

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

For $i = 1, 2, \dots, n$, we define the discrete random variable X_i by

$$X_i(\omega) = \begin{cases} 1 & \text{if the } i\text{th entry in } \omega \text{ is H}, \\ 0 & \text{if the } i\text{th entry in } \omega \text{ is T}. \end{cases}$$

Each X_i takes values in $\{0, 1\}$ and has mass function given by

$$\mathbb{P}(X_i = 0) = \mathbb{P}\big(\{\omega \in \Omega : \omega_i = \text{T}\}\big),$$

where ω_i is the ith entry in ω. Thus

$$\begin{aligned}\mathbb{P}(X_i = 0) &= \sum_{\omega:\,\omega_i = \mathrm{T}} p^{h(\omega)} q^{n-h(\omega)} \\ &= \sum_{h=0}^{n-1} \sum_{\substack{\omega:\,\omega_i = \mathrm{T},\\ h(\omega)=h}} p^h q^{n-h} = \sum_{h=0}^{n-1} \binom{n-1}{h} p^h q^{n-h} \\ &= q(p+q)^{n-1} = q\end{aligned}$$

and

$$\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p.$$

Hence, each X_i has the Bernoulli distribution with parameter p. We have derived this fact in a cumbersome manner, but we believe these details to be instructive.

Let

$$S_n = X_1 + X_2 + \cdots + X_n,$$

which is to say that $S_n(\omega) = X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)$. Clearly, S_n is the total number of heads which occur, and S_n takes values in $\{0, 1, \ldots, n\}$ since each X_i equals 0 or 1. Also, for $k = 0, 1, \ldots, n$, we have that

$$\begin{aligned}\mathbb{P}(S_n = k) &= \mathbb{P}\big(\{\omega \in \Omega : h(\omega) = k\}\big) \\ &= \sum_{\omega:\,h(\omega)=k} \mathbb{P}(\omega) \\ &= \binom{n}{k} p^k q^{n-k}, \end{aligned} \tag{2.19}$$

and so S_n has the binomial distribution with parameters n and p.

If n is very large and p is very small but np is a 'reasonable size' ($np = \lambda$, say) then the distribution of S_n may be approximated by the Poisson distribution with parameter λ, as follows. For *fixed* $k \geq 0$, write $p = \lambda/n$ and suppose that n is large to find that

$$\begin{aligned}\mathbb{P}(S_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &\sim \frac{n^k}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &\sim \frac{1}{k!} \lambda^k e^{-\lambda}. \end{aligned} \tag{2.20}$$

This approximation may be useful in practice. For example, consider a single page of the Guardian newspaper containing, say, 10^6 characters, and suppose that the typesetter flips a coin before setting each character and then deliberately mis-sets this character whenever the coin comes up heads. If the coin comes up heads with probability 10^{-5} on each flip, then this is the equivalent to taking $n = 10^6$ and $p = 10^{-5}$ in the above example, giving that the number S_n of deliberate mistakes has the binomial distribution with parameters 10^6 and

10^{-5}. It may be easier (and not too inaccurate) to use (2.20) rather than (2.19) to calculate probabilities. In this case, $\lambda = np = 10$ and so, for example,

$$\mathbb{P}(S_n = 10) \approx \frac{1}{10!}(10e^{-1})^{10} \approx 0.125. \qquad \triangle$$

Example 2.21 Suppose that we toss the coin of the previous example until the first head turns up, and then we stop. The sample space now is

$$\Omega = \{\mathrm{H}, \mathrm{TH}, \mathrm{T}^2\mathrm{H}, \dots\} \cup \{\mathrm{T}^\infty\},$$

where $\mathrm{T}^k\mathrm{H}$ represents the outcome of k tails followed by a head, and T^∞ represents an infinite sequence of tails with no head. As before, $\mathcal{F}$ is the set of all subsets of Ω, and $\mathbb{P}$ is given by the observation that

$$\mathbb{P}(\mathrm{T}^k\mathrm{H}) = pq^k \qquad \text{for } k = 0, 1, 2 \dots,$$
$$\mathbb{P}(\mathrm{T}^\infty) = \begin{cases} 1 & \text{if } p = 0, \\ 0 & \text{if } p > 0. \end{cases}$$

Let Y be the total number of tosses in this experiment, so that $Y(\mathrm{T}^k\mathrm{H}) = k+1$ for $0 \le k < \infty$ and $Y(\mathrm{T}^\infty) = \infty$. If $p > 0$, then

$$\mathbb{P}(Y = k) = \mathbb{P}(\mathrm{T}^{k-1}\mathrm{H}) = pq^{k-1} \qquad \text{for } k = 1, 2, \dots,$$

showing that Y has the geometric distribution with parameter p. $\triangle$

Example 2.22 If we carry on tossing the coin in the previous example until the nth head has turned up, then a similar argument shows that, if $p \in (0, 1)$, the total number of tosses required has the negative binomial distribution with parameters n and p. $\triangle$

Exercise 2.23 If X is a discrete random variable having the Poisson distribution with parameter λ, show that the probability that X is even is $e^{-\lambda}\cosh\lambda$.

Exercise 2.24 If X is a discrete random variable having the geometric distribution with parameter p, show that the probability that X is greater than k is $(1-p)^k$.

2.3 Functions of discrete random variables

Let X be a discrete random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $g : \mathbb{R} \to \mathbb{R}$. It is easy to check that $Y = g(X)$ is a discrete random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ also, defined by

$$Y(\omega) = g(X(\omega)) \qquad \text{for } \omega \in \Omega.$$

Simple examples are

$$\text{if } g(x) = ax + b \quad \text{then } g(X) = aX + b,$$
$$\text{if } g(x) = cx^2 \quad \text{then } g(X) = cX^2.$$

If $Y = g(X)$, the mass function of Y is given by

$$\begin{aligned} p_Y(y) = \mathbb{P}(Y = y) &= \mathbb{P}(g(X) = y) \\ &= \mathbb{P}(X \in g^{-1}(y)) \\ &= \sum_{x \in g^{-1}(y)} \mathbb{P}(X = x), \end{aligned} \tag{2.25}$$

since there are only countably many non-zero contributions to this sum. Thus, if $Y = aX + b$ with $a \neq 0$, then

$$\mathbb{P}(Y = y) = \mathbb{P}(aX + b = y) = \mathbb{P}\big(X = a^{-1}(y - b)\big) \qquad \text{for } y \in \mathbb{R},$$

while if $Y = X^2$, then

$$\mathbb{P}(Y = y) = \begin{cases} \mathbb{P}\big(X = \sqrt{y}\big) + \mathbb{P}\big(X = -\sqrt{y}\big) & \text{if } y > 0, \\ \mathbb{P}(X = 0) & \text{if } y = 0, \\ 0 & \text{if } y < 0. \end{cases}$$

Exercise 2.26 Let X be a discrete random variable having the Poisson distribution with parameter λ, and let $Y = |\sin(\frac{1}{2}\pi X)|$. Find the mass function of Y.

2.4 Expectation

Consider a fair die. If it were thrown a large number of times, each of the possible outcomes $1, 2, \ldots, 6$ would appear on about one-sixth of the throws, and the average of the numbers observed would be approximately

$$\tfrac{1}{6} \cdot 1 + \tfrac{1}{6} \cdot 2 + \cdots + \tfrac{1}{6} \cdot 6 = \tfrac{7}{2},$$

which we call the *mean value*. This notion of mean value is easily extended to more general distributions as follows.

Definition 2.27 *If X is a discrete random variable, the **expectation** of X is denoted by $\mathbb{E}(X)$ and defined by*

$$\mathbb{E}(X) = \sum_{x \in \operatorname{Im} X} x\mathbb{P}(X = x) \tag{2.28}$$

whenever this sum converges absolutely, in that $\sum_x |x\mathbb{P}(X = x)| < \infty$.

Equation (2.28) is often written

$$\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x) = \sum_x x p_X(x),$$

and the expectation of X is often called the *expected value* or *mean* of X.[3] The reason for requiring absolute convergence in (2.28) is that the image $\operatorname{Im} X$ may be an infinite set, and we

[3] One should be careful to avoid ambiguity in the use (or not) of parentheses. For example, we shall sometimes write $\mathbb{E}(X)^2$ for $[\mathbb{E}(X)]^2$, and $\mathbb{E}|X|$ for $\mathbb{E}(|X|)$.

need the summation in (2.28) to take the same value irrespective of the order in which we add up its terms.

The physical analogy of 'expectation' is the idea of 'centre of gravity'. If masses with weights $\pi_1, \pi_2, \ldots$ are placed at the points $x_1, x_2, \ldots$ of $\mathbb{R}$, then the position of the centre of gravity is $\sum \pi_i x_i \big/ \sum \pi_i$, or $\sum x_i p_i$, where $p_i = \pi_i \big/ \sum_j \pi_j$ is the proportion of the total weight allocated to position x_i.

If X is a discrete random variable (on some probability space) and $g : \mathbb{R} \to \mathbb{R}$, then $Y = g(X)$ is a discrete random variable also. According to the above definition, we need to know the mass function of Y before we can calculate its expectation. The following theorem provides a useful way of avoiding this tedious calculation.

Theorem 2.29 (Law of the subconscious statistician) *If X is a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}(g(X)) = \sum_{x \in \operatorname{Im} X} g(x)\mathbb{P}(X = x),$$

whenever this sum converges absolutely.

Intuitively, this result is rather clear, since $g(X)$ takes the value $g(x)$ when X takes the value x, an event which has probability $\mathbb{P}(X = x)$. A more formal proof proceeds as follows.

Proof Writing I for the image of X, we have that $Y = g(X)$ has image $g(I)$. Thus

$$\begin{aligned}
\mathbb{E}(Y) &= \sum_{y \in g(I)} y\mathbb{P}(Y = y) \\
&= \sum_{y \in g(I)} y \sum_{x \in I:\ g(x) = y} \mathbb{P}(X = x) \qquad \text{by (2.25)} \\
&= \sum_{x \in I} g(x)\mathbb{P}(X = x)
\end{aligned}$$

if the last sum converges absolutely. □

Two simple but useful properties of expectation are as follows.

Theorem 2.30 *Let X be a discrete random variable and let $a, b \in \mathbb{R}$.*

(a) *If $\mathbb{P}(X \geq 0) = 1$ and $\mathbb{E}(X) = 0$, then $\mathbb{P}(X = 0) = 1$.*
(b) *We have that $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.*

Proof (a) Suppose the assumptions hold. By the definition (2.28) of $\mathbb{E}(X)$, we have that $x\mathbb{P}(X = x) = 0$ for all $x \in \operatorname{Im} X$. Therefore, $\mathbb{P}(X = x) = 0$ for $x \neq 0$, and the claim follows.
(b) This is a simple consequence of Theorem 2.29 with $g(x) = ax + b$. □

Here is an example of Theorem 2.29 in action.

Example 2.31 Suppose that X is a random variable with the Poisson distribution, parameter λ, and we wish to find the expected value of $Y = e^X$. Without Theorem 2.29 we would have to find the mass function of Y. Actually this is not difficult, but it is even easier to apply the theorem to find that

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}(e^X) \\ &= \sum_{k=0}^{\infty} e^k \mathbb{P}(X = k) = \sum_{k=0}^{\infty} e^k \frac{1}{k!} \lambda^k e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{k!} (\lambda e)^k = e^{\lambda(e-1)}.\end{aligned}$$ △

The expectation $\mathbb{E}(X)$ of a discrete random variable X is an indication of the 'centre' of the distribution of X. Another important quantity associated with X is the 'variance' of X, and this is a measure of the degree of dispersion of X about its expectation $\mathbb{E}(X)$.

Definition 2.32 *The **variance*** $\operatorname{var}(X)$ *of a discrete random variable X is defined by*

$$\operatorname{var}(X) = \mathbb{E}\big([X - \mathbb{E}(X)]^2\big). \tag{2.33}$$

We note that, by Theorem 2.29,

$$\operatorname{var}(X) = \sum_{x \in \operatorname{Im} X} (x - \mu)^2 \mathbb{P}(X = x), \tag{2.34}$$

where $\mu = \mathbb{E}(X)$. A rough motivation for this definition is as follows. If the dispersion of X about its expectation is very small, then $|X - \mu|$ tends to be small, giving that $\operatorname{var}(X) = \mathbb{E}(|X - \mu|^2)$ is small also; on the other hand, if there is often a considerable difference between X and its mean, then $|X - \mu|$ may be large, giving that $\operatorname{var}(X)$ is large also.

Equation (2.34) is not always the most convenient way to calculate the variance of a discrete random variable. We may expand the term $(x - \mu)^2$ in (2.34) to obtain

$$\begin{aligned}\operatorname{var}(X) &= \sum_x (x^2 - 2\mu x + \mu^2) \mathbb{P}(X = x) \\ &= \sum_x x^2 \mathbb{P}(X = x) - 2\mu \sum_x x \mathbb{P}(X = x) + \mu^2 \sum_x \mathbb{P}(X = x) \\ &= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 \qquad \text{by (2.28) and (2.6)} \\ &= \mathbb{E}(X^2) - \mu^2,\end{aligned}$$

where $\mu = \mathbb{E}(X)$ as before. Thus we obtain the useful formula

$$\operatorname{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \tag{2.35}$$

Example 2.36 If X has the geometric distribution with parameter p $(= 1 - q)$, the mean of X is

$$\begin{aligned}\mathbb{E}(X) &= \sum_{k=1}^{\infty} kpq^{k-1} \\ &= \frac{p}{(1-q)^2} = \frac{1}{p},\end{aligned}$$

and the variance of X is

$$\operatorname{var}(X) = \sum_{k=1}^{\infty} k^2 pq^{k-1} - \frac{1}{p^2}$$

by (2.35).[4] Now,

$$\begin{aligned}\sum_{k=1}^{\infty} k^2 q^{k-1} &= q \sum_{k=1}^{\infty} k(k-1)q^{k-2} + \sum_{k=1}^{\infty} kq^{k-1} \\ &= \frac{2q}{(1-q)^3} + \frac{1}{(1-q)^2}\end{aligned}$$

by Footnote 4, giving that

$$\begin{aligned}\operatorname{var}(X) &= p\left(\frac{2q}{p^3} + \frac{1}{p^2}\right) - \frac{1}{p^2} \\ &= qp^{-2}.\end{aligned}$$

△

Exercise 2.37 If X has the binomial distribution with parameters n and $p = 1 - q$, show that

$$\mathbb{E}(X) = np, \qquad \mathbb{E}(X^2) = npq + n^2p^2,$$

and deduce the variance of X.

Exercise 2.38 Show that $\operatorname{var}(aX + b) = a^2 \operatorname{var}(X)$ for $a, b \in \mathbb{R}$.

Exercise 2.39 Find $\mathbb{E}(X)$ and $\mathbb{E}(X^2)$ when X has the Poisson distribution with parameter λ, and hence show that the Poisson distribution has variance equal to its mean.

2.5 Conditional expectation and the partition theorem

Suppose that X is a discrete random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and that B is an event with $\mathbb{P}(B) > 0$. If we are given that B occurs, then this information affects the probability distribution of X. That is, probabilities such as $\mathbb{P}(X = x)$ are replaced by conditional probabilities such as $\mathbb{P}(X = x \mid B) = \mathbb{P}\big(\{X(\omega) = x\} \cap B\big)/\mathbb{P}(B)$.

[4] To sum a series such as $\sum_{k=0}^{\infty} kx^{k-1}$, just note that, if $|x| < 1$, then $\sum_k kx^{k-1} = (d/dx)\sum_k x^k$, and hence $\sum_{k=0}^{\infty} kx^{k-1} = (d/dx)(1-x)^{-1} = (1-x)^{-2}$. The relevant property of power series is that they may be differentiated term by term within their circle of convergence. Repeated differentiation of $(1-x)^{-1}$ yields formulae for $\sum_k k(k-1)x^{k-2}$ and similar expressions.

Definition 2.40 *If X is a discrete random variable and $\mathbb{P}(B) > 0$, the **conditional expectation of X given B** is denoted by $\mathbb{E}(X \mid B)$ and defined by*

$$\mathbb{E}(X \mid B) = \sum_{x \in \operatorname{Im} X} x\mathbb{P}(X = x \mid B), \tag{2.41}$$

whenever this sum converges absolutely.

Just as the partition theorem, Theorem 1.48, expressed probabilities in terms of conditional probabilities, so expectations may be expressed in terms of conditional expectations.

Theorem 2.42 (Partition theorem) *If X is a discrete random variable and $\{B_1, B_2, \dots\}$ is a partition of the sample space such that $\mathbb{P}(B_i) > 0$ for each i, then*

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X \mid B_i)\mathbb{P}(B_i), \tag{2.43}$$

whenever this sum converges absolutely.

Proof The right-hand side of (2.43) equals, by (2.41),

$$\begin{aligned}\sum_i \sum_x x\mathbb{P}\big(\{X = x\} \cap B_i\big) &= \sum_x x\mathbb{P}\left(\{X = x\} \cap \left(\bigcup_i B_i\right)\right) \\ &= \sum_x x\mathbb{P}(X = x).\end{aligned}$$

□

We close this chapter with an example of this partition theorem in use.

Example 2.44 A coin is tossed repeatedly, and heads appears at each toss with probability p, where $0 < p = 1 - q < 1$. Find the expected length of the initial run (this is a run of heads if the first toss gives heads, and of tails otherwise).

Solution Let H be the event that the first toss gives heads and H^{c} the event that the first toss gives tails. The pair H, H^{c} forms a partition of the sample space. Let X be the length of the initial run. It is easy to see that

$$\mathbb{P}(X = k \mid H) = p^{k-1}q \qquad \text{for } k = 1, 2, \dots,$$

since if H occurs, then $X = k$ if and only if the first toss is followed by exactly $k - 1$ heads and then a tail. Similarly,

$$\mathbb{P}(X = k \mid H^{\mathrm{c}}) = q^{k-1}p \qquad \text{for } k = 1, 2, \dots.$$

Therefore,

$$\mathbb{E}(X \mid H) = \sum_{k=1}^{\infty} kp^{k-1}q = \frac{q}{(1-p)^2} = \frac{1}{q},$$

and similarly,

$$\mathbb{E}(X \mid H^{\mathrm{c}}) = \frac{1}{p}.$$

By the partition theorem, Theorem 2.42,

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(X \mid H)\mathbb{P}(H) + \mathbb{E}(X \mid H^{\mathrm{c}})\mathbb{P}(H^{\mathrm{c}}) \\ &= \frac{1}{q}p + \frac{1}{p}q \\ &= \frac{1}{pq} - 2.\end{aligned}$$

△

Exercise 2.45 Let X be a discrete random variable and let g be a function from $\mathbb{R}$ to $\mathbb{R}$. If x is a real number such that $\mathbb{P}(X = x) > 0$, show formally that

$$\mathbb{E}(g(X) \mid X = x) = g(x),$$

and deduce from the partition theorem, Theorem 2.42, that

$$\mathbb{E}(g(X)) = \sum_{x} g(x)\mathbb{P}(X = x).$$

Exercise 2.46 Let N be the number of tosses of a fair coin up to and including the appearance of the first head. By conditioning on the result of the first toss, show that $\mathbb{E}(N) = 2$.

2.6 Problems

1. If X has the Poisson distribution with parameter λ, show that

$$\mathbb{E}\big(X(X-1)(X-2)\cdots(X-k)\big) = \lambda^{k+1}$$

for $k = 0, 1, 2, \ldots$.

2. Each toss of a coin results in heads with probability p (> 0). If $m(r)$ is the mean number of tosses up to and including the rth head, show that

$$m(r) = p\big[1 + m(r-1)\big] + (1-p)\big[1 + m(r)\big]$$

for $r = 1, 2, \ldots$, with the convention that $m(0) = 0$. Solve this difference equation by the method described in Appendix B.

3. If X is a discrete random variable and $\mathbb{E}(X^2) = 0$, show that $\mathbb{P}(X = 0) = 1$. Deduce that, if $\mathrm{var}(X) = 0$, then $\mathbb{P}(X = \mu) = 1$, whenever $\mu = \mathbb{E}(X)$ is finite.

4. For what values of c and α is the function p, defined by

$$p(k) = \begin{cases} ck^{\alpha} & \text{for } k = 1, 2, \ldots, \\ 0 & \text{otherwise,} \end{cases}$$

a mass function?

5. *Lack-of-memory property.* If X has the geometric distribution with parameter p, show that

$$\mathbb{P}\big(X > m + n \,\big|\, X > m\big) = \mathbb{P}(X > n)$$

for $m, n = 0, 1, 2, \ldots$.

We say that X has the 'lack-of-memory property' since, if we are given that $X - m > 0$, then the distribution of $X - m$ is the same as the original distribution of X. Show that the geometric distribution is the only distribution concentrated on the positive integers with the lack-of-memory property.

6. The random variable N takes non-negative integer values. Show that

$$\mathbb{E}(N) = \sum_{k=0}^{\infty} \mathbb{P}(N > k)$$

provided that the series on the right-hand side converges.

A fair die having two faces coloured blue, two red and two green, is thrown repeatedly. Find the probability that not all colours occur in the first k throws.

Deduce that, if N is the random variable which takes the value n if all three colours occur in the first n throws but only two of the colours in the first $n - 1$ throws, then the expected value of N is $\frac{11}{2}$. (Oxford 1979M)

7. *Coupon-collecting problem*. There are c different types of coupon, and each coupon obtained is equally likely to be any one of the c types. Find the probability that the first n coupons which you collect do not form a complete set, and deduce an expression for the mean number of coupons you will need to collect before you have a complete set.

* 8. An ambidextrous student has a left and a right pocket, each initially containing n humbugs. Each time he feels hungry, he puts a hand into one of his pockets and, if it is not empty, he takes a humbug from it and eats it. On each occasion, he is equally likely to choose either the left or right pocket. When he first puts his hand into an empty pocket, the other pocket contains H humbugs.

Show that if p_h is the probability that $H = h$, then

$$p_h = \binom{2n-h}{n} \frac{1}{2^{2n-h}} \qquad \text{for } h = 0, 1, \ldots, n,$$

and find the expected value of H, by considering

$$\sum_{h=0}^{n} p_h, \quad \sum_{h=0}^{n} h p_h, \quad \sum_{h=0}^{n} (n-h) p_h,$$

or otherwise. (Oxford 1982M)

9. The probability of obtaining a head when a certain coin is tossed is p. The coin is tossed repeatedly until n heads occur in a row. Let X be the total number of tosses required for this to happen. Find the expected value of X.

10. A population of N animals has had a certain number a of its members captured, marked, and then released. Show that the probability P_n that it is necessary to capture n animals in order to obtain m which have been marked is

$$P_n = \frac{a}{N} \binom{a-1}{m-1} \binom{N-a}{n-m} \Big/ \binom{N-1}{n-1},$$

where $m \leq n \leq N - a + m$. Hence, show that

$$\frac{a}{N}\binom{a-1}{m-1}\frac{(N-a)!}{(N-1)!}\sum_{n=m}^{N-a+m}\frac{(n-1)!\,(N-n)!}{(n-m)!\,(N-a+m-n)!} = 1,$$

and that the expectation of n is $\dfrac{N+1}{a+1}m$. (Oxford 1972M)

3

Multivariate discrete distributions and independence

Summary. Following an extension of the theory of discrete random variables to discrete random *vectors*, the independence of a family of random variables is explored. Much of probability theory is concerned with sums of random variables, and it is shown here how to study the sums of independent variables. Properties of indicator functions are presented, and it is shown how they may be used to facilitate certain calculations.

3.1 Bivariate discrete distributions

Let X and Y be discrete random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Instead of treating X and Y separately, it is often necessary to regard the pair (X, Y) as a random vector taking values in $\mathbb{R}^2$.

Definition 3.1 *If X and Y are discrete random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, the **joint (probability) mass function** $p_{X,Y}$ of X and Y is the function $p_{X,Y} : \mathbb{R}^2 \to [0, 1]$ defined by*

$$p_{X,Y}(x, y) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}\right), \tag{3.2}$$

usually abbreviated to $p_{X,Y}(x, y) = \mathbb{P}(X = x,\ Y = y)$.

It is clear that

$$p_{X,Y}(x, y) = 0 \quad \text{unless } x \in \operatorname{Im} X \text{ and } y \in \operatorname{Im} Y, \tag{3.3}$$

$$\sum_{x \in \operatorname{Im} X} \sum_{y \in \operatorname{Im} Y} p_{X,Y}(x, y) = 1. \tag{3.4}$$

The individual mass functions p_X and p_Y of X and Y may be found from $p_{X,Y}$ thus:

$$\begin{aligned} p_X(x) = \mathbb{P}(X = x) &= \sum_{y \in \operatorname{Im} Y} \mathbb{P}(X = x,\ Y = y) \\ &= \sum_{y} p_{X,Y}(x, y), \end{aligned} \tag{3.5}$$

and similarly,

$$p_Y(y) = \sum_x p_{X,Y}(x, y). \tag{3.6}$$

These mass functions, given by (3.5) and (3.6), are called the *marginal* mass functions of X and Y, respectively, since, if we think of (X, Y) as a randomly chosen point in the plane, then X and Y are the projections of this point onto the coordinate axes.

	$x = 1$	$x = 2$	$x = 3$
$y = 1$	$\frac{1}{12}$	$\frac{3}{18}$	$\frac{1}{6}$
$y = 2$	$\frac{1}{18}$	0	$\frac{5}{18}$
$y = 3$	0	$\frac{3}{18}$	$\frac{1}{12}$

Table 3.1 The joint mass function of the pair X, Y.

Example 3.7 Suppose that X and Y are random variables each taking the values 1, 2, or 3, and that the probability that the pair (X, Y) equals (x, y) is given in Table 3.1 for all relevant values of x and y.

Then, for example,

$$\begin{aligned}\mathbb{P}(X = 3) &= \mathbb{P}(X = 3,\ Y = 1) + \mathbb{P}(X = 3,\ Y = 2) + \mathbb{P}(X = 3,\ Y = 3)\\ &= \tfrac{1}{6} + \tfrac{5}{18} + \tfrac{1}{12} = \tfrac{19}{36}.\end{aligned}$$

Similarly,

$$\mathbb{P}(Y = 2) = \tfrac{1}{18} + 0 + \tfrac{5}{18} = \tfrac{1}{3}. \qquad \triangle$$

Similar ideas apply to families $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of discrete random variables on a probability space. For example, the *joint mass function* of $\mathbf{X}$ is the function $p_{\mathbf{X}}$ defined by

$$p_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(X_1 = x_1,\ X_2 = x_2, \ldots,\ X_n = x_n)$$

for $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

Exercise 3.8 Two cards are drawn at random from a deck of 52 cards. If X denotes the number of aces drawn and Y denotes the number of kings, display the joint mass function of X and Y in the tabular form of Table 3.1.

Exercise 3.9 The pair of discrete random variables (X, Y) has joint mass function

$$\mathbb{P}(X = i,\ Y = j) = \begin{cases} \theta^{i+j+1} & \text{if } i, j = 0, 1, 2,\\ 0 & \text{otherwise,}\end{cases}$$

for some value of θ. Show that θ satisfies the equation

$$\theta + 2\theta^2 + 3\theta^3 + 2\theta^4 + \theta^5 = 1,$$

and find the marginal mass function of X in terms of θ.

3.2 Expectation in the multivariate case

If X and Y are discrete random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ and $g : \mathbb{R}^2 \to \mathbb{R}$, it is easy to check that $Z = g(X, Y)$ is a discrete random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ also, defined formally by $Z(\omega) = g(X(\omega), Y(\omega))$ for $\omega \in \Omega$. The expectation of Z may be calculated directly from the joint mass function $p_{X,Y}(x, y) = \mathbb{P}(X = x,\ Y = y)$, as the following theorem indicates; the proof is exactly analogous to that of Theorem 2.29.

Theorem 3.10 *We have that*

$$\mathbb{E}(g(X, Y)) = \sum_{x \in \operatorname{Im} X} \sum_{y \in \operatorname{Im} Y} g(x, y)\mathbb{P}(X = x,\ Y = y),$$

whenever this sum converges absolutely.

One particular consequence of this is of great importance: the expectation operator $\mathbb{E}$ acts linearly on the set of discrete random variables. That is to say, if X and Y are discrete random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, and $a, b \in \mathbb{R}$, then

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y), \tag{3.11}$$

whenever $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ exist.[1] To see this, we use Theorem 3.10 with $g(x, y) = ax + by$ to obtain that

$$\begin{aligned}
\mathbb{E}(aX + bY) &= \sum_x \sum_y (ax + by)\mathbb{P}(X = x,\ Y = y) \\
&= a \sum_x x \sum_y \mathbb{P}(X = x,\ Y = y) + b \sum_y y \sum_x \mathbb{P}(X = x,\ Y = y) \\
&= a \sum_x x\mathbb{P}(X = x) + b \sum_y y\mathbb{P}(Y = y) \qquad \text{by (3.5) and (3.6)} \\
&= a\mathbb{E}(X) + b\mathbb{E}(Y).
\end{aligned}$$

Exercise 3.12 Suppose that (X, Y) has joint mass function

$$\mathbb{P}(X = i,\ Y = j) = \theta^{i+j+1} \qquad \text{for } i, j = 0, 1, 2.$$

Show that

$$\mathbb{E}(XY) = \theta^3 + 4\theta^4 + 4\theta^5$$

and

$$\mathbb{E}(X) = \theta^2 + 3\theta^3 + 3\theta^4 + 2\theta^5.$$

[1] The linearity of $\mathbb{E}$ extends beyond finite sums of the form (3.11), but the full property involves the convergence of infinite series of random variables, and is therefore beyond the scope of this text. We give an example. If $X_1, X_2, \ldots$ is a sequence of non-negative random variables with sum S, then $\mathbb{E}(S) = \sum_{i=1}^{\infty} \mathbb{E}(X_i)$ regardless of whether this sum converges or diverges. See Grimmett and Stirzaker (2001, eqn (5.6.13)) for further details.

3.3 Independence of discrete random variables

In a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, events A and B are called independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Discrete random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ are called 'independent' if the value taken by X is independent of the value taken by Y.

Definition 3.13 *Two discrete random variables X and Y are **independent** if the pair of events $\{X = x\}$ and $\{Y = y\}$ are independent for all $x, y \in \mathbb{R}$, and we normally write this condition as*

$$\mathbb{P}(X = x,\ Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \qquad \text{for } x, y \in \mathbb{R}. \tag{3.14}$$

*Random variables which are not independent are called **dependent**.*

Condition (3.14) may be expressed as

$$p_{X,Y}(x, y) = \left(\sum_y p_{X,Y}(x, y)\right)\left(\sum_x p_{X,Y}(x, y)\right) \qquad \text{for } x, y \in \mathbb{R}, \tag{3.15}$$

in terms of the joint mass function of X and Y. This latter condition may be simplified as indicated by the following theorem.

Theorem 3.16 *Discrete random variables X and Y are independent if and only if there exist functions $f, g : \mathbb{R} \to \mathbb{R}$ such that the joint mass function of X and Y satisfies*

$$p_{X,Y}(x, y) = f(x)g(y) \qquad \text{for } x, y \in \mathbb{R}. \tag{3.17}$$

Of course, we need only check (3.17) for $x \in \operatorname{Im} X$ and $y \in \operatorname{Im} Y$.

Proof We need only to prove the sufficiency of the condition. Suppose that (3.17) holds for some f and g. By (3.5) and (3.6),

$$p_X(x) = f(x)\sum_y g(y), \qquad p_Y(y) = g(y)\sum_x f(x),$$

and by (3.4)

$$\begin{aligned} 1 = \sum_{x,y} p_{X,Y}(x, y) &= \sum_{x,y} f(x)g(y) \\ &= \sum_x f(x)\sum_y g(y). \end{aligned}$$

Therefore,

$$\begin{aligned} p_{X,Y}(x, y) = f(x)g(y) &= f(x)g(y)\sum_x f(x)\sum_y g(y) \\ &= p_X(x)p_Y(y). \end{aligned}$$

□

Example 3.18 Suppose that X and Y are random variables taking values in the non-negative integers with joint mass function

$$p_{X,Y}(i,\, j) = \mathbb{P}(X = i,\ Y = j) = \frac{1}{i!\, j!}\lambda^i \mu^j e^{-(\lambda+\mu)} \qquad \text{for } i,\, j = 0, 1, 2, \ldots.$$

Immediate from Theorem 3.16 is the fact that X and Y are independent, since their joint mass function may be factorized in the form

$$p_{X,Y}(i,\, j) = \left(\frac{1}{i!}\lambda^i\right)\left(\frac{1}{j!}\mu^j e^{-(\lambda+\mu)}\right),$$

as a function of i multiplied by a function of j. Such a factorization is not unique, and it is more natural to write

$$p_{X,Y}(i,\, j) = \left(\frac{1}{i!}\lambda^i e^{-\lambda}\right)\left(\frac{1}{j!}\mu^j e^{-\mu}\right)$$

as the product of the marginal mass functions: X and Y are independent random variables each having a Poisson distribution, with parameters λ and μ respectively. △

The following is an important property of independent pairs of random variables.

Theorem 3.19 *If X and Y are independent discrete random variables with expectations $\mathbb{E}(X)$ and $\mathbb{E}(Y)$, then*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Proof By Theorem 3.10,

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{x,y} xy\mathbb{P}(X = x,\ Y = y)\\ &= \sum_{x,y} xy\mathbb{P}(X = x)\mathbb{P}(Y = y) \qquad \text{by independence}\\ &= \sum_x x\mathbb{P}(X = x)\sum_y y\mathbb{P}(Y = y) = \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

It is the existence of $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ which authorizes us to interchange the summations as we have done. □

The converse of Theorem 3.19 is false: if $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, then it does not follow that X and Y are independent (see Example 3.22 below). The correct converse is given next.

Theorem 3.20 *Discrete random variables X and Y on $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if and only if*

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \tag{3.21}$$

for all functions $g, h : \mathbb{R} \to \mathbb{R}$ for which the last two expectations exist.

Proof The necessity of (3.21) follows just as in the proof of Theorem 3.19. To prove sufficiency, let $a, b \in \mathbb{R}$ and define g and h by

$$g(x) = \begin{cases} 1 & \text{if } x = a, \\ 0 & \text{if } x \neq a, \end{cases} \qquad h(y) = \begin{cases} 1 & \text{if } y = b, \\ 0 & \text{if } y \neq b. \end{cases}$$

Then

$$\mathbb{E}(g(X)h(Y)) = \mathbb{P}(X = a,\ Y = b)$$

and

$$\mathbb{E}(g(X))\mathbb{E}(h(Y)) = \mathbb{P}(X = a)\mathbb{P}(Y = b),$$

giving by (3.21) that $p_{X,Y}(a, b) = p_X(a)p_Y(b)$. □

Here is an example of two discrete random variables X and Y which are not independent but which satisfy $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Example 3.22 Suppose that X has distribution given by

$$\mathbb{P}(X = -1) = \mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \tfrac{1}{3}$$

and Y is given by

$$Y = \begin{cases} 0 & \text{if } X = 0, \\ 1 & \text{if } X \neq 0. \end{cases}$$

It is easy to find a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, together with two random variables having these distributions. For example, take $\Omega = \{-1, 0, 1\}$, $\mathcal{F}$ the set of all subsets of Ω, $\mathbb{P}$ given by $\mathbb{P}(-1) = \mathbb{P}(0) = \mathbb{P}(1) = \frac{1}{3}$, and $X(\omega) = \omega$, $Y(\omega) = |\omega|$. Then X and Y are dependent since

$$\mathbb{P}(X = 0,\ Y = 1) = 0$$

but

$$\mathbb{P}(X = 0)\mathbb{P}(Y = 1) = \tfrac{1}{3} \cdot \tfrac{2}{3} = \tfrac{2}{9}.$$

On the other hand,

$$\begin{aligned} \mathbb{E}(XY) &= \sum_{x,y} xy\mathbb{P}(X = x,\ Y = y) \\ &= (-1) \cdot \tfrac{1}{3} + 0 \cdot \tfrac{1}{3} + 1 \cdot \tfrac{1}{3} = 0 \end{aligned}$$

and

$$\mathbb{E}(X)\mathbb{E}(Y) = 0 \cdot \tfrac{2}{3} = 0.$$ △

In this section so far, we have considered pairs of random variables only, but the same ideas apply to families $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ of random variables with $n > 2$. For example, the family $\mathbf{X}$ is called *independent* if

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n),$$

or, equivalently, $p_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} p_{X_i}(x_i)$, for $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. Furthermore, if $X_1, X_2, \ldots, X_n$ are independent, then

$$\mathbb{E}(X_1 X_2 \cdots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2) \cdots \mathbb{E}(X_n),$$

just as in Theorem 3.19. Finally, the family is called *pairwise independent* if X_i and X_j are independent whenever $i \neq j$. See Problem 3.6.2 for an example of pairwise-independent random variables that are not independent.

Exercise 3.23 Let X and Y be independent discrete random variables. Prove that

$$\mathbb{P}(X \geq x \text{ and } Y \geq y) = \mathbb{P}(X \geq x)\mathbb{P}(Y \geq y)$$

for all $x, y \in \mathbb{R}$.

Exercise 3.24 The *indicator function* of an event A is the function 1_A defined by

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Show that two events A and B are independent if and only if their indicator functions are independent random variables.

Exercise 3.25 If X and Y are independent discrete random variables, show that the two random variables $g(X)$ and $h(Y)$ are independent also, for any functions g and h which map $\mathbb{R}$ into $\mathbb{R}$.

3.4 Sums of random variables

Much of probability theory is concerned with sums of random variables, and so we need an answer to the following question: if X and Y are discrete random variables with a certain joint mass function, what is the mass function of $Z = X + Y$? Clearly, Z takes the value z if and only if $X = x$ and $Y = z - x$ for some value of x, and so

$$\begin{aligned} \mathbb{P}(Z = z) &= \mathbb{P}\left(\bigcup_x \big(\{X = x\} \cap \{Y = z - x\}\big)\right) \\ &= \sum_{x \in \operatorname{Im} X} \mathbb{P}(X = x,\ Y = z - x) \qquad \text{for } z \in \mathbb{R}. \end{aligned} \tag{3.26}$$

If X and Y are independent, their joint mass function factorizes, and we obtain the following result.

Theorem 3.27 (Convolution formula) *If X and Y are independent discrete random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, then $Z = X + Y$ has mass function*

$$\mathbb{P}(Z = z) = \sum_{x \in \operatorname{Im} X} \mathbb{P}(X = x)\mathbb{P}(Y = z - x) \qquad \text{for } z \in \mathbb{R}. \tag{3.28}$$

In the language of analysis, equation (3.28) says that the mass function of $X+Y$ is the *convolution* of the mass functions of X and Y. Formula (3.28) is rather inconvenient in practice, since it involves a summation. Soon we shall see a better way of treating sums of independent random variables.

Exercise 3.29 If X and Y are independent discrete random variables, X having the Poisson distribution with parameter λ and Y having the Poission distribution with parameter μ, show that $X + Y$ has the Poisson distribution with parameter $\lambda + \mu$. Give an example to show that the conclusion is not generally true if X and Y are dependent.

Exercise 3.30 If X has the binomial distribution with parameters m and p, Y has the binomial distribution with parameters n and p, and X and Y are independent, show that $X + Y$ has the binomial distribution with parameters $m + n$ and p.

Exercise 3.31 Show by induction that the sum of n independent random variables, each having the Bernoulli distribution with parameter p, has the binomial distribution with parameters n and p.

3.5 Indicator functions

Indicator functions have been encountered already in Exercises 2.10 and 3.24.

Definition 3.32 *The **indicator function** of an event A is the random variable denoted 1_A and given by*

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

The function 1_A *indicates* whether or not A occurs.[2] It is a discrete random variable with expectation given by

$$\mathbb{E}(1_A) = \mathbb{P}(A).$$

Indicator functions have two basic properties, namely,

$$1_{A \cap B} = 1_A 1_B, \tag{3.33}$$

$$1_A + 1_{A^c} = 1, \tag{3.34}$$

each of which is easily checked by considering the various possibilities for given $\omega \in \Omega$.

[2]Probability theory is based customarily on events, followed by random variables. By representing events via their indicator functions, one may rework the entire theory with random variables in the principal role, and with expectation taking the role of probability. See Whittle (2000).

Indicator functions provide a useful tool for calculating probabilities and expectations. Here is an elementary example. By (3.33)–(3.34),

$$\begin{aligned} 1_{A\cup B} &= 1 - 1_{A^c\cap B^c} \\ &= 1 - 1_{A^c}1_{B^c} = 1 - (1-1_A)(1-1_B) \\ &= 1_A + 1_B - 1_{A\cap B}. \end{aligned} \tag{3.35}$$

Now take expectations to deduce the standard fact that $\mathbb{P}(A\cup B) = \mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A\cap B)$.

Example 3.36 (Inclusion–exclusion formula) Let $A_1, A_2, \ldots, A_n$ be events, and let 1_{A_i} be the indicator function of A_i. As in (3.35), the union $A = A_1 \cup A_2 \cup \cdots \cup A_n$ has indicator function

$$1_A = 1 - \prod_{i=1}^{n}(1 - 1_{A_i}).$$

The product may be expanded and the terms grouped to obtain

$$\begin{aligned} 1_A = \sum_i 1_{A_i} - \sum_{i<j} 1_{A_i}1_{A_j} + \sum_{i<j<k} 1_{A_i}1_{A_j}1_{A_k} - \cdots \\ + (-1)^{n+1}1_{A_1}1_{A_2}\cdots 1_{A_n}. \end{aligned}$$

On taking expectations, we obtain the *inclusion–exclusion formula*

$$\begin{aligned} \mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j}\mathbb{P}(A_i\cap A_j) + \sum_{i<j<k}\mathbb{P}(A_i\cap A_j\cap A_k) - \cdots \\ + (-1)^{n+1}\mathbb{P}\left(\bigcap_i A_i\right) \end{aligned}$$

of Problem 1.11.14. △

Example 3.37 The $2n$ seats around a circular table are numbered clockwise. The guests at dinner form n king/queen pairs. The queens sit at random in the odd-numbered seats, with the kings at random between them. Let N be the number of queens sitting next to their king. Find the mean and variance of N.

Solution Let A_i be the event that the ith king/queen pair are seated adjacently. Then

$$N = \sum_{i=1}^{n} 1_{A_i}, \tag{3.38}$$

so that

$$\mathbb{E}(N) = \sum_{i=1}^{n}\mathbb{E}(1_{A_i}) = \sum_{i=1}^{n}\mathbb{P}(A_i) = n\mathbb{P}(A_1),$$

by symmetry. It is easily seen (by conditional probability, or simply by counting) that $\mathbb{P}(A_1) = 2/n$, and hence $\mathbb{E}(N) = n(2/n) = 2$ regardless of the value of n.

In order to find the variance, we should calculate $\mathbb{E}(N^2)$. By (3.38),

$$\mathbb{E}(N^2) = \mathbb{E}\left(\left[\sum_i 1_{A_i}\right]^2\right) = \mathbb{E}\left(\sum_i 1_{A_i}^2 + 2\sum_{i<j} 1_{A_i} 1_{A_j}\right). \tag{3.39}$$

Now $1_{A_i}^2 = 1_{A_i}$, since an indicator function takes only the values 0 and 1, and also $1_{A_i} 1_{A_j} = 1_{A_i \cap A_j}$. Therefore, by symmetry,

$$\mathbb{E}(N^2) = \mathbb{E}\left(\sum_i 1_{A_i} + 2\sum_{i<j} 1_{A_i \cap A_j}\right) = n\mathbb{P}(A_1) + n(n-1)\mathbb{P}(A_1 \cap A_2). \tag{3.40}$$

Using conditional probability,

$$\begin{aligned}\mathbb{P}(A_1 \cap A_2) &= \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1) \\ &= \frac{2}{n}\left(\frac{1}{n-1} \cdot \frac{1}{n-1} + \frac{n-2}{n-1} \cdot \frac{2}{n-1}\right) = \frac{2(2n-3)}{n(n-1)^2},\end{aligned} \tag{3.41}$$

where the two terms correspond to whether or not the second queen sits next to the first couple. By (3.39)–(3.41),

$$\mathbb{E}(N^2) = 2 + n(n-1) \cdot \frac{2(2n-3)}{n(n-1)^2},$$

and hence

$$\operatorname{var}(N) = \mathbb{E}(N^2) - \mathbb{E}(N)^2 = \frac{2(n-2)}{n-1}. \qquad \triangle$$

Exercise 3.42 Let N be the number of the events $A_1, A_2, \ldots, A_n$ which occur. Show that[3]

$$\mathbb{E}(N) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

3.6 Problems

1. Let X and Y be independent discrete random variables, each having mass function given by

$$\mathbb{P}(X = k) = \mathbb{P}(Y = k) = pq^k \qquad \text{for } k = 0, 1, 2 \ldots,$$

where $0 < p = 1 - q < 1$. Show that

$$\mathbb{P}(X = k \mid X + Y = n) = \frac{1}{n+1} \qquad \text{for } k = 0, 1, 2 \ldots, n.$$

[3] A similar fact is valid for an *infinite* sequence $A_1, A_2, \ldots$, namely that the mean number of events that occur is $\sum_{i=1}^{\infty} \mathbb{P}(A_i)$. This is, however, harder to prove. See the footnote on p. 40.

2. Independent random variables U and V each take the values -1 or 1 only, and

$$\mathbb{P}(U = 1) = a, \qquad \mathbb{P}(V = 1) = b,$$

where $0 < a, b < 1$. A third random variable W is defined by $W = UV$. Show that there are unique values of a and b such that U, V, and W are pairwise independent. For these values of a and b, are U, V, and W independent? Justify your answer. (Oxford 1971F)

3. If X and Y are discrete random variables, each taking only two distinct values, prove that X and Y are independent if and only if $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

4. Let $X_1, X_2, \ldots, X_n$ be independent discrete random variables, each having mass function

$$\mathbb{P}(X_i = k) = \frac{1}{N} \qquad \text{for } k = 1, 2, \ldots, N.$$

Find the mass functions of U_n and V_n, given by

$$U_n = \min\{X_1, X_2, \ldots, X_n\}, \qquad V_n = \max\{X_1, X_2, \ldots, X_n\}.$$

5. Let X and Y be independent discrete random variables, X having the geometric distribution with parameter p and Y having the geometric distribution with parameter r. Show that $U = \min\{X, Y\}$ has the geometric distribution with parameter $p + r - pr$.

6. Hugo's bowl of spaghetti contains n strands. He selects two ends at random and joins them. He does this until no ends are left. What is the expected number of spaghetti hoops in his bowl?

7. Let $X_1, X_2, \ldots$ be discrete random variables, each having mean μ, and let N be a random variable which takes values in the non-negative integers and which is independent of the X_i. By conditioning on the value of N, show that

$$\mathbb{E}(X_1 + X_2 + \cdots + X_N) = \mu\mathbb{E}(N).$$

8. Let $X_1, X_2, \ldots$ be independent, identically distributed random variables, and $S_n = X_1 + X_2 + \cdots + X_n$. Show that $\mathbb{E}(S_m/S_n) = m/n$ if $m \leq n$, and $\mathbb{E}(S_m/S_n) = 1 + (m - n)\mu\mathbb{E}(1/S_n)$ if $m > n$, where $\mu = \mathbb{E}(X_1)$. You may assume that all the expectations are finite.

9. The random variables U and V each take the values ± 1. Their joint distribution is given by

$$\begin{aligned}
\mathbb{P}(U = +1) = \mathbb{P}(U = -1) &= \tfrac{1}{2}, \\
\mathbb{P}(V = +1 \mid U = 1) = \tfrac{1}{3} &= \mathbb{P}(V = -1 \mid U = -1), \\
\mathbb{P}(V = -1 \mid U = 1) = \tfrac{2}{3} &= \mathbb{P}(V = +1 \mid U = -1).
\end{aligned}$$

   (a) Find the probability that $x^2 + Ux + V = 0$ has at least one real root.
   (b) Find the expected value of the larger root, given that there is at least one real root.
   (c) Find the probability that $x^2 + (U + V)x + U + V = 0$ has at least one real root.

   (Oxford 1980M)

10. A number N of balls are thrown at random into M boxes, with multiple occupancy permitted. Show that the expected number of empty boxes is $(M - 1)^N / M^{N-1}$.

11. We are provided with a coin which comes up heads with probability p at each toss. Let $v_1, v_2, \ldots, v_n$ be n distinct points on a unit circle. We examine each unordered pair v_i, v_j in turn and toss the coin; if it comes up heads, we join v_i and v_j by a straight line segment (called an *edge*), otherwise we do nothing. The resulting network is called a *random graph*.

Prove that

(a) the expected number of edges in the random graph is $\frac{1}{2}n(n-1)p$,
(b) the expected number of triangles (triples of points each pair of which is joined by an edge) is $\frac{1}{6}n(n-1)(n-2)p^3$.

12. *Coupon-collecting problem*. There are c different types of coupon, and each coupon obtained is equally likely to be any one of the c types. Let Y_i be the additional number of coupons collected, after obtaining i distinct types, before a new type is collected. Show that Y_i has the geometric distribution with parameter $(c-i)/c$, and deduce the mean number of coupons you will need to collect before you have a complete set.

13. In Problem 3.6.12 above, find the expected number of different types of coupon in the first n coupons received.

14. Each time you flip a certain coin, heads appears with probability p. Suppose that you flip the coin a random number N of times, where N has the Poisson distribution with parameter λ and is independent of the outcomes of the flips. Find the distributions of the numbers X and Y of resulting heads and tails, respectively, and show that X and Y are independent.

15. Let $(Z_n : 1 \le n < \infty)$ be a sequence of independent, identically distributed random variables with
$$\mathbb{P}(Z_n = 0) = q, \qquad \mathbb{P}(Z_n = 1) = p,$$
where $p + q = 1$. Let A_i be the event that $Z_i = 0$ and $Z_{i-1} = 1$. If U_n is the number of times A_i occurs for $2 \le i \le n$, prove that $\mathbb{E}(U_n) = (n-1)pq$, and find the variance of U_n. (Oxford 1977F)

16. I throw two dice and record the scores S_1 and S_2. Let X be the sum $S_1 + S_2$ and Y the difference $S_1 - S_2$.

(a) Suppose the dice are fair, so that the values $1, 2, \ldots, 6$ are equally likely. Calculate the mean and variance of both X and Y. Find all the values of x and y at which the probabilities $\mathbb{P}(X = x)$, $\mathbb{P}(Y = y)$ are each either greatest or least. Determine whether the random variables X and Y are independent.
(b) Now suppose the dice give the values $1, 2, \ldots, 6$ with probabilities $p_1, p_2, \ldots, p_6$ and $q_1, q_2, \ldots, q_6$, respectively. Write down the values of $\mathbb{P}(X = 2)$, $\mathbb{P}(X = 7)$, and $\mathbb{P}(X = 12)$. By comparing $\mathbb{P}(X = 7)$ with $\sqrt{\mathbb{P}(X = 2)\mathbb{P}(X = 12)}$ and applying the arithmetic/geometric mean inequality,[4] or otherwise, show that X cannot be uniformly distributed on the set $\{2, 3, \ldots, 12\}$.

(Cambridge 2009)

[4] See the forthcoming Example 7.70 also.

4

Probability generating functions

Summary. Generating functions provide a powerful tool for studying random variables that take integer values. The moments of a random variable are defined, and it is shown how they may be derived from its generating function. Generating functions are especially useful in understanding sums of random variables. The chapter ends with an account of the random sum formula.

4.1 Generating functions

One way to record a sequence $u_0, u_1, u_2, \ldots$ of real numbers is to write down a general formula for the nth term u_n. Another is to write down the *generating function* of the sequence, defined to be the sum of the power series

$$u_0 + u_1 s + u_2 s^2 + \cdots . \tag{4.1}$$

For example, the sequence $1, 2, 4, 8, \ldots$ has generating function

$$1 + 2s + 4s^2 + \cdots = \sum_{n=0}^{\infty} (2s)^n = \frac{1}{1-2s},$$

valid whenever $|s| < \frac{1}{2}$. Similarly the sequence $1, 2, 3, \ldots$ has generating function

$$1 + 2s + 3s^2 + \cdots = \sum_{n=0}^{\infty} (n+1)s^n = \frac{1}{(1-s)^2}, \tag{4.2}$$

valid whenever $|s| < 1$. Such generating functions are useful ways of dealing with real sequences since they specify the sequence uniquely. That is to say, given the real sequence $u_0, u_1, u_2, \ldots$, we may find its generating function (4.1); conversely if the generating function $U(s)$ has a convergent Taylor series

$$U(s) = u_0 + u_1 s + u_2 s^2 + \cdots$$

for all small s, then this expansion is unique, and so $U(s)$ generates the sequence $u_0, u_1, u_2, \ldots$ only.

One may define other types of generating functions also. For example, the *exponential generating function* of the real sequence $u_0, u_1, u_2, \ldots$ is defined to be the sum of the power series

$$u_0 + u_1 s + \frac{1}{2!}u_2 s^2 + \frac{1}{3!}u_3 s^3 + \cdots$$

whenever this series converges. We do not consider such generating functions in this chapter, but shall return to them in Chapter 7.

When dealing with generating functions of real sequences, it is important that the underlying power series converges for certain $s \neq 0$, but so long as this is the case we will not normally go to the length of saying for which values of s the power series is absolutely convergent. For example, we say that $(1 - s)^{-2}$ is the generating function of the sequence $1, 2, 3, \ldots$ without explicit reference to the fact that the series in (4.2) converges absolutely only if $|s| < 1$.

Example 4.3 The sequence given by

$$u_n = \begin{cases} \binom{N}{n} & \text{if } n = 0, 1, 2, \ldots, N \\ 0 & \text{otherwise} \end{cases}$$

has generating function

$$U(s) = \sum_{n=0}^{N} \binom{N}{n} s^n = (1 + s)^N. \qquad \triangle$$

Exercise 4.4 If $u_0, u_1, \ldots$ has generating function $U(s)$ and $v_0, v_1, \ldots$ has generating function $V(s)$, find $V(s)$ in terms of $U(s)$ when (a) $v_n = 2u_n$, (b) $v_n = u_n + 1$, (c) $v_n = nu_n$.

Exercise 4.5 Let $0 < p = 1 - q < 1$. Of which sequence is $U(s) = \sqrt{1 - 4pqs^2}$ the generating function?

4.2 Integer-valued random variables

Many random variables of interest take values in the set of non-negative integers (all the examples in Section 2.2 are of this form). We may think of the mass function of such a random variable X as a sequence $p_0, p_1, p_2, \ldots$ of numbers, where

$$p_k = \mathbb{P}(X = k) \qquad \text{for } k = 0, 1, 2, \ldots,$$

satisfying

$$p_k \geq 0 \text{ for all } k, \quad \text{and} \quad \sum_{k=0}^{\infty} p_k = 1. \tag{4.6}$$

Definition 4.7 *The **probability generating function** (or **pgf**) of X is the function $G_X(s)$ defined by*

$$G_X(s) = p_0 + p_1 s + p_2 s^2 + \cdots, \tag{4.8}$$

for all values of s for which the right-hand side converges absolutely.

In other words, the probability generating function $G_X(s)$ of X is the generating function of the sequence $p_0, p_1, \ldots$. From (4.6) and (4.8) we see that

$$G_X(0) = p_0 \qquad \text{and} \qquad G_X(1) = 1, \tag{4.9}$$

and, by Theorem 2.29,

$$G_X(s) = \mathbb{E}(s^X) \tag{4.10}$$

whenever this expectation exists. It is immediate that $G_X(s)$ exists for all values of s satisfying $|s| \leq 1$, since in this case,[1]

$$\sum_{k=0}^{\infty} |p_k s^k| \leq \sum_{k=0}^{\infty} p_k = 1. \tag{4.11}$$

Example 4.12 Let X be a random variable having the geometric distribution with parameter p. Then

$$\mathbb{P}(X = k) = pq^{k-1} \qquad \text{for } k = 1, 2, 3, \ldots,$$

where $p + q = 1$, and X has probability generating function

$$\begin{aligned} G_X(s) &= \sum_{k=1}^{\infty} pq^{k-1} s^k \\ &= ps \sum_{k=0}^{\infty} (qs)^k = \frac{ps}{1 - qs} \qquad \text{if } |s| < q^{-1}. \end{aligned}$$

△

A crucially important property of probability generating functions is the following uniqueness theorem.

Theorem 4.13 (Uniqueness theorem for probability generating functions) *Suppose X and Y have probability generating functions G_X and G_Y, respectively. Then*

$$G_X(s) = G_Y(s) \qquad \textit{for all } s$$

if and only if

$$\mathbb{P}(X = k) = \mathbb{P}(Y = k) \qquad \textit{for } k = 0, 1, 2, \ldots.$$

In other words, integer-valued random variables have the same probability generating function if and only if they have the same mass function.

[1] Regarded as a power series, G_X has radius of convergence at least 1.

Proof We need only show that $G_X = G_Y$ implies that $\mathbb{P}(X = k) = \mathbb{P}(Y = k)$ for all k. By (4.11), G_X and G_Y have radii of convergence at least 1, and therefore they have unique power series expansions about the origin:

$$G_X(s) = \sum_{k=0}^{\infty} s^k \mathbb{P}(X = k), \qquad G_Y(s) = \sum_{k=0}^{\infty} s^k \mathbb{P}(Y = k).$$

If $G_X = G_Y$, these two power series have identical coefficients. □

We saw in Example 4.12 that a random variable with the geometric distribution, parameter p, has probability generating function $ps(1 - qs)^{-1}$, where $p + q = 1$. Only by an appeal to the above theorem can we deduce the converse: if X has probability generating function $ps(1 - qs)^{-1}$, then X has the geometric distribution with parameter p.

Here is a list of some common probability generating functions. Let $p = 1 - q \in [0, 1]$.

Bernoulli distribution. If X has the Bernoulli distribution with parameter p, then

$$G_X(s) = q + ps. \tag{4.14}$$

Binomial distribution. If X has the binomial distribution with parameters n and p, then

$$G_X(s) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} s^k = (q + ps)^n. \tag{4.15}$$

Poisson distribution. If X has the Poisson distribution with parameter λ, then

$$G_X(s) = \sum_{k=0}^{\infty} \frac{1}{k!} \lambda^k e^{-\lambda} s^k = e^{\lambda(s-1)}. \tag{4.16}$$

Negative binomial distribution. If X has the negative binomial distribution with parameters n and p, then

$$G_X(s) = \sum_{k=n}^{\infty} \binom{k-1}{n-1} p^n q^{k-n} s^k = \left(\frac{ps}{1 - qs} \right)^n \qquad \text{if } |s| < q^{-1}. \tag{4.17}$$

We have used the negative binomial expansion here, see Theorem A.3.

There are two principal reasons why it is often more convenient to work with probability generating functions than with mass functions, and we discuss these in the next two sections.

Exercise 4.18 If X is a random variable with probability generating function $G_X(s)$, and k is a positive integer, show that $Y = kX$ and $Z = X + k$ have probability generating functions

$$G_Y(s) = G_X(s^k), \qquad G_Z(s) = s^k G_X(s).$$

Exercise 4.19 If X is uniformly distributed on $\{0, 1, 2, \ldots, a\}$, in that

$$\mathbb{P}(X = k) = \frac{1}{a+1} \qquad \text{for } k = 0, 1, 2, \ldots, a,$$

show that X has probability generating function

$$G_X(s) = \frac{1 - s^{a+1}}{(a+1)(1-s)}.$$

4.3 Moments

For any discrete random variable X, the mean value $\mathbb{E}(X)$ is an indication of the 'centre' of the distribution of X. This is only the first of a collection of numbers containing information about the distribution of X, the whole collection being the sequence $\mathbb{E}(X), \mathbb{E}(X^2), \mathbb{E}(X^3), \ldots$ of means of powers of X. These numbers are called the *moments* of X.

Definition 4.20 *Let $k \geq 1$. The kth **moment** of the random variable X is the quantity $\mathbb{E}(X^k)$.*

Possibly the two most important quantities which arise from the moments of X are the mean $\mathbb{E}(X)$ of X, and the variance of X, defined in (2.33) to be

$$\operatorname{var}(X) = \mathbb{E}\big([X - \mathbb{E}(X)]^2\big). \tag{4.21}$$

To see the relationship between $\operatorname{var}(X)$ and the moments of X, just note that

$$\begin{aligned} \operatorname{var}(X) &= \mathbb{E}\big(X^2 - 2X\mathbb{E}(X) + \mathbb{E}(X)^2\big) \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2 \qquad \text{by (3.11)} \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2, \end{aligned} \tag{4.22}$$

in agreement with (2.35).

If X is a random variable with values in the non-negative integers, the moments of X are easily found from the probability generating function of X by calculating the derivatives of this function at the point $s = 1$. The basic observation is as follows.

Theorem 4.23 *Let X be a random variable with probability generating function $G_X(s)$. The rth derivative of $G_X(s)$ at $s = 1$ equals $\mathbb{E}\big(X[X-1]\cdots[X-r+1]\big)$ for $r = 1, 2, \ldots$. That is to say,*

$$G_X^{(r)}(1) = \mathbb{E}\big(X[X-1]\cdots[X-r+1]\big). \tag{4.24}$$

Proof To see this when $r = 1$, we use the following non-rigorous argument:

$$\begin{aligned} G'_X(s) &= \frac{d}{ds}\sum_{k=0}^{\infty} s^k \mathbb{P}(X = k) \\ &= \sum_{k=0}^{\infty} \frac{d}{ds} s^k \mathbb{P}(X = k) = \sum_{k=0}^{\infty} k s^{k-1} \mathbb{P}(X = k) \end{aligned} \tag{4.25}$$

so that

$$G'_X(1) = \sum_{k=0}^{\infty} k\mathbb{P}(X = k) = \mathbb{E}(X)$$

as required. A similar argument holds for the rth derivative of $G_X(s)$ at $s = 1$. The difficulty in (4.25) is to justify the interchange of the differential operator and the summation, but this may be shown to be valid[2] if $|s| < 1$, and then Abel's lemma[3] enables us to conclude that (4.25) is correct. □

It is easy to see how to calculate the moments of X from (4.24). For example

$$\mathbb{E}(X) = G'_X(1), \tag{4.26}$$

and

$$\begin{aligned} \mathbb{E}(X^2) &= \mathbb{E}\big(X[X-1] + X\big) \\ &= \mathbb{E}\big(X[X-1]\big) + \mathbb{E}(X) \\ &= G''_X(1) + G'_X(1), \end{aligned} \tag{4.27}$$

and similarly, by (4.22),

$$\operatorname{var}(X) = G''_X(1) + G'_X(1) - G_X(1)^2. \tag{4.28}$$

Example 4.29 Let X have the geometric distribution with parameter $p \in (0, 1)$. It has probability generating function $G_X(s) = ps(1 - qs)^{-1}$ for $|s| < q^{-1}$, where $p + q = 1$. Hence

$$\begin{aligned} \mathbb{E}(X) &= G'_X(1) = \frac{1}{p}, \\ \mathbb{E}(X^2) &= G''_X(1) + G'_X(1) = \frac{q+1}{p^2}, \\ \operatorname{var}(X) &= \frac{q}{p^2}, \end{aligned}$$

in agreement with the calculations of Example 2.36. △

[2] Recall that G_X is a power series with radius of convergence at least 1.

[3] Abel's lemma is a classical result of real analysis. It says that if $u_0, u_1, \ldots$ is a real non-negative sequence such that the power series $\sum_{k=0}^{\infty} u_k s^k$ converges with sum $U(s)$ if $|s| < 1$, then $\sum_{k=0}^{\infty} u_k = \lim_{s \uparrow 1} U(s)$, where we allow the possibility that both sides equal $+\infty$.

Exercise 4.30 Use the method of generating functions to show that a random variable having the Poisson distribution, parameter λ, has both mean and variance equal to λ.

Exercise 4.31 If X has the negative binomial distribution with parameters n and p, show that

$$\mathbb{E}(X) = n/p, \qquad \text{var}(X) = nq/p^2,$$

where $q = 1 - p$.

Exercise 4.32 Let X be a random variable taking values in the finite set $\{1, 2, \ldots, N\}$. The *Dirichlet* probability generating function of X is defined as the function $\Delta(s) = \mathbb{E}(X^{-s})$. Express the mean of X in terms of Δ.

Similarly, express the mean of $\log X$ in terms of Δ. You may find it useful to recall that $(x^y - 1)/y \to \log x$ as $y \to 0$.

4.4 Sums of independent random variables

Much of probability theory is concerned with sums of independent random variables, and we need a way of dealing with such sums. The convolution formula of Theorem 3.27 is usually inconvenient, since $n - 1$ convolutions are required to find the mass function of the sum of n independent random variables, and each such operation can be rather complicated. It is in this respect that probability generating functions are a powerful tool.

Theorem 4.33 *If X and Y are independent random variables, each taking values in the set $\{0, 1, 2, \ldots\}$, then their sum has probability generating function*

$$G_{X+Y}(s) = G_X(s)G_Y(s). \tag{4.34}$$

Proof We have that

$$\begin{aligned} G_{X+Y}(s) &= \mathbb{E}(s^{X+Y}) = \mathbb{E}(s^X s^Y) \\ &= \mathbb{E}(s^X)\mathbb{E}(s^Y) \qquad \text{by Theorem 3.20} \\ &= G_X(s)G_Y(s). \end{aligned}$$

□

It follows that the sum $S_n = X_1 + X_2 + \cdots + X_n$ of n independent random variables, each taking values in $\{0, 1, 2, \ldots\}$, has probability generating function given by

$$G_{S_n}(s) = G_{X_1}(s)G_{X_2}(s)\cdots G_{X_n}(s). \tag{4.35}$$

We shall make much use of this formula. An important extension of (4.35) deals with the sum of a random number of independent random variables.

Theorem 4.36 (Random sum formula) *Let N and $X_1, X_2, \ldots$ be independent random variables, each taking values in $\{0, 1, 2, \ldots\}$. If the X_i are identically distributed with common probability generating function G_X, then the sum*

$$S = X_1 + X_2 + \cdots + X_N$$

has probability generating function

$$G_S(s) = G_N(G_X(s)). \tag{4.37}$$

Proof We use the partition theorem, Theorem 2.42, with the events $B_n = \{N = n\}$ to find that

$$\begin{aligned}
G_S(s) &= \mathbb{E}(s^{X_1+\cdots+X_N}) \\
&= \sum_{n=0}^{\infty} \mathbb{E}\big(s^{X_1+\cdots+X_N} \,\big|\, N = n\big)\mathbb{P}(N = n) && \text{by Theorem 2.42} \\
&= \sum_{n=0}^{\infty} \mathbb{E}(s^{X_1+\cdots+X_n})\mathbb{P}(N = n) \\
&= \sum_{n=0}^{\infty} G_X(s)^n \mathbb{P}(N = n) && \text{by (4.35)} \\
&= G_N(G_X(s))
\end{aligned}$$

by the definition of G_N. □

Formula (4.37) enables us to say quite a lot about such a random sum. For example, to find the mean vaue of S, in the notation of Theorem 4.36, we merely calculate $G'_S(1)$ as follows. By (4.37),

$$G'_S(s) = G'_N(G_X(s))G'_X(s).$$

Now set $s = 1$ to obtain

$$G'_S(1) = G'_N(G_X(1))G'_X(1) = G'_N(1)G'_X(1),$$

since $G_X(1) = 1$. By (4.26),

$$\mathbb{E}(S) = G'_S(1) = \mathbb{E}(N)\mathbb{E}(X), \tag{4.38}$$

where $\mathbb{E}(X)$ is the mean of a typical X_i.

Example 4.39 One evening, the hutch in the garden contains 20 pregnant rabbits. The hutch is insecure and each rabbit has a chance of $\frac{1}{2}$ of escaping overnight. The next morning, each remaining rabbit gives birth to a litter, with each mother having a random number of offspring with the Poisson distribution, parameter 3 (this is a very unlikely tale). Assuming as much

independence as necessary, determine the probability generating function of the total number of offspring.

Solution The number N of rabbits who do not escape has the binomial distribution with parameters 20 and $\frac{1}{2}$, and consequently N has probability generating function

$$G_N(s) = \mathbb{E}(s^N) = \left(\tfrac{1}{2} + \tfrac{1}{2}s\right)^{20}.$$

Let X_i be the number of offspring of the ith of these rabbits. Each X_i has the Poisson distribution with probability generating function

$$G_X(s) = e^{3(s-1)}.$$

Assuming that N and the X_i are independent, we conclude from the random sum formula (4.37) that the total number $S = X_1 + X_2 + \cdots + X_N$ of offspring has probability generating function

$$G_S(s) = G_N(G_X(s)) = \left(\tfrac{1}{2} + \tfrac{1}{2}e^{3(s-1)}\right)^{20}. \qquad \triangle$$

Exercise 4.40 Use Theorem 4.33 to show that the sum of two independent random variables, having the Poisson distribution with parameters λ and μ respectively, has the Poisson distribution also, with parameter $\lambda + \mu$. Compare your solution to that of Exercise 3.29.

Exercise 4.41 Use generating functions to find the distribution of $X + Y$, where X and Y are independent random variables, X having the binomial distribution with parameters m and p, and Y having the binomial distribution with parameters n and p. Deduce that the sum of n independent random variables, each having the Bernoulli distribution with parameter p, has the binomial distribution with parameters n and p.

Exercise 4.42 Each egg laid by a hen falls onto the concrete floor of the henhouse and cracks with probability p. If the number of eggs laid today by the hen has the Poisson distribution, parameter λ, use generating functions to show that the number of uncracked eggs has the Poisson distribution with parameter $\lambda(1 - p)$.

4.5 Problems

1. Let X have probability generating function $G_X(s)$ and let $u_n = \mathbb{P}(X > n)$. Show that the generating function $U(s)$ of the sequence $u_0, u_1, \ldots$ satisfies

$$(1 - s)U(s) = 1 - G_X(s),$$

whenever the series defining these generating functions converge.

2. A symmetrical die is thrown independently seven times. What is the probability that the total number of points obtained is 14? (Oxford 1974M)

3. Three players, Alan, Bob, and Cindy, throw a perfect die in turn independently in the order $A, B, C, A, \ldots$ until one wins by throwing a 5 or a 6. Show that the probability generating function $F(s)$ for the random variable X which takes the value r if the game ends on the rth throw can be written as

$$F(s) = \frac{9s}{27 - 8s^3} + \frac{6s^2}{27 - 8s^3} + \frac{4s^3}{27 - 8s^3}.$$

Hence find the probabilities of winning for Alan, Bob, and Cindy. Find the mean duration of the game. (Oxford 1973M)

4. A player undertakes trials, and the probability of success at each trial is p. A turn consists of a sequence of trials up to the first failure. Obtain the probability generating function for the total number of successes in N turns. Show that the mean of this distribution is $Np(1-p)^{-1}$ and find its variance. (Oxford 1974M)

5. Each year a tree of a particular type flowers once, and the probability that it has n flowers is $(1-p)p^n$, $n = 0, 1, 2, \ldots$, where $0 < p < 1$. Each flower has probability $\frac{1}{2}$ of producing a ripe fruit, independently of all other flowers. Find the probability that in a given year

   (a) the tree produces r ripe fruits,
   (b) the tree had n flowers, given that it produces r ripe fruits.

   (Oxford 1982M)

6. An unfair coin is tossed n times, each outcome is independent of all the others, and on each toss a head is shown with probability p. The total number of heads shown is X. Use the probability generating function of X to find

   (a) the mean and variance of X,
   (b) the probability that X is even,
   (c) the probability that X is divisible by 3.

   (Oxford 1980M)

7. Let X and Y be independent random variables having Poisson distributions with parameters λ and μ, respectively. Prove that $X + Y$ has a Poisson distribution and that $\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y)$. Find the conditional probability $\mathbb{P}(X = k \mid X + Y = n)$ for $0 \le k \le n$, and hence show that the conditional expectation of X given that $X + Y = n$, that is,

$$\mathbb{E}\big(X \,\big|\, X + Y = n\big) = \sum_{k=0}^{\infty} k\mathbb{P}\big(X = k \,\big|\, X + Y = n\big),$$

   is $n\lambda/(\lambda+\mu)$. (Oxford 1983M)

* 8. A fair coin is tossed a random number N of times, giving a total of X heads and Y tails. You showed in Problem 3.6.14 that X and Y are independent if N has the Poisson distribution. Use generating functions to show that the converse is valid too: if X and Y are independent and the generating function $G_N(s)$ of N is assumed to exist for values of s in a neighbourhood of $s = 1$, then N has the Poisson distribution.

9. *Coupon-collecting problem.* Each packet of a certain breakfast cereal contains one token, coloured either red, blue, or green. The coloured tokens are distributed randomly among the packets, each colour being equally likely. Let X be the random variable which takes the value j when I find my first red token in the jth packet which I open. Obtain the probability generating function of X, and hence find its expectation.

   More generally, suppose that there are tokens of m different colours, all equally likely. Let Y be the random variable which takes the value j when I first obtain a full set, of at least one token of each colour, when I open my jth packet. Find the generating function of Y, and show that its expectation is $m\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m}\right)$. (Oxford 1985M)

10. Define the mean value of a discrete random variable and the probability generating function ϕ. Show that the mean value is $\phi'(1)$. If $\phi(s)$ has the form $p(s)/q(s)$ show that the mean value is $(p'(1) - q'(1))/q(1)$.

    Two duellists, A and B, fire at each other in turn until one hits the other. Each duellist has the same probability of obtaining a hit with each shot fired, these probabilities being a for A and b for B. If A fires the first shot, calculate the probability that A wins the duel. Find also the probability distribution of the number of shots fired before the duel terminates. What is the expected number of shots fired? (Oxford 1976M)

11. There is a random number N of foreign objects in my soup, with mean μ and finite variance. Each object is a fly with probability p, and otherwise a spider; different objects have independent types. Let F be the number of flies and S the number of spiders.

    (a) Show that $G_F(s) = G_N(ps + 1 - p)$. [You should present a clear statement of any general result used.]
    (b) Suppose N has the Poisson distribution with parameter μ. Show that F has the Poisson distribution with parameter μp, and that F and S are independent.
    (c) Let $p = \frac{1}{2}$ and suppose F and S are independent. [You are given nothing about the distribution of N.] Show that $G_N(s) = G_N\left(\frac{1}{2}[1 + s]\right)^2$. By working with the function $H(s) = G_N(1 - s)$ or otherwise, deduce that N has a Poisson distribution.

    You may assume that $\left[1 + (x/n) + o(n^{-1})\right]^n \to e^x$ as $n \to \infty$. (Cambridge 2002)

5

Distribution functions and density functions

Summary. One of the basic objects associated with a random variable is its distribution function, which summarizes the probabilities of different values. Continuous random variables are defined, together with their probability density functions. There is an account of how to find the density of a function of a continuous random variable, and several examples are presented of the use of density functions in practice. The final section is devoted to three problems in geometrical probability, namely Bertrand's paradox, Buffon's needle, and the problem of stick breaking.

5.1 Distribution functions

Discrete random variables may take only *countably* many values. This condition is too restrictive for many situations, and accordingly we make a broader definition[1]: a *random variable* X on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a mapping $X : \Omega \to \mathbb{R}$ such that

$$\big\{\omega \in \Omega : X(\omega) \leq x\big\} \in \mathcal{F} \qquad \text{for all } x \in \mathbb{R}. \tag{5.1}$$

Henceforth, we abbreviate events of the form $\{\omega \in \Omega : X(\omega) \leq x\}$ to the simpler expression $\{X \leq x\}$.

We require that random variables satisfy (5.1) for very much the same reason as we required (2.2) for discrete random variables. That is, we are interested in the values taken by a random variable X, and the probabilities associated with these values. A convenient way to do this is to fix $x \in \mathbb{R}$ and ask for the probability that X takes a value in the interval $(-\infty, x]$. This probability exists only if its inverse image $X^{-1}\big((-\infty, x]\big) = \{X \leq x\}$ lies in the event space $\mathcal{F}$, and so we postulate that this holds for all $x \in \mathbb{R}$. Note that every discrete random variable X is a random variable. To see this, observe that if X is discrete, then

$$\{X \leq x\} = \bigcup_{y \in \operatorname{Im} X:\, y \leq x} \{X = y\},$$

which is the countable union of events in $\mathcal{F}$ and therefore belongs to $\mathcal{F}$.

Whereas discrete random variables were studied via their *mass* functions, random variables in the broader sense are studied via their *distribution* functions, defined as follows.

[1] If (5.1) holds, the function X is said to be $\mathcal{F}$*-measurable*.

Definition 5.2 *If X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, the* ***distribution function***[2] *of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by*

$$F_X(x) = \mathbb{P}(X \leq x). \tag{5.3}$$

Example 5.4 Suppose that X is a discrete random variable taking non-negative integer values, with mass function

$$\mathbb{P}(X = k) = p_k \qquad \text{for } k = 0, 1, 2, \ldots.$$

For $x \in \mathbb{R}$, it is the case that $X \leq x$ if and only if X takes one of the values $0, 1, 2, \ldots, \lfloor x \rfloor$, where $\lfloor x \rfloor$ denotes the greatest integer not greater than x. Hence

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ p_0 + p_1 + \cdots + p_{\lfloor x \rfloor} & \text{if } x \geq 0, \end{cases}$$

and a sketch of this function is displayed in Figure 5.1. △

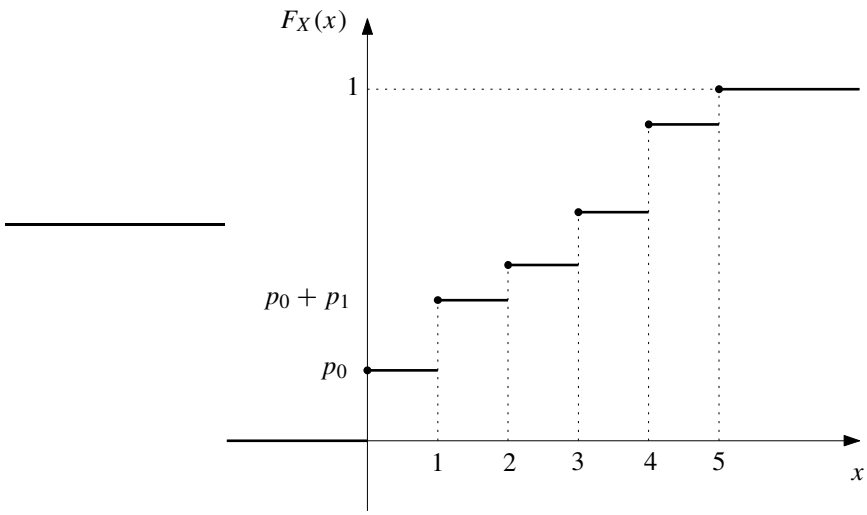


Fig. 5.1 The distribution function of a random variable that takes values in the set $\{0, 1, 2, 3, 4, 5\}$.

The distribution function F_X has various general and elementary properties, the first of which is

$$F_X(x) \leq F_X(y) \qquad \text{if } x \leq y, \tag{5.5}$$

which is to say that F_X is monotonic non-decreasing. This holds because

[2]Sometimes referred to as the *cumulative* distribution function of X.

$$\{X \leq x\} \subseteq \{X \leq y\}$$

whenever $x \leq y$, since if X takes a value not exceeding x, then this value cannot exceed y. Other elementary properties of $F_X(x)$ concern its behaviour when x is near $-\infty$ or $+\infty$. It is intuitively clear that

$$F_X(x) \to 0 \qquad \text{as } x \to -\infty, \tag{5.6}$$

$$F_X(x) \to 1 \qquad \text{as } x \to \infty, \tag{5.7}$$

since in the first case, as $x \to -\infty$ the event that X is smaller than x becomes less and less likely, whilst in the second case, as $x \to \infty$ this event becomes overwhelmingly likely. At an intuitive level, (5.6) and (5.7) are obvious, since they resemble the trivial remarks

$$\mathbb{P}(X \leq -\infty) = 0, \qquad \mathbb{P}(X \leq \infty) = 1,$$

but a formal verification of (5.6) and (5.7) relies on the continuity of $\mathbb{P}$, Theorem 1.54.

In the same way, Theorem 1.54 is needed to prove the third general property of distribution functions:

$$F_X \text{ is continous from the right,} \tag{5.8}$$

which is to say that[3]

$$F_X(x + \epsilon) \to F_X(x) \qquad \text{as } \epsilon \downarrow 0. \tag{5.9}$$

A glance at Figure 5.1 confirms that distribution functions need not be continuous from the left. Properties (5.5)–(5.8) characterize distribution functions completely, in the sense that if F is a function which satisfies (5.5)–(5.8), there exists a probability space and a random variable X on this space such that X has distribution function F. The proof is omitted, but this fact should be noted since, in many circumstances, it allows us to avoid the rather tedious business of writing down probability spaces and random variables explicitly.

Before we give examples of distribution functions, here is a final property. The probability $F_X(x) = \mathbb{P}(X \leq x)$ is the probability that X takes a value in the infinite interval $(-\infty, x]$. To find the probability that X takes a value in a bounded interval $(a, b]$, we proceed in the following way. For $a < b$,

$$\begin{aligned} \mathbb{P}(a < X \leq b) &= \mathbb{P}\big(\{X \leq b\} \setminus \{X \leq a\}\big) \\ &= \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) \end{aligned}$$

since the event $\{X \leq a\}$ is a subset of the event $\{X \leq b\}$. Hence

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a). \tag{5.10}$$

Exercise 5.11 Let X be a random variable taking integer values such that $\mathbb{P}(X = k) = p_k$ for $k = \ldots, -1, 0, 1, \ldots$. Show that the distribution function of X satisfies

$$F_X(b) - F_X(a) = p_{a+1} + p_{a+2} + \cdots + p_b$$

for all integers a, b with $a < b$.

[3]The limit in (5.9) is taken as ϵ tends down to 0 through positive values only.

Exercise 5.12 If X is a random variable and c is a real number such that $\mathbb{P}(X = c) > 0$, show that the distribution function $F_X(x)$ of X is discontinuous at the point $x = c$. Is the converse true?

Exercise 5.13 Express the distribution function of $Y = \max\{0, X\}$ in terms of the distribution function F_X of X.

Exercise 5.14 The real number m is called a *median* of the random variable X if

$$\mathbb{P}(X < m) \leq \tfrac{1}{2} \leq \mathbb{P}(X \leq m).$$

Show that every random variable has at least one median.

5.2 Examples of distribution functions

Example 5.4 contains our first example of a distribution function. Note the general features of this function: non-decreasing, continuous from the right, tending to 0 as $x \to -\infty$ and to 1 as $x \to \infty$. Other distribution functions contrast starkly to this function by being continuous, and our next example is such a function.

Example 5.15 (Uniform distribution) Let $a, b \in \mathbb{R}$ and $a < b$. The function

$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ \dfrac{x-a}{b-a} & \text{if } a \leq x \leq b, \\ 1 & \text{if } x > b, \end{cases}$$

is sketched in Figure 5.2. It has the properties (5.5)–(5.8) and is thus a distribution function. A random variable with this distribution function is said to have the *uniform distribution* on the interval (a, b); some people call this the uniform distribution on $[a, b]$. △

Example 5.16 (Exponential distribution) Let $\lambda > 0$ and let F be given by

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - e^{-\lambda x} & \text{if } x > 0, \end{cases} \tag{5.17}$$

as sketched in Figure 5.3. Clearly F is a distribution function. A random variable with this distribution is said to have the *exponential distribution* with parameter λ. △

The two distribution functions above are very important in probability theory. There are many other distribution functions including, for example, any non-negative function F which is continuous and non-decreasing and satisfies

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

Exercise 5.18 Show that if F_1 and F_2 are distribution functions, then so is the function $F(x) = \alpha F_1(x) + (1 - \alpha)F_2(x)$ for any α satisfying $0 \leq \alpha \leq 1$.

Exercise 5.19 Let

$$F(x) = c \int_{-\infty}^{x} e^{-|u|}\, du \qquad \text{for } x \in \mathbb{R}.$$

For what value of c is F a distribution function?

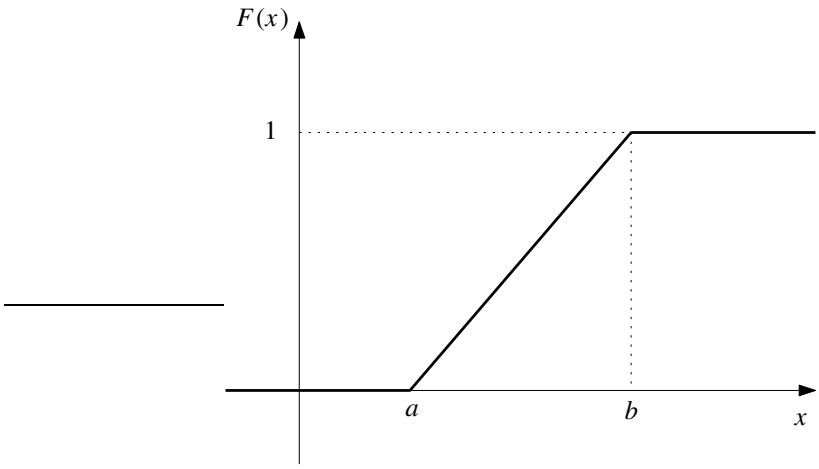


Fig. 5.2 The distribution function of the uniform distribution.

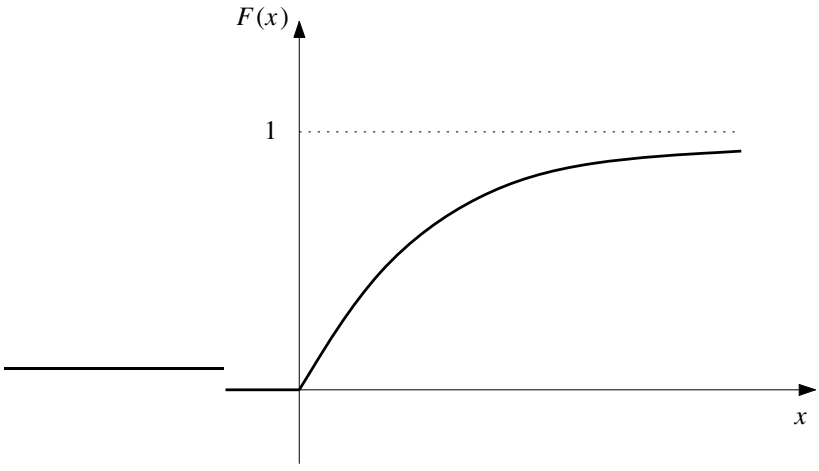


Fig. 5.3 The distribution function of the exponential distribution.

5.3 Continuous random variables

Random variables come in many shapes, but there are two classes of random variables which are particularly important:

I. discrete random variables,
II. continuous random variables.

Discrete random variables take only countably many values, and their distribution functions generally look like step functions (remember Figure 5.1). At the other extreme, there are random variables whose distribution functions are very smooth (remember Figures 5.2–5.3), and we call such random variables ‘continuous’.

Definition 5.20 *A random variable X is **continuous** if its distribution function F_X may be written in the form*

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)\,du \qquad \text{for } x \in \mathbb{R}, \tag{5.21}$$

for some non-negative function f_X.[4] *In this case, we say that X has **(probability) density function** (or **pdf**) f_X.*

Example 5.22 If X has the exponential distribution with parameter λ, then

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - e^{-\lambda x} & \text{if } x > 0, \end{cases}$$

with density function

$$f_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ \lambda e^{-\lambda x} & \text{if } x > 0. \end{cases}$$ △

Provided that X is a continuous random variable and F_X is well behaved in (5.21), we can take

$$f_X(x) = \begin{cases} \dfrac{d}{dx} F_X(x) & \text{if this derivative exists at } x, \\ 0 & \text{otherwise,} \end{cases} \tag{5.23}$$

as the density function of X. We shall normally do this, although we should point out that there are some difficulties over mathematical rigour here. However, for almost all practical purposes (5.23) is adequate, and the reader of a text at this level should seldom get into trouble if he or she uses (5.23) when finding density functions of continuous random variables.

Density functions serve continuous random variables in very much the same way as mass functions serve discrete random variables, and it is not surprising that the general properties of density functions and mass functions are very similar. For example, it is clear that the density function f_X of X satisfies

$$f_X(x) \geq 0 \text{ for } x \in \mathbb{R}, \qquad (p_Y(x) \geq 0) \tag{5.24}$$

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1, \qquad \left(\sum_x p_Y(x) = 1\right), \tag{5.25}$$

where the parentheses contain the corresponding properties, (2.5) and (2.6), of a mass function p_Y. However, this analogy can be dangerous, since $f_X(x)$ is *not* a probability and may well even exceed 1 in value. On the other hand, $f_X(x)$ is indeed a 'measure' of probability in the

[4] More advanced textbooks call such random variables 'absolutely continuous'.

following sense. If δx is small and positive, then, roughly speaking, the probability that X is 'near' x satisfies

$$\begin{aligned}\mathbb{P}(x < X \leq x + \delta x) &= F(x + \delta x) - F(x) && \text{by (5.10)} \\ &= \int_x^{x+\delta x} f_X(u)\, du && \text{by (5.21)} \\ &\approx f_X(x)\delta x && \text{for small } \delta x. \end{aligned} \tag{5.26}$$

So the true analogy is not between a density function $f_X(x)$ and a mass function $p_Y(x)$ but instead between $f_X(x)\delta x$ and $p_Y(x)$. This is borne out by comparing (5.25) with (2.6): values of the mass function are replaced by $f_X(x)\delta x$, and the summation (since, for discrete random variables, only countably many values are positive) is replaced by the integral. A startling difference between discrete and continuous random variables is given in the first part of the next theorem.

Theorem 5.27 *If X is continuous with density function f_X, then*

$$\mathbb{P}(X = x) = 0 \qquad \text{for } x \in \mathbb{R}, \tag{5.28}$$

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(u)\, du \qquad \text{for } a, b \in \mathbb{R} \text{ with } a \leq b. \tag{5.29}$$

Proof We argue as follows:

$$\begin{aligned}\mathbb{P}(X = x) &= \lim_{\epsilon \downarrow 0} \mathbb{P}(x - \epsilon < X \leq x) \\ &= \lim_{\epsilon \downarrow 0} \big[F_X(x) - F_X(x - \epsilon) \big] && \text{by (5.10)} \\ &= \lim_{\epsilon \downarrow 0} \int_{x-\epsilon}^{x} f_X(u)\, du && \text{by (5.21)} \\ &= 0. \end{aligned}$$

The first equality here cannot be justified without an appeal to the continuity of $\mathbb{P}$, Theorem 1.54. For the second part of the theorem, if $a \leq b$, then

$$\begin{aligned}\mathbb{P}(a \leq X \leq b) &= \mathbb{P}(a < X \leq b) && \text{by (5.28)} \\ &= F_X(b) - F_X(a) && \text{by (5.10)} \\ &= \int_a^b f_X(u)\, du. \end{aligned}$$

□

To recap, all random variables have a distribution function. In addition, discrete random variables have a mass function, and continuous random variables have a density function. There are many random variables which are neither discrete nor continuous, and we shall come across some of these later.

Exercise 5.30 A random variable X has density function

$$f(x) = \begin{cases} 2x & \text{if } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the distribution function of X.

Exercise 5.31 If X has density function

$$f(x) = \tfrac{1}{2}e^{-|x|} \qquad \text{for } x \in \mathbb{R},$$

find the distribution function of X. This is called the *bilateral* (or *double*) *exponential* distribution.

Exercise 5.32 If X has distribution function

$$F(x) = \begin{cases} \dfrac{1}{2(1+x^2)} & \text{for } -\infty < x \leq 0, \\ \dfrac{1+2x^2}{2(1+x^2)} & \text{for } 0 < x < \infty, \end{cases}$$

show that X is continuous and find its density function.

Exercise 5.33 Find the distribution function of the so-called 'extreme value' density function

$$f(x) = \exp(-x - e^{-x}) \qquad \text{for } x \in \mathbb{R}.$$

5.4 Some common density functions

It is fairly clear that any function f which satisfies

$$f(x) \geq 0 \qquad \text{for } x \in \mathbb{R} \tag{5.34}$$

and

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{5.35}$$

is the density function of some random variable. To confirm this, simply define

$$F(x) = \int_{-\infty}^{x} f(u)\,du$$

and check that F is a distribution function by verifying (5.5)–(5.8). There are several such functions f which are especially important in practice, and we list these below.

The **uniform distribution** on the interval (a, b) has density function

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases} \tag{5.36}$$

The **exponential distribution** with parameter $\lambda > 0$ has density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases} \tag{5.37}$$

The **normal (or Gaussian) distribution** with parameters μ and σ^2, sometimes written as $\mathrm{N}(\mu, \sigma^2)$, has density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \qquad \text{for } x \in \mathbb{R}. \tag{5.38}$$

The **Cauchy distribution** has density function

$$f(x) = \frac{1}{\pi(1+x^2)} \qquad \text{for } x \in \mathbb{R}. \tag{5.39}$$

The **gamma distribution** with parameters w (> 0) and λ (> 0) has density function

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(w)} \lambda^w x^{w-1} e^{-\lambda x} & \text{if } x > 0, \\ 0 & \text{if } x \leq 0, \end{cases} \tag{5.40}$$

where $\Gamma(w)$ is the *gamma function*, defined by

$$\Gamma(w) = \int_0^\infty x^{w-1} e^{-x}\, dx. \tag{5.41}$$

Note that, for positive integers w, $\Gamma(w) = (w-1)!$ (see Exercise 5.46).

The **beta distribution** with parameters s, t (> 0) has density function

$$f(x) = \frac{1}{B(s,t)} x^{s-1}(1-x)^{t-1} \qquad \text{for } 0 \leq x \leq 1. \tag{5.42}$$

The *beta function*

$$B(s,t) = \int_0^1 x^{s-1}(1-x)^{t-1}\, dx \tag{5.43}$$

is chosen so that f has integral equal to one. You may care to prove that

$$B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$$

(see (6.44)). If $s = t = 1$, then X is uniform on $[0, 1]$.

The **chi-squared distribution with *n* degrees of freedom** (sometimes written χ^2_n) has density function

$$f(x) = \begin{cases} \dfrac{1}{2\Gamma(\frac{1}{2}n)} (\frac{1}{2}x)^{\frac{1}{2}n-1} e^{-\frac{1}{2}x} & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases} \tag{5.44}$$

A comparison of (5.44) with (5.40) shows that the χ^2_n distribution is the same as the gamma distribution with parameters $\frac{1}{2}n$ and $\frac{1}{2}$, but we list the distribution separately here because of its common occurrence in statistics.

The above list is a dull compendium of some of the commoner density functions, and we do not expect it to inspire the reader in this form. It is difficult to motivate these density functions adequately at this stage, but we shall need to refer back to this section later when we meet these functions in action.

It is not always a trivial task to show that these functions are actually density functions. The condition (5.34) of non-negativity is no problem, but some care is required in checking that the functions integrate to 1. For example, to check this for the function given in (5.38) we require the standard definite integral

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}.$$

An outline of the proof of this may be found in Exercise 5.47.

The constant terms in (5.36)–(5.44) have been chosen solely so that the resulting functions integrate to 1. For example, it is clear that the function

$$g(x) = \frac{1}{1+x^2} \qquad \text{for } x \in \mathbb{R},$$

is not a density function since

$$\int_{-\infty}^{\infty} g(x)\,dx = \pi,$$

but it follows that the 'normalized' function

$$f(x) = \frac{1}{\pi} g(x)$$

is a density function.

Exercise 5.45 For what values of its parameters is the gamma distribution also an exponential distribution?

Exercise 5.46 Show that the gamma function $\Gamma(w)$ satisfies $\Gamma(w) = (w-1)\Gamma(w-1)$ for $w > 1$, and deduce that $\Gamma(n) = (n-1)!$ for $n = 1, 2, 3, \ldots$.

Exercise 5.47 Let

$$I = \int_{-\infty}^{\infty} e^{-x^2}\,dx.$$

By changing variables to polar coordinates, show that

$$I^2 = \iint_{\mathbb{R}^2} e^{-x^2-y^2}\,dx\,dy = \int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} e^{-r^2} r\,dr\,d\theta,$$

and deduce that $I = \sqrt{\pi}$.

Exercise 5.48 Show that the density function

$$f(x) = \begin{cases} \dfrac{1}{\pi\sqrt{x(1-x)}} & \text{if } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

has distribution function with the form

$$F(x) = c \sin^{-1} \sqrt{x} \qquad \text{if } 0 < x < 1,$$

and find the constant c.

5.5 Functions of random variables

Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and suppose that $g : \mathbb{R} \to \mathbb{R}$. Then $Y = g(X)$ is a mapping from Ω into $\mathbb{R}$, defined by $Y(\omega) = g[X(\omega)]$ for $\omega \in \Omega$. Actually, Y is not generally a random variable since it need not satisfy condition (5.1). It turns out, however, that (5.1) is valid for Y whenever g is sufficiently well behaved (such as g is a continuous function, or a monotone function, or ...), and so we neglect this difficulty, assuming henceforth that *all quantities of the form $Y = g(X)$ are random variables*. The main question is now the following: if we know the distribution of X, then how do we find the distribution of $Y = g(X)$? If X is discrete with mass function p_X, then (2.25) provides the answer, and we consider next the case when X is continuous with density function f_X. We begin with an example.

Example 5.49 If X is continuous with density function f_X, and $g(x) = ax + b$ when $a > 0$, then $Y = g(X) = aX + b$ has distribution function given by

$$\begin{aligned} \mathbb{P}(Y \le y) &= \mathbb{P}(aX + b \le y) \\ &= \mathbb{P}\big(X \le a^{-1}(y - b)\big) \\ &= F_X\big(a^{-1}(y - b)\big). \end{aligned}$$

By differentiation with respect to y,

$$f_Y(y) = a^{-1} f_X\big(a^{-1}(y - b)\big) \qquad \text{for } y \in \mathbb{R}. \qquad \triangle$$

The next theorem generalizes the result of this example.

Theorem 5.50 *If X is a continuous random variable with density function f_X, and g is a strictly increasing and differentiable function from $\mathbb{R}$ into $\mathbb{R}$, then $Y = g(X)$ has density function*

$$f_Y(y) = f_X\big(g^{-1}(y)\big) \frac{d}{dy}[g^{-1}(y)] \qquad \textit{for } y \in \mathbb{R}, \tag{5.51}$$

where g^{-1} is the inverse function of g.

Proof First, we find the distribution function of Y:

$$\begin{aligned}\mathbb{P}(Y \le y) &= \mathbb{P}\big(g(X) \le y\big) \\ &= \mathbb{P}\big(X \le g^{-1}(y)\big) \quad \text{since } g \text{ is increasing.}\end{aligned}$$

We differentiate this with respect to y, noting that the right-hand side is a function of a function, to obtain (5.51). □

If, in Theorem 5.50, g were strictly *decreasing*, then the same argument gives that $Y = g(X)$ has density function

$$f_Y(y) = -f_X\big(g^{-1}(y)\big)\frac{d}{dy}[g^{-1}(y)] \qquad \text{for } y \in \mathbb{R}. \tag{5.52}$$

Formulae (5.51) and (5.52) rely heavily on the monotonicity of g. Other cases are best treated on their own merits, and actually there is a lot to be said for using the method of the next example always, rather than taking recourse in the general results (5.51)–(5.52).

Example 5.53 If X has density function f_X, and $g(x) = x^2$, then $Y = g(X) = X^2$ has distribution function

$$\begin{aligned}\mathbb{P}(Y \le y) &= \mathbb{P}(X^2 \le y) \\ &= \begin{cases} 0 & \text{if } y < 0, \\ \mathbb{P}\big(-\sqrt{y} \le X \le \sqrt{y}\big) & \text{if } y \ge 0. \end{cases}\end{aligned}$$

Hence $f_Y(y) = 0$ if $y \le 0$, while for $y > 0$

$$\begin{aligned}f_Y(y) &= \frac{d}{dy}\mathbb{P}(Y \le y) \qquad \text{if this derivative exists} \\ &= \frac{d}{dy}\big[F_X(\sqrt{y}) - F_X(-\sqrt{y})\big] \\ &= \frac{1}{2\sqrt{y}}\big[f_X(\sqrt{y}) + f_X(-\sqrt{y})\big].\end{aligned}$$

△

Exercise 5.54 Let X be a random variable with the exponential distribution, parameter λ. Find the density function of

(a) $A = 2X + 5$,

(b) $B = e^X$,

(c) $C = (1 + X)^{-1}$,

(d) $D = (1 + X)^{-2}$.

Exercise 5.55 Show that if X has the normal distribution with parameters 0 and 1, then $Y = X^2$ has the χ^2 distribution with one degree of freedom.

5.6 Expectations of continuous random variables

If a one-dimensional metal rod has density $\rho(x)$ at point x, then its mass is $m = \int \rho(x)\,dx$, and its centre of gravity is at the position $m^{-1}\int x\rho(x)\,dx$. This leads naturally to the idea of the expectation of a continuous random variable.

Definition 5.56 *If X is a continuous random variable with density function f_X, the **expectation** of X is denoted by $\mathbb{E}(X)$ and defined by*

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx, \tag{5.57}$$

whenever this integral converges absolutely, in that $\int_{-\infty}^{\infty} |x f_X(x)|\,dx < \infty$.

As in the case of discrete variables, the expectation of X is often called the *expected value* or *mean* of X.

If X is a continuous variable and $g : \mathbb{R} \to \mathbb{R}$, then $Y = g(X)$ is a random variable also (so long as g is sufficiently well behaved). It may be difficult to calculate $\mathbb{E}(Y)$ from first principles, not least since Y may be neither discrete nor continuous and so neither of formulae (2.28) and (5.57) may apply. Of great value here is the following result, which enables us to calculate $\mathbb{E}(Y)$ directly from knowledge of f_X and g.

Theorem 5.58 (Law of the subconscious statistician) *If X is a continuous random variable with density function f_X, and $g : \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx, \tag{5.59}$$

whenever this integral converges absolutely.

Sketch proof The theorem is not too difficult to prove, but the full proof is a little long (see Grimmett and Stirzaker (2001, p. 93) for a discussion). We think that it is more important to understand *why* it is true rather than to see a formal proof. If Y is a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$, then

$$\mathbb{E}(g(Y)) = \sum_x g(x)\mathbb{P}(Y = x), \tag{5.60}$$

as in Theorem 2.29. Remember the analogy between mass functions and density functions: in (5.60), replace $\mathbb{P}(Y = x)$ by $f_X(x)\,dx$ and the summation by the integral, to obtain (5.59). A more complete proof is outlined at Problem 5.8.8. □

As in the case of discrete random variables, the mean $\mathbb{E}(X)$ of a continuous random variable X is an indication of the 'centre' of the distribution of X. As a measure of the degree of dispersion of X about this mean, we normally take the *variance* of X, defined to be

$$\operatorname{var}(X) = \mathbb{E}\big([X - \mathbb{E}(X)]^2\big). \tag{5.61}$$

By Theorem 5.58,

$$\operatorname{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx,$$

where $\mu = \mathbb{E}(X)$. Therefore,

$$\begin{aligned}\operatorname{var}(X) &= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f_X(x)\,dx \\ &= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx - 2\mu \int_{-\infty}^{\infty} x f_X(x)\,dx + \mu^2 \int_{-\infty}^{\infty} f_X(x)\,dx \\ &= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx - \mu^2\end{aligned}$$

by (5.57) and (5.35). Thus we obtain the usual formula

$$\operatorname{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \tag{5.62}$$

Example 5.63 If X has the *uniform distribution* on (a, b), then

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} x f_X(x)\,dx \qquad \text{by (5.57)} \\ &= \int_a^b x \frac{1}{b-a}\,dx = \tfrac{1}{2}(a+b).\end{aligned}$$

As an example of a function of a random variable, let $Y = \sin X$. Then

$$\begin{aligned}\mathbb{E}(Y) &= \int_{-\infty}^{\infty} \sin x \; f_X(x)\,dx \qquad \text{by (5.59)} \\ &= \int_a^b \frac{\sin x}{b-a}\,dx = \frac{\cos a - \cos b}{b-a}.\end{aligned}$$

△

Example 5.64 If X has the *exponential distribution* with parameter λ, then

$$\begin{aligned}\mathbb{E}(X) &= \int_0^{\infty} x\lambda e^{-\lambda x}\,dx = \frac{1}{\lambda}, \\ \mathbb{E}(X^2) &= \int_0^{\infty} x^2\lambda e^{-\lambda x}\,dx = \frac{2}{\lambda^2}\end{aligned}$$

by (5.59), giving by (5.62) that the variance of X is

$$\begin{aligned}\operatorname{var}(X) &= \mathbb{E}\big([X - \mathbb{E}(X)]^2\big) \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{1}{\lambda^2}.\end{aligned}$$

△

Example 5.65 If X has the *normal distribution* with parameters $\mu = 0$ and $\sigma^2 = 1$, then

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx = 0,$$

by symmetry properties of the integrand. Hence

$$\text{var}(X) = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx = 1.$$

Similar integrations show that the normal distribution with parameters μ and σ^2 has mean μ and variance σ^2, as we may have expected. △

Example 5.66 If X has the *Cauchy distribution*, then

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \frac{1}{\pi(1+x^2)}\, dx,$$

so long as this integral exists. *It does not exist*, since

$$\begin{aligned}\int_{-M}^{N} x \frac{1}{\pi(1+x^2)}\, dx &= \left[\frac{1}{2\pi}\log(1+x^2)\right]_{-M}^{N} \\ &= \frac{1}{2\pi}\log\frac{1+N^2}{1+M^2} = l(M,N),\end{aligned}$$

say, and the limit of $l(M, N)$ as $M, N \to \infty$ depends on the way in which M and N approach ∞. If $M \to \infty$ and $N \to \infty$ in that order, then $l(M, N) \to -\infty$, while if the limit is taken in the other order, then $l(M, N) \to \infty$. Hence *the Cauchy distribution does not have a mean value*. On the other hand, there are many functions of X with finite expectations. For example, if $Y = \tan^{-1} X$, then

$$\begin{aligned}\mathbb{E}(Y) &= \int_{-\infty}^{\infty} \tan^{-1} x \ \frac{1}{\pi(1+x^2)}\, dx \\ &= \int_{-\frac{1}{2}\pi}^{\frac{1}{2}\pi} \frac{v}{\pi}\, dv \qquad \text{where } v = \tan^{-1} x \\ &= 0.\end{aligned}$$

△

Exercise 5.67 Show that a random variable with density function

$$f(x) = \begin{cases} \dfrac{1}{\pi\sqrt{x(1-x)}} & \text{if } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

has mean $\frac{1}{2}$.

Exercise 5.68 The random variable X has density function

$$f(x) = cx(1-x) \qquad \text{for } 0 \le x \le 1.$$

Determine c, and find the mean and variance of X.

Exercise 5.69 If X has the normal distribution with mean 0 and variance 1, find the mean value of $Y = e^{2X}$.

5.7 Geometrical probability

This chapter closes with three examples of the use of probability in simple geometrical calculations, namely Bertrand's paradox, Buffon's needle, and stick breaking.

The paradox of Joseph Louis François Bertrand. A chord of the unit circle is picked at random. What is the probability that an equilateral triangle with the chord as base fits inside the circle?

There are several ways of 'choosing a chord at random', and the 'paradox' lies in the multiplicity of answers arising from different interpretations. Here are three such interpretations, and a fourth is found in Exercise 5.70.

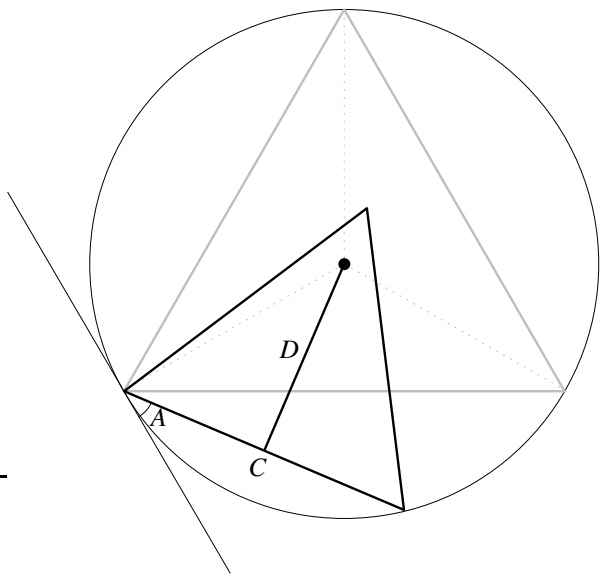


Fig. 5.4 The triangle can be drawn inside the circle if and only if $D \ge \frac{1}{2}$.

I. Let D be the perpendicular distance between the centre of the circle and the chord, as illustrated in Figure 5.4. The triangle can be drawn inside the circle if and only if $D \ge \frac{1}{2}$. *Assume* that D is uniformly distributed on the interval $(0, 1)$. Then the answer is

$$\mathbb{P}(D \ge \tfrac{1}{2}) = \tfrac{1}{2}.$$

II. *Assume* the acute angle A between the chord and the tangent is uniform on the interval $(0, \frac{1}{2}\pi)$. Since $D = \frac{1}{2}$ when $A = \frac{1}{3}\pi$, the answer is

$$\mathbb{P}(A \le \tfrac{1}{3}\pi) = \tfrac{1}{3}\pi \big/ \tfrac{1}{2}\pi = \tfrac{2}{3}.$$

III. *Assume* the centre C of the chord is chosen uniformly in the interior of the circle. Then $D \le d$ if and only if C lies within a circle of radius d, so that $\mathbb{P}(D \le d) = \pi d^2/\pi = d^2$. The answer is

$$\mathbb{P}(D \ge \tfrac{1}{2}) = 1 - \mathbb{P}(D \le \tfrac{1}{2}) = \tfrac{3}{4}.$$

The needle of Georges Louis Leclerc, Comte de Buffon. This is more interesting. A plane is ruled by straight lines which are unit distance apart, as in Figure 5.5. What is the probability that a unit needle, dropped at random, intersects a line?[5]

We may position the plane so that the x-axis is along a line. Let (X, Y) be the coordinates of the midpoint of the needle, and let Θ be its inclination to horizontal, as in the figure. It is reasonable to assume that

(a) $Z := Y - \lfloor Y \rfloor$ is uniformly distributed on $[0, 1]$,
(b) Θ is uniform on $[0, \pi]$, and
(c) Z and Θ are independent.

Thus,

$$f_{Z,\Theta}(z, \theta) = \frac{1}{\pi} \qquad \text{for } z \in [0, 1],\ \theta \in [0, \pi].$$

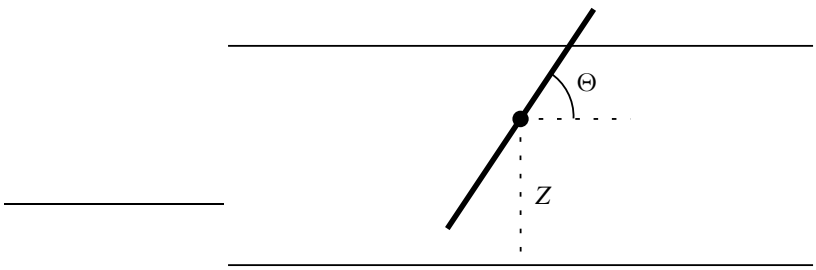


Fig. 5.5 The needle and the ruled plane in Buffon's problem.

An intersection occurs if and only if $(z, \theta) \in B$, where

$$B = \big\{(z, \theta) : \text{either } z \le \tfrac{1}{2}\sin\theta \text{ or } 1 - z \le \tfrac{1}{2}\sin\theta\big\}.$$

Therefore, the probability of an intersection satisfies

[5]The original problem of Buffon was slightly different, namely the following. A coin is dropped on a tiled floor. What is the probability it falls on a crack? See Problem 5.8.12.

$$\begin{aligned}\mathbb{P}(\text{intersection}) &= \iint_B \frac{1}{\pi}\, dz\, d\theta \\ &= \frac{1}{\pi}\int_{\theta=0}^{\pi} d\theta \left(\int_0^{\frac{1}{2}\sin\theta} dz + \int_{1-\frac{1}{2}\sin\theta}^{1} dz \right) \\ &= \frac{1}{\pi}\int_0^{\pi} d\theta\, \sin\theta = \frac{2}{\pi}.\end{aligned}$$

This motivates a Monte Carlo experiment to estimate the value of π. Drop the needle n times, and let I_n be the number of throws that result in intersections. The natural estimate of π is $\widehat{\pi}_n := (2n)/I_n$. It may be shown that $\mathbb{E}(\widehat{\pi}_n) \to \pi$ as $n \to \infty$, and furthermore the variance of $\widehat{\pi}_n$ is of order n^{-1}. Thus, there is a sense in which the accuracy of this experiment increases as $n \to \infty$. There are, however, better ways to estimate π than by Monte Carlo methods.

Stick breaking. Here is an everyday problem of broken sticks. A stick of unit length is broken at the two places X, Y, each chosen uniformly at random along the stick. What is the probability that the three pieces can be used to make a triangle? We assume that X and Y are independent.

The lengths of the three substicks are

$$U = \min\{X, Y\}, \quad V = |Y - X|, \quad W = 1 - U - V,$$

and we are asked for the probability that no substick is longer than the sum of the other two lengths. Since $U + V + W = 1$, this is equivalent to requiring that $U, V, W \le \frac{1}{2}$. The region of the (X, Y)-plane satisfying $U, V, 1 - U - V \le \frac{1}{2}$ is shaded in Figure 5.6, and it has area $\frac{1}{4}$. Therefore, the answer is $\frac{1}{4}$.

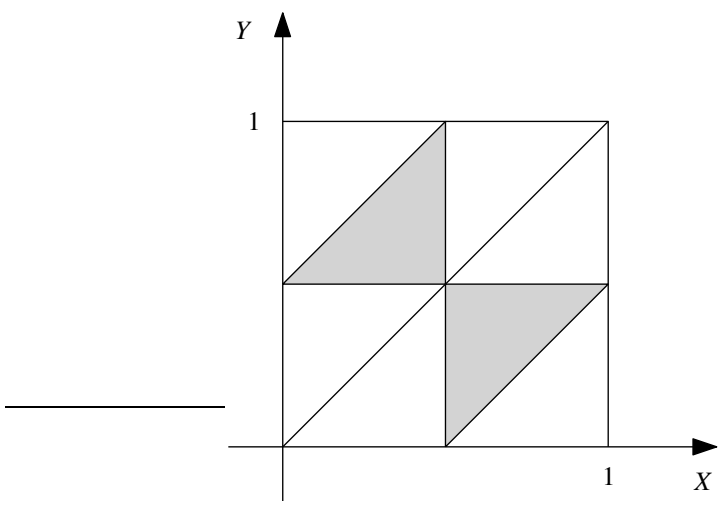


Fig. 5.6 The shaded area corresponds to the region where (U, V) satisfies the given conditions.

Exercise 5.70 What is the answer to Bertrand's question if the chord is PQ, where P and Q are chosen uniformly and independently at random on the circumference of the circle?

Exercise 5.71 Suppose Buffon junior uses a needle of length ℓ (< 1). Show that the probability of an intersection is $2\ell/\pi$.

Exercise 5.72 One of Hugo's longer noodles, of length ℓ, falls at random from his bowl onto Buffon's ruled plane. Show that the mean number of intersections of the noodle with lines is $2\ell/\pi$.

5.8 Problems

1. The *bilateral* (or *double*) *exponential* distribution has density function

$$f(x) = \tfrac{1}{2}ce^{-c|x|} \qquad \text{for } x \in \mathbb{R},$$

where c (> 0) is a parameter of the distribution. Show that the mean and variance of this distribution are 0 and $2c^{-2}$, respectively.

2. Let X be a random variable with the Poisson distribution, parameter λ. Show that, for $w = 1, 2, 3, \ldots$,

$$\mathbb{P}(X \geq w) = \mathbb{P}(Y \leq \lambda),$$

where Y is a random variable having the gamma distribution with parameters w and 1.

3. The random variable X has density function proportional to $g(x)$, where g is a function satisfying

$$g(x) = \begin{cases} |x|^{-n} & \text{if } |x| \geq 1, \\ 0 & \text{otherwise,} \end{cases}$$

and n (≥ 2) is an integer. Find and sketch the density function of X, and determine the values of n for which both the mean and variance of X exist.

4. If X has the normal distribution with mean 0 and variance 1, find the density function of $Y = |X|$, and find the mean and variance of Y.

5. Let X be a random variable whose distribution function F is a continuous function. Show that the random variable Y, defined by $Y = F(X)$, is uniformly distributed on the interval (0, 1).

* 6. Let F be a distribution function, and let X be a random variable which is uniformly distributed on the interval (0, 1). Let F^{-1} be the inverse function of F, defined by

$$F^{-1}(y) = \inf\{x : F(x) \geq y\}.$$

Show that the random variable $Y = F^{-1}(X)$ has distribution function F. This observation may be used in practice to generate pseudorandom numbers drawn from any given distribution.

7. If X is a continuous random variable taking non-negative values only, show that

$$\mathbb{E}(X) = \int_0^\infty [1 - F_X(x)]\, dx,$$

whenever this integral exists.

* 8. Use the result of Problem 5.8.7 to show that

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

whenever X and $g(X)$ are continuous random variables and $g : \mathbb{R} \to [0, \infty)$.

9. The random variable X' is said to be obtained from the random variable X by *truncation* at the point a if

$$X'(\omega) = \begin{cases} X(\omega) & \text{if } X(\omega) \le a, \\ a & \text{if } X(\omega) > a. \end{cases}$$

Express the distribution function of X' in terms of the distribution function of X.

10. Let X have the exponential distribution with parameter 1. Find the density function of $Y = (X - 2)/(X + 1)$.

11. William Tell is a very bad shot. In practice, he places a small green apple on top of a straight wall which stretches to infinity in both directions. He then takes up position at a distance of one perch from the apple, so that his line of sight to the target is perpendicular to the wall. He now selects an angle uniformly at random from his entire field of view and shoots his arrow in this direction. Assuming that his arrow hits the wall somewhere, what is the distribution function of the horizontal distance (measured in perches) between the apple and the point which the arrow strikes? There is no wind.

* 12. *Buffon–Laplace needle.* Let $a, b > 0$. The Cartesian plane is ruled with two sets of parallel lines of the form $x = ma$ and $y = nb$ for integers m and n. A needle of length ℓ ($< \min\{a, b\}$) is dropped at random. Show that the probability it intersects some line is $\ell(2a+2b-\ell)/(\pi ab)$.

* 13. A unit stick is broken at n random places, each uniform on $[0, 1]$, and different breaks are chosen independently. Show that the resulting $n + 1$ substicks can form a closed polygon with probability $1 - (n + 1)/2^n$.

14. The random variable X is uniformly distributed on the interval $[0, 1]$. Find the distribution and probability density function of Y, where

$$Y = \frac{3X}{1 - X}.$$

(Cambridge 2003)

Part B

Further Probability

6

Multivariate distributions and independence

Summary. A random vector is studied via its joint distribution function, and this leads to a discussion of the independence of random variables. The joint, marginal, and conditional density functions of continuous variables are defined, and their theory explored. Sums of independent variables are studied via the convolution formula, and transformations of random vectors via the Jacobian method. The basic properties of the bivariate normal distribution are described.

6.1 Random vectors and independence

Given two random variables X and Y, acting on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, it is often useful to think of them acting together as a random vector (X, Y) taking values in $\mathbb{R}^2$. If X and Y are discrete, we may study this random vector by using the joint mass function of X and Y, but this method is not always available. In the general case of arbitrary random variables X, Y, we study instead their *joint distribution function*, defined as follows.

Definition 6.1 *The **joint distribution function** of the pair X, Y of random variables is the mapping $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ given by*

$$F_{X,Y}(x, y) = \mathbb{P}(X \le x,\ Y \le y). \tag{6.2}$$

Joint distribution functions have certain elementary properties which are exactly analogous to those of ordinary distribution functions. For example, it is easy to see that

$$\lim_{x,y\to-\infty} F_{X,Y}(x, y) = 0, \tag{6.3}$$

$$\lim_{x,y\to\infty} F_{X,Y}(x, y) = 1, \tag{6.4}$$

just as in (5.6) and (5.7). Similarly, $F_{X,Y}$ is non-increasing in each variable in that

$$F_{X,Y}(x_1, y_1) \le F_{X,Y}(x_2, y_2) \quad \text{if} \quad x_1 \le x_2 \text{ and } y_1 \le y_2. \tag{6.5}$$

The joint distribution function $F_{X,Y}$ contains a great deal more information than the two ordinary distribution functions F_X and F_Y, since it tells us how X and Y behave *together*. In

particular, the distribution functions of X and Y may be found from their joint distribution function in a routine way. It is intuitively attractive to write

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \le x) \\ &= \mathbb{P}(X \le x,\ Y \le \infty) = F_{X,Y}(x, \infty) \end{aligned}$$

and similarly,

$$F_Y(y) = F_{X,Y}(\infty, y),$$

but the mathematically correct way of expressing this is

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x, y), \qquad F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x, y). \tag{6.6}$$

These distribution functions are called the *marginal* distribution functions of the joint distribution function $F_{X,Y}$.

The idea of 'independence' of random variables X and Y follows naturally from this discussion.

Definition 6.7 *We call X and Y **independent** if, for all $x, y \in \mathbb{R}$, the events $\{X \le x\}$ and $\{Y \le y\}$ are independent.*

That is to say, X and Y are independent if and only if

$$\mathbb{P}(X \le x,\ Y \le y) = \mathbb{P}(X \le x)\mathbb{P}(Y \le y) \qquad \text{for } x, y \in \mathbb{R},$$

which is to say that their joint distribution function factorizes as the product of the two marginal distribution functions:

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \qquad \text{for } x, y \in \mathbb{R}. \tag{6.8}$$

It is a straightforward exercise to show that this is a genuine extension of the notion of independent *discrete* random variables. Random variables which are not independent are called *dependent*.

We study families of random variables in very much the same way. Briefly, if $X_1, X_2, \ldots, X_n$ are random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, their *joint distribution function* is the function $F_{\mathbf{X}} : \mathbb{R}^n \to [0, 1]$ given by

$$F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}\big(X_1 \le x_1,\ X_2 \le x_2, \ldots,\ X_n \le x_n\big) \tag{6.9}$$

for $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. The variables $X_1, X_2, \ldots, X_n$ are called *independent* if

$$\mathbb{P}\big(X_1 \le x_1, \ldots, X_n \le x_n\big) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n) \qquad \text{for } \mathbf{x} \in \mathbb{R}^n,$$

or equivalently if

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1}(x_1) \cdots F_{X_n}(x_n) \qquad \text{for } \mathbf{x} \in \mathbb{R}^n. \tag{6.10}$$

Example 6.11 Suppose that X and Y are random variables on some probability space, each taking values in the integers $\{\ldots, -1, 0, 1, \ldots\}$ with joint mass function

$$\mathbb{P}(X = i,\ Y = j) = p(i, j) \qquad \text{for } i, j = 0, \pm 1, \pm 2, \ldots.$$

Their joint distribution function is given by

$$F_{X,Y}(x, y) = \sum_{i \le x,\ j \le y} p(i, j) \qquad \text{for } (x, y) \in \mathbb{R}^2. \qquad \triangle$$

Example 6.12 Suppose that X and Y are random variables with joint distribution function

$$F_{X,Y}(x, y) = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-x-y} & \text{if } x, y \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

The (marginal) distribution function of X is

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y) = \begin{cases} 1 - e^{-x} & \text{if } x \ge 0, \\ 0 & \text{otherwise,} \end{cases}$$

so that X has the exponential distribution with parameter 1. A similar calculation shows that Y has this distribution also. Hence

$$F_{X,Y}(x, y) = F_X(x) F_Y(y) \qquad \text{for } x, y \in \mathbb{R},$$

and so X and Y are independent. $\triangle$

Exercise 6.13 Show that two random variables X and Y are independent if and only if

$$\mathbb{P}(X > x,\ Y > y) = \mathbb{P}(X > x)\mathbb{P}(Y > y) \qquad \text{for } x, y \in \mathbb{R}.$$

Exercise 6.14 Let the pair (X, Y) of random variables have joint distribution function $F(x, y)$. Prove that

$$\mathbb{P}(a < X \le b,\ c < Y \le d) = F(b, d) + F(a, c) - F(a, d) - F(b, c)$$

for any $a, b, c, d \in \mathbb{R}$ such that $a < b$ and $c < d$.

Exercise 6.15 Prove that two random variables X and Y are independent if and only if

$$\mathbb{P}(a < X \le b,\ c < Y \le d) = \mathbb{P}(a < X \le b)\mathbb{P}(c < Y \le d)$$

for all $a, b, c, d \in \mathbb{R}$ satisfying $a < b$ and $c < d$.

6.2 Joint density functions

Recall that a random variable X is *continuous* if its distribution function may be expressed in the form

$$F_X(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f(u)\, du \qquad \text{for } x \in \mathbb{R}.$$

Definition 6.16 *The pair X, Y of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ is called **(jointly) continuous** if its joint distribution function is expressible in the form*[1]

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x,\ Y \leq y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f(u, v)\, du\, dv \tag{6.17}$$

*for $x, y \in \mathbb{R}$ and some function $f : \mathbb{R}^2 \to [0, \infty)$. If this holds, we say that X and Y have **joint (probability) density function** f, and we usually denote this function by $f_{X,Y}$.*

As in Section 5.3, if X and Y are jointly continuous, we may take their joint density function to be given by

$$f_{X,Y}(x, y) = \begin{cases} \dfrac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) & \text{if this derivative exists at } (x, y), \\ 0 & \text{otherwise,} \end{cases} \tag{6.18}$$

and we shall normally do this in future. There are the usual problems here over mathematical rigour but, as noted after (5.23), you should not get into trouble at this level if you take this as the definition of the joint density function of X and Y.

The elementary properties of the joint density function $f_{X,Y}$ are consequences of properties (6.3)–(6.5) of joint distribution functions:

$$f_{X,Y}(x, y) \geq 0 \qquad \text{for } x, y \in \mathbb{R}, \tag{6.19}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy = 1. \tag{6.20}$$

Once again, we note an analogy between joint density functions and joint mass functions. This may be expressed rather crudely by saying that for any $(x, y) \in \mathbb{R}^2$ and small positive δx and δy, the probability that the random vector (X, Y) lies in the small rectangle with bottom left-hand corner at (x, y) and side lengths δx and δy is

$$\mathbb{P}\big(x < X \leq x + \delta x,\ y < Y \leq y + \delta y\big) \approx f_{X,Y}(x, y)\, \delta x\, \delta y \tag{6.21}$$

(see Figure 6.1). This holds for very much the same reasons as the one-dimensional case (5.26). It is not difficult to see how this leads to the next theorem.

Theorem 6.22 *If A is any regular subset of $\mathbb{R}^2$ and X and Y are jointly continuous random variables with joint density function $f_{X,Y}$, then*

$$\mathbb{P}\big((X, Y) \in A\big) = \iint_{(x,y)\in A} f_{X,Y}(x, y)\, dx\, dy. \tag{6.23}$$

[1] We ought to say exactly what we mean by the integral on the right-hand side of (6.17). At this level, it is perhaps enough to say that this double integral may be interpreted in any standard way, and that there is a result (called Fubini's theorem) which says that, under certain wide conditions, it does not matter whether we integrate over u first or over v first when we calculate its value.

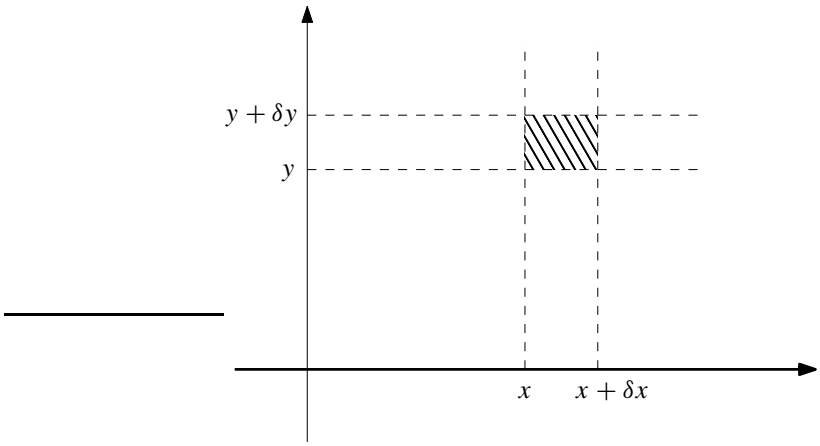


Fig. 6.1 The probability that (X, Y) lies in the shaded region is approximately $f_{X,Y}(x, y)\,\delta x\,\delta y$.

This is really a result about integration rather than about probability theory, and so we omit the proof. We do not even attempt to explain the term 'regular', noting only that it covers sets such as rectangles, discs, regions bounded by closed Jordan curves, and so on. On the other hand, it is easy to see why (6.23) should hold. The set A may be split up into the union of lots of small non-overlapping rectangles, and $f_{X,Y}(x, y)\,\delta x\,\delta y$ is the probability that (X, Y) takes a value in a typical rectangle. Roughly speaking, the probability that (X, Y) takes a value in A is the sum of these small probabilities.

Example 6.24 It is not too difficult to check that a function $f : \mathbb{R}^2 \to \mathbb{R}$ is the joint density function of some pair of random variables if and only if f satisfies (6.19) and (6.20):

$$f(x, y) \geq 0 \quad \text{for } x, y \in \mathbb{R}, \qquad \text{and} \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy = 1.$$

This holds in just the same way as the corresponding properties (5.34) and (5.35) were necessary and sufficient for a function of one variable to be a density function. It follows that the function

$$f(x, y) = \begin{cases} \dfrac{1}{ab} & \text{if } 0 < x < a \text{ and } 0 < y < b, \\ 0 & \text{otherwise,} \end{cases}$$

is a joint density function. If X and Y have joint density function f, the vector (X, Y) is said to be *uniformly distributed* on the rectangle $B = (0, a) \times (0, b)$. For any region A of the plane,

$$\begin{aligned} \mathbb{P}\big((X, Y) \in A\big) &= \iint_A f(x, y)\,dx\,dy \\ &= \iint_{A\cap B} \frac{1}{ab}\,dx\,dy = \frac{\text{area } (A \cap B)}{\text{area } (B)}. \end{aligned}$$

△

Exercise 6.25 Random variables X and Y have joint density function

$$f(x, y) = \begin{cases} c(x^2 + \frac{1}{2}xy) & \text{if } 0 < x < 1,\ 0 < y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of the constant c and the joint distribution function of X and Y.

Exercise 6.26 Random variables X and Y have joint density function

$$f(x, y) = \begin{cases} e^{-x-y} & \text{if } x, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{P}(X + Y \leq 1)$ and $\mathbb{P}(X > Y)$.

6.3 Marginal density functions and independence

Whenever the pair X, Y has joint density function $f_{X,Y}$, the ordinary density functions of X and Y may be retrieved immediately since (at points of differentiability)

$$\begin{aligned} f_X(x) &= \frac{d}{dx}\mathbb{P}(X \leq x) \\ &= \frac{d}{dx}\int_{u=-\infty}^{x}\int_{v=-\infty}^{\infty} f_{X,Y}(u, v)\,du\,dv \qquad \text{by Theorem 6.22} \\ &= \int_{v=-\infty}^{\infty} f_{X,Y}(x, v)\,dv, \end{aligned} \tag{6.27}$$

and similarly,

$$f_Y(y) = \int_{u=-\infty}^{\infty} f_{X,Y}(u, y)\,du. \tag{6.28}$$

These density functions are called the *marginal* density functions of X and of Y, since they are obtained by 'projecting' the random vector (X, Y) onto the two coordinate axes of the plane.

Recall that X and Y are *independent* if their distribution functions satisfy

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \qquad \text{for } x, y \in \mathbb{R}. \tag{6.29}$$

If X and Y are jointly continuous, then differentiation of this relation with respect to both x and y yields a condition on their density functions,

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \qquad \text{for } x, y \in \mathbb{R}, \tag{6.30}$$

and it is easy to see that (6.29) holds if and only if (6.30) holds—certainly (6.29) implies (6.30), and conversely, if we integrate both sides of (6.30) as x ranges over $(-\infty, u]$ and y ranges over $(-\infty, v]$, then we obtain (6.29). Thus, jointly continuous random variables are independent if and only if their joint density function factorizes as the product of the two marginal density functions. This is exactly analogous to the case of discrete random variables, discussed in Section 3.3. Just as in the case of discrete random variables, there is a more general result.

Theorem 6.31 *Jointly continuous random variables X and Y are independent if and only if their joint density function may be expressed in the form*

$$f_{X,Y}(x, y) = g(x)h(y) \qquad \text{for } x, y \in \mathbb{R},$$

as the product of a function of the first variable and a function of the second.

We do not prove this, but we suggest that the reader adapts the proof of Theorem 3.16, replacing summations by integrals.

We do not wish to spend a lot of time going over the case when there are three or more random variables. Roughly speaking, all the ideas of this chapter so far have analogues in more than two dimensions. For example, three random variables X, Y, Z are called *jointly continuous* if

$$\mathbb{P}\big(X \leq x,\ Y \leq y,\ Z \leq z\big) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} \int_{w=-\infty}^{z} f(u, v, w)\, du\, dv\, dw$$

for $x, y, z \in \mathbb{R}$ and some function f. If this holds, we may take

$$f(x, y, z) = \frac{\partial^3}{\partial x \partial y \partial z} \mathbb{P}\big(X \leq x,\ Y \leq y,\ Z \leq z\big)$$

to be the joint density function of the triple (X, Y, Z), whenever these derivatives exist. The random variables are *independent* if and only if f factorizes as the product of the marginal density functions:

$$f(x, y, z) = f_X(x) f_Y(y) f_Z(z) \qquad \text{for } x, y, z \in \mathbb{R}.$$

Example 6.32 Suppose that X and Y have joint density function

$$f(x, y) = \begin{cases} e^{-x-y} & \text{if } x, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y)\, dy \\ &= \begin{cases} \displaystyle\int_0^{\infty} e^{-x-y}\, dy & \text{if } x > 0, \\ 0 & \text{otherwise,} \end{cases} \\ &= \begin{cases} e^{-x} & \text{if } x > 0, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

giving that X has the exponential distribution with parameter 1. The random variable Y has this distribution also, and

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \qquad \text{for } x, y \in \mathbb{R},$$

so that X and Y are independent. △

Example 6.33 Suppose that X and Y have joint density function

$$f(x, y) = \begin{cases} ce^{-x-y} & \text{if } 0 < x < y, \\ 0 & \text{otherwise,} \end{cases} \tag{6.34}$$

for some constant c. Find c and ascertain whether X and Y are independent.

Solution Joint density functions integrate to 1, so that

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = \int_{x=0}^{\infty} \int_{y=x}^{\infty} e^{-x-y}\, dx\, dy \\ &= c \int_0^{\infty} e^{-2x}\, dx = \tfrac{1}{2}c, \end{aligned}$$

giving that $c = 2$. Clearly, X and Y are dependent, since f cannot be factorized as the product of a function of x and a function of y (look at the domain of f in (6.34)). More explicitly, by Theorem 6.22,

$$\mathbb{P}(X > 2,\ Y < 1) = \int_{x=2}^{\infty} \int_{y=-\infty}^{1} f(x, y)\, dx\, dy = 0$$

since $f(x, y) = 0$ if $y < x$. On the other hand,

$$\mathbb{P}(X > 2) > 0 \qquad \text{and} \qquad \mathbb{P}(Y < 1) > 0,$$

so that

$$\mathbb{P}(X > 2,\ Y < 1) \neq \mathbb{P}(X > 2)\mathbb{P}(Y < 1),$$

implying that X and Y are dependent. △

Exercise 6.35 Let X and Y have joint density function

$$f(x, y) = \begin{cases} cx & \text{if } 0 < y < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of the constant c and the marginal density functions of X and Y. Are X and Y independent?

Exercise 6.36 Random variables X, Y, and Z have joint density function

$$f(x, y, z) = \begin{cases} 8xyz & \text{if } 0 < x, y, z < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Are X, Y, and Z independent? Find $\mathbb{P}(X > Y)$ and $\mathbb{P}(Y > Z)$.

6.4 Sums of continuous random variables

We often need to know the density function of the sum $Z = X + Y$ of two jointly continuous random variables. The density function of Z is the derivative of the distribution function of Z, and so we calculate this first. Suppose that X and Y have joint density function $f_{X,Y}$. Then

$$\begin{aligned}\mathbb{P}(Z \le z) &= \mathbb{P}(X + Y \le z) \\ &= \iint_A f_{X,Y}(x, y)\, dx\, dy\end{aligned}$$

by Theorem 6.22, where $A = \{(x, y) \in \mathbb{R}^2 : x + y \le z\}$. Writing in the limits of integration, we find that

$$\begin{aligned}\mathbb{P}(Z \le z) &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{z-x} f_{X,Y}(x, y)\, dx\, dy \\ &= \int_{v=-\infty}^{z} \int_{u=-\infty}^{\infty} f_{X,Y}(u, v-u)\, du\, dv\end{aligned}$$

by the substitution $u = x$, $v = x + y$.[2] Differentiate this equation with respect to z, where possible, to obtain

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(u, z-u)\, du. \tag{6.37}$$

An important special case is when X and Y are independent, for which the following theorem is an immediate consequence of (6.37).

Theorem 6.38 (Convolution formula) *If the random variables X and Y are independent and continuous with density functions f_X and f_Y, then the density function of $Z = X + Y$ is*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx \qquad \text{for } z \in \mathbb{R}. \tag{6.39}$$

In the language of analysis, equation (6.39) says that f_Z is the *convolution* of f_X and f_Y, written $f_Z = f_X * f_Y$.

Example 6.40 Let X and Y be independent random variables having, respectively, the gamma distribution with parameters s and λ, and the gamma distribution with parameters t and λ. Then $Z = X + Y$ has density function

$$\begin{aligned}f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx \\ &= \begin{cases} \displaystyle\int_0^z f_X(x) f_Y(z-x)\, dx & \text{if } z > 0, \\ 0 & \text{otherwise,} \end{cases}\end{aligned}$$

[2] This is a simple change of variables in two dimensions. Those readers not familiar with such transformations in more than one dimension should read on to the next section.

since $f_X(x)f_Y(z-x) = 0$ unless $x > 0$ and $z - x > 0$. Thus, for $z > 0$,

$$\begin{aligned} f_Z(z) &= \int_0^z \frac{1}{\Gamma(s)}\lambda(\lambda x)^{s-1}e^{-\lambda x}\frac{1}{\Gamma(t)}\lambda[\lambda(z-x)]^{t-1}e^{-\lambda(z-x)}\,dx \\ &= Ae^{-\lambda z}\int_0^z x^{s-1}(z-x)^{t-1}\,dx, \end{aligned}$$

where

$$A = \frac{1}{\Gamma(s)\Gamma(t)}\lambda^{s+t}.$$

Substitute $y = x/z$ in the last integral to obtain

$$f_Z(z) = Bz^{s+t-1}e^{-\lambda z} \qquad \text{for } z > 0, \tag{6.41}$$

where B is a constant given by

$$B = \frac{1}{\Gamma(s)\Gamma(t)}\lambda^{s+t}\int_0^1 y^{s-1}(1-y)^{t-1}\,dy. \tag{6.42}$$

The only distribution with density function of the form (6.41) is the gamma distribution with parameters $s + t$ and λ, and it follows that the constant B satisfies

$$B = \frac{1}{\Gamma(s+t)}\lambda^{s+t}. \tag{6.43}$$

A glance at (5.40) confirms this. Our principal conclusion is that the sum of two independent gamma-distributed random variables, with parameters s, λ and t, λ, respectively, has the gamma distribution with parameters $s + t, \lambda$. We have a subsidiary conclusion also: by comparison of (6.42) and (6.43),

$$\int_0^1 y^{s-1}(1-y)^{t-1}\,dy = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)} \qquad \text{for } s, t > 0. \tag{6.44}$$

This well known formula arose earlier as the normalizing constant for the beta distribution (5.42). △

Exercise 6.45 If X and Y have joint density function

$$f(x, y) = \begin{cases} \frac{1}{2}(x+y)e^{-x-y} & \text{if } x, y > 0, \\ 0 & \text{otherwise,} \end{cases}$$

find the density function of $X + Y$.

Exercise 6.46 If X and Y are independent random variables having the χ^2 distribution with m and n degrees of freedom, respectively, prove that $X+Y$ has the χ^2 distribution with $m+n$ degrees of freedom.

Exercise 6.47 If X and Y are independent random variables, each having the normal distribution with mean 0 and variance 1, find the distribution of $X + Y$.

6.5 Changes of variables

The following type of question arises commonly: if X and Y are random variables and $u, v : \mathbb{R}^2 \to \mathbb{R}$, what can be said about the joint distribution of the pair (U, V) of random variables given by $U = u(X, Y)$, $V = v(X, Y)$? We present an answer to this question in the particular case when X and Y are jointly continuous and the functions u and v satisfy certain conditions which allow us to use the usual theory of changes of variables within an integral. Let T be the mapping from $\mathbb{R}^2$ into $\mathbb{R}^2$ given by $T(x, y) = (u, v)$, where $u = u(x, y)$ and $v = v(x, y)$, and suppose that T is a bijection between some domain $D \subseteq \mathbb{R}^2$ and some range $S \subseteq \mathbb{R}^2$. Then T may be inverted to obtain a bijection $T^{-1} : S \to D$. That is, for each $(u, v) \in S$, there exists a point $(x, y) = T^{-1}(u, v)$ in D, and we write $x = x(u, v)$ and $y = y(u, v)$. The *Jacobian* of T^{-1} is defined to be the determinant

$$J(u, v) = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v}, \tag{6.48}$$

and we suppose that these derivatives exist and are continuous at all points in S. The standard theory of multiple integrals tells us how to change variables within the integral: if $g : \mathbb{R}^2 \to \mathbb{R}^2$, then, for any sufficiently regular subset A of D and any integrable function g,

$$\iint_A g(x, y)\, dx\, dy = \iint_{T(A)} g\big(x(u, v), y(u, v)\big)|J(u, v)|\, du\, dv, \tag{6.49}$$

where $T(A)$ is the image of A under T.

Theorem 6.50 (Jacobian formula) *Let X and Y be jointly continuous with joint density function $f_{X,Y}$, and let $D = \{(x, y) : f_{X,Y}(x, y) > 0\}$. If the mapping T given by $T(x, y) = (u(x, y), v(x, y))$ is a bijection from D to the set $S \subseteq \mathbb{R}^2$, then (subject to the previous conditions) the pair $(U, V) = (u(X, Y), v(X, Y))$ is jointly continuous with joint density function*

$$f_{U,V}(u, v) = \begin{cases} f_{X,Y}\big(x(u, v), y(u, v)\big)|J(u, v)| & \text{if } (u, v) \in S, \\ 0 & \text{otherwise.} \end{cases} \tag{6.51}$$

Proof You should not worry overmuch about the details of this argument. Suppose that $A \subseteq D$ and $T(A) = B$. Since $T : D \to S$ is a bijection,

$$\mathbb{P}\big((U, V) \in B\big) = \mathbb{P}\big((X, Y) \in A\big). \tag{6.52}$$

However,

$$\begin{aligned} \mathbb{P}\big((X, Y) \in A\big) &= \iint_A f_{X,Y}(x, y)\, dx\, dy && \text{by Theorem 6.22} \\ &= \iint_B f_{X,Y}\big(x(u, v), y(u, v)\big)|J(u, v)|\, du\, dv && \text{by (6.49)} \\ &= \mathbb{P}\big((U, V) \in B\big) && \text{by (6.52).} \end{aligned}$$

This holds for any $B \subseteq S$, and another glance at Theorem 6.22 gives the result. □

Although the statement of Theorem 6.50 looks forbidding, it is not difficult to apply in practice, although it is necessary to check that the mapping in question is a bijection. Here is an example.

Example 6.53 Let X and Y have joint density function

$$f(x, y) = \begin{cases} e^{-x-y} & \text{if } x, y > 0, \\ 0 & \text{otherwise,} \end{cases}$$

and let $U = X + Y$ and $V = X/(X + Y)$. Find the joint density function of U and V and the marginal density function of V.

Solution The mapping T of this problem is given by $T(x, y) = (u, v)$, where

$$u = x + y, \qquad v = \frac{x}{x + y},$$

and T is a bijection from $D = \{(x, y) : x, y > 0\}$ to $S = \{(u, v) : 0 < u < \infty,\ 0 < v < 1\}$. It has inverse $T^{-1}(u, v) = (x, y)$, where

$$x = uv, \qquad y = u(1 - v).$$

The Jacobian of T^{-1} is

$$\begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ (1 - v) & -u \end{vmatrix} = -u,$$

giving by (6.51) that U and V have joint density function

$$f_{U,V}(u, v) = \begin{cases} ue^{-u} & \text{if } u > 0 \text{ and } 0 < v < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The marginal density function of V is

$$\begin{aligned} f_V(v) &= \int_{-\infty}^{\infty} f_{U,V}(u, v)\, du \\ &= \begin{cases} \displaystyle\int_0^{\infty} ue^{-u}\, du = 1 & \text{if } 0 < v < 1, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

so that V is uniformly distributed on $(0, 1)$. It may in addition be shown that U and V are independent, and U has the gamma distribution with parameters 2 and 1. △

Exercise 6.54 Let X and Y be independent random variables, each having the normal distribution with mean μ and variance σ^2. Find the joint density function of $U = X - Y$ and $V = X + Y$. Are U and V independent?

Exercise 6.55 Let X and Y be random variables with joint density function

$$f(x, y) = \begin{cases} \frac{1}{4}e^{-\frac{1}{2}(x+y)} & \text{if } x, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Show that the joint density function of $U = \frac{1}{2}(X - Y)$ and $V = Y$ is

$$f_{U,V}(u, v) = \begin{cases} \frac{1}{2}e^{-u-v} & \text{if } (u, v) \in A, \\ 0 & \text{otherwise,} \end{cases}$$

where A is a region of the (u, v) plane to be determined. Deduce that U has the bilateral exponential distribution with density function

$$f_U(u) = \tfrac{1}{2}e^{-|u|} \qquad \text{for } u \in \mathbb{R}.$$

6.6 Conditional density functions

Let us suppose that X and Y are jointly continuous random variables with joint density function $f_{X,Y}$. To obtain the marginal density function f_Y of Y, we 'average' over all possible values of X,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx,$$

and this is the calculation which we perform if we care about Y only and have no information about the value taken by X. A contrasting situation arises if we have full information about the value taken by X, say if we are given that X takes the value x. This information has consequences for the distribution of Y, and it is this '*conditional* distribution of Y given that $X = x$' that concerns us in this section. We cannot calculate $\mathbb{P}(Y \leq y \mid X = x)$ from the usual formula $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ since $\mathbb{P}(B) = 0$ in this case, and so we proceed as follows. Instead of conditioning on the event that $X = x$, we condition on the event that $x \leq X \leq x + \delta x$ and take the limit as $\delta x \downarrow 0$. Thus,

$$\begin{aligned} \mathbb{P}\big(Y \leq y \,\big|\, x \leq X \leq x + \delta x\big) &= \frac{\mathbb{P}(Y \leq y,\ x \leq X \leq x + \delta x)}{\mathbb{P}(x \leq X \leq x + \delta x)} \\ &= \frac{\int_{u=x}^{x+\delta x} \int_{v=-\infty}^{y} f_{X,Y}(u, v)\, du\, dv}{\int_{x}^{x+\delta x} f_X(u)\, du} \end{aligned}$$

by Theorems 6.22 and 5.27. We divide both the numerator and the denominator by δx and take the limit as $\delta x \downarrow 0$ to obtain

$$\begin{aligned} \mathbb{P}\big(Y \leq y \,\big|\, x \leq X \leq x + \delta x\big) &\to \int_{-\infty}^{y} \frac{f_{X,Y}(x, v)}{f_X(x)}\, dv \\ &= G(y), \end{aligned} \tag{6.56}$$

say. It is clear from (6.56) that G is a distribution function with density function

$$g(y) = \frac{f_{X,Y}(x, y)}{f_X(x)} \qquad \text{for } y \in \mathbb{R},$$

and we call G and g the 'conditional distribution function' and the 'conditional density function' of Y given that X equals x. The above discussion is valid only for values of x such that $f_X(x) > 0$, and so we make the following formal definition.

Definition 6.57 *The **conditional density function of** Y **given that** $X = x$ is denoted by $f_{Y|X}(\cdot \mid x)$ and defined by*

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} \tag{6.58}$$

for $y \in \mathbb{R}$ and x satisfying $f_X(x) > 0$.

We emphasize that expressions such as $\mathbb{P}(Y \leq y \mid X = x)$ cannot be interpreted in the usual way by using the formula for $\mathbb{P}(A \mid B)$. The only way of giving meaning to such a quantity is to make a new definition, such as: $\mathbb{P}(Y \leq y \mid X = x)$ is defined to be the conditional distribution function $G(y)$ of Y given $X = x$, as in (6.56).

If X and Y are independent, then $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. By (6.58), $f_{Y|X}(y \mid x) = f_Y(y)$, which is to say that information about X is irrelevant when studying Y.

Example 6.59 Let X and Y have joint density function

$$f(x, y) = \begin{cases} 2e^{-x-y} & \text{if } 0 < x < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

The marginal density functions are

$$f_X(x) = 2e^{-2x} \qquad \text{for } x > 0,$$
$$f_Y(y) = 2e^{-y}(1 - e^{-y}) \qquad \text{for } y > 0,$$

where it is understood that these functions take the value 0 off the specified domains. The conditional density function of Y given $X = x$ (> 0) is

$$f_{Y|X}(y \mid x) = \frac{2e^{-x-y}}{2e^{-2x}} = e^{x-y} \qquad \text{for } y > x.$$

The conditional density function of X given $Y = y$ is

$$f_{X|Y}(x \mid y) = \frac{e^{-x}}{1 - e^{-y}} \qquad \text{for } 0 < x < y.$$

It is clear that both these conditional density functions equal 0 if $x > y$. △

Exercise 6.60 Suppose that X and Y have joint density function

$$f(x, y) = \begin{cases} e^{-y} & \text{if } 0 < x < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Find the conditional density functions of X given that $Y = y$, and of Y given that $X = x$.

Exercise 6.61 Let X and Y be independent random variables, each having the exponential distribution with parameter λ. Find the joint density function of X and $X+Y$, and deduce that the conditional density function of X, given that $X + Y = a$, is uniform on the interval $(0, a)$ for each $a > 0$. In other words, the knowledge that $X + Y = a$ provides no useful clue about the position of X in the interval $(0, a)$.

6.7 Expectations of continuous random variables

Let X and Y be jointly continuous random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $g : \mathbb{R}^2 \to \mathbb{R}$. As in the discussion of Section 5.5, we shall suppose that the mapping $Z : \Omega \to \mathbb{R}$, defined by $Z(\omega) = g(X(\omega), Y(\omega))$, is a random variable (this is certainly the case if g is sufficiently well behaved). In calculating the expectation of Z, we do not have to find the distribution of Z explicitly.

Theorem 6.62 *We have that*

$$\mathbb{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, dx\, dy,$$

whenever this integral converges absolutely.

We do not prove this, but note that the result follows intuitively from the corresponding result, Theorem 3.10, for discrete random variables, by exploiting the analogy between joint mass functions and joint density functions.

Using Theorem 6.62, we find that the expectation operator acts linearly on the space of continuous random variables, which is to say that

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y) \tag{6.63}$$

whenever $a, b \in \mathbb{R}$ and X and Y are jointly continuous random variables with means $\mathbb{E}(X)$ and $\mathbb{E}(Y)$. This follows from Theorem 6.62 by writing

$$\begin{aligned} \mathbb{E}(aX + bY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by) f_{X,Y}(x, y)\, dx\, dy \\ &= a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y)\, dx\, dy + b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y)\, dx\, dy \\ &= a \int_{-\infty}^{\infty} x f_X(x)\, dx + b \int_{-\infty}^{\infty} y f_Y(y)\, dy \\ &= a\mathbb{E}(X) + b\mathbb{E}(Y). \end{aligned}$$

We mention one common error here. It is a mistake to demand that X and Y be independent in order that $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. This equation holds *whether or not* X and Y are independent.

Next we discuss the relationship between expectation and independence, noting the excellent analogy with Theorems 3.19 and 3.20 dealing with discrete random variables. First, if X and Y are independent random variables with joint density function $f_{X,Y}$, then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \tag{6.64}$$

whenever these expectations exist, since

$$\begin{aligned}\mathbb{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} x f_X(x)\,dx \int_{-\infty}^{\infty} y f_Y(y)\,dy \qquad \text{by independence} \\ &= \mathbb{E}(X)\mathbb{E}(Y). \end{aligned} \tag{6.65}$$

The converse is false: there exist jointly continuous dependent random variables X and Y for which $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. The correct and full result here is the next theorem.

Theorem 6.66 *Jointly continuous random variables X and Y are independent if and only if*

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)) \tag{6.67}$$

for all functions $g, h : \mathbb{R} \to \mathbb{R}$ for which these expectations exist.

Proof If X and Y are independent, then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, and (6.67) holds by Theorem 6.62. Conversely, if (6.67) holds for all appropriate functions g and h, then it holds in particular for the functions given by

$$g(u) = \begin{cases} 1 & \text{if } u \le x, \\ 0 & \text{if } u > x, \end{cases} \qquad h(v) = \begin{cases} 1 & \text{if } v \le y, \\ 0 & \text{if } v > y, \end{cases}$$

for fixed values of x and y. In this case, $g(X)h(Y)$ is a discrete random variable with the Bernoulli distribution, parameter $p_1 = \mathbb{P}(X \le x,\ Y \le y)$, and $g(X)$ and $h(Y)$ are Bernoulli random variables with parameters $p_2 = \mathbb{P}(X \le x)$ and $p_3 = \mathbb{P}(Y \le y)$, respectively. Hence

$$\mathbb{E}(g(X)h(Y)) = \mathbb{P}(X \le x,\ Y \le y)$$

by (2.28), and

$$\mathbb{E}(g(X)) = \mathbb{P}(X \le x), \qquad \mathbb{E}(h(Y)) = \mathbb{P}(Y \le y),$$

giving by (6.67) that

$$\mathbb{P}(X \le x,\ Y \le y) = \mathbb{P}(X \le x)\mathbb{P}(Y \le y) \qquad \text{for } x, y \in \mathbb{R},$$

as required. □

We turn now to conditional expectation. Suppose that X and Y are jointly continuous random variables with joint density function $f_{X,Y}$, and that we are given that $X = x$. In light of this information, the new density function of Y is the conditional density function $f_{Y|X}(\cdot \mid x)$.

Definition 6.68 *The **conditional expectation of** Y **given** $X = x$, written $\mathbb{E}(Y \mid X = x)$, is the mean of the conditional density function,*

$$\mathbb{E}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{X|Y}(y \mid x)\, dy = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x, y)}{f_X(x)}\, dy,$$

valid for any value of x for which $f_X(x) > 0$.

Possibly the most useful application of conditional expectation is the next theorem, a form of the partition theorem which enables us to calculate $\mathbb{E}(Y)$ in situations where the conditional expectations $\mathbb{E}(Y \mid X = x)$ are easily calculated.

Theorem 6.69 *If X and Y are jointly continuous random variables, then*

$$\mathbb{E}(Y) = \int \mathbb{E}(Y \mid X = x) f_X(x)\, dx,$$

where the integral is over all values of x such that $f_X(x) > 0$.

In other words, in calculating $\mathbb{E}(Y)$ we may first fix the value of X and then average over this value later.

Proof This is straightforward:

$$\begin{aligned}
\mathbb{E}(Y) &= \int y f_Y(y)\, dy = \iint y f_{X,Y}(x, y)\, dx\, dy \\
&= \iint y f_{Y|X}(y \mid x) f_X(x)\, dx\, dy \qquad \text{by (6.58)} \\
&= \int \left(\int y f_{Y|X}(y \mid x)\, dy \right) f_X(x)\, dx
\end{aligned}$$

as required. The integrals here range over all appropriate values of x and y. □

Exercise 6.70 Let the pair (X, Y) be uniformly distributed on the unit disc, so that

$$f_{X,Y}(x, y) = \begin{cases} \pi^{-1} & \text{if } x^2 + y^2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{E}\sqrt{X^2 + Y^2}$ and $\mathbb{E}(X^2 + Y^2)$.

Exercise 6.71 Give an example of a pair of dependent and jointly continuous random variables X, Y for which $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Exercise 6.72 If X and Y have joint density function

$$f(x, y) = \begin{cases} e^{-y} & \text{if } 0 < x < y < \infty, \\ 0 & \text{otherwise,} \end{cases}$$

find $\mathbb{E}(X \mid Y = y)$ and $\mathbb{E}(Y \mid X = x)$.

6.8 Bivariate normal distribution

The 'univariate' normal distribution N(0, 1) has density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad \text{for } x \in \mathbb{R}.$$

The corresponding bivariate distribution has a joint density function of a similar form. Let $\rho \in (-1, 1)$, and let f be the function of two variables given by

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2) \right) \qquad \text{for } x, y \in \mathbb{R}. \quad (6.73)$$

Clearly, $f(x, y) \geq 0$ for all x and y, and it is the case that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

(the reader should check this), giving that f is a joint density function; it is called the joint density function of the *standard bivariate normal (or Gaussian) distribution*. Suppose that X and Y are random variables with the standard bivariate normal density function f. We calculate next

(a) the marginal density function of X,
(b) the conditional density function of Y given $X = x$,
(c) the conditional expectation of Y given $X = x$,
(d) a condition under which X and Y are independent.

Marginals. The marginal density function of X is

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y)\, dy \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[(y - \rho x)^2 + x^2(1-\rho^2) \right] \right) dy \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)} \right) dy. \end{aligned}$$

Note the completion of the square in the exponent. The function within the final integral is the density function of the normal distribution with mean ρx and variance $1 - \rho^2$, and therefore this final integral equals 1, giving that

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad \text{for } x \in \mathbb{R}. \tag{6.74}$$

We conclude that X has the normal distribution with mean 0 and variance 1. By symmetry, Y has this distribution also.

Conditional density function. The conditional density function of Y given that $X = x$ is

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)} = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)} \right),$$

and so the conditional distribution of Y given $X = x$ is the normal distribution with mean ρx and variance $1 - \rho^2$.

Conditional expectation. By the above, the conditional expectation of Y given $X = x$ is

$$\mathbb{E}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\, dy = \rho x. \tag{6.75}$$

Independence. The random variables X and Y are independent if and only if $f(x, y)$ factorizes as the product of a function of x and a function of y. This happens (by a glance at (6.73)) if and only if $\rho = 0$. The constant ρ occurs in another way also. We may calculate $\mathbb{E}(XY)$ by applying Theorem 6.69 to the random variables X and XY to obtain

$$\begin{aligned}
\mathbb{E}(XY) &= \int_{-\infty}^{\infty} \mathbb{E}(XY \mid X = x) f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} x \mathbb{E}(Y \mid X = x) f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \rho x^2 f_X(x)\, dx && \text{by (6.75)} \\
&= \rho \mathbb{E}(X^2) = \rho \operatorname{var}(X) && \text{since } \mathbb{E}(X) = 0 \\
&= \rho
\end{aligned}$$

by the remarks after (6.74). Also, $\mathbb{E}(X) = \mathbb{E}(Y) = 0$, giving that

$$\rho = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

We deduce that X and Y are independent if and only if $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Thus, by the discussion around (6.65), for random variables X and Y with the bivariate normal distribution, *X and Y are independent if and only if* $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. This is commonly expressed by saying that such random variables are independent if and only if they are uncorrelated (see the account of covariance and correlation in the forthcoming Section 7.3).

The marginals of the standard bivariate normal distribution are N(0, 1). Here is a more general bivariate distribution. Let g be the function of two variables given by

$$g(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}Q(x,y)} \qquad \text{for } x, y \in \mathbb{R}, \tag{6.76}$$

where Q is the quadratic form

$$Q(x, y) = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right] \tag{6.77}$$

and $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$, $-1 < \rho < 1$. The standard bivariate normal distribution is obtained by setting $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$. It is not difficult but slightly tedious to show that g is a joint density function, and the corresponding distribution is called the *bivariate normal distribution* with the appropriate parameters. We leave it for Exercise 6.79 to show that, if X and Y have joint density function g, then the pair U, V given by

$$U = \frac{X-\mu_1}{\sigma_1}, \qquad V = \frac{Y-\mu_2}{\sigma_2}, \tag{6.78}$$

has the standard bivariate normal distribution with parameter ρ.

The general bivariate normal distribution of (6.76)–(6.77) has a complicated form. There is however a simpler definition of substantial appeal, namely the following. A pair (X, Y) of random variables is said to have a bivariate normal distribution if, for all $a, b \in \mathbb{R}$, the linear combination $aX + bY$ has a univariate normal distribution. It is not difficult to show that this is equivalent to the definition given above, so long as one allows degenerate normal distributions with zero variances.

This characterization of normal distributions is valuable, especially when extending the theory from two variables to a general number. More generally, a vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is said to have a multivariate normal distribution if, for all $\mathbf{a} \in \mathbb{R}^n$, the scalar product $\mathbf{a} \cdot \mathbf{X}'$ has a univariate normal distribution.

Exercise 6.79 Let the pair (X, Y) have the bivariate normal density function of (6.76), and let U and V be given by (6.78). Show that U and V have the standard bivariate normal distribution. Hence or otherwise show that

$$\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \rho\sigma_1\sigma_2,$$

and that

$$\mathbb{E}(Y \mid X = x) = \mu_2 + \rho\sigma_2(x - \mu_1)/\sigma_1.$$

Exercise 6.80 Let the pair (X, Y) have the bivariate normal distribution of (6.76), and let $a, b \in \mathbb{R}$. Show that $aX + bY$ has a univariate normal distribution, possibly wth zero variance.

6.9 Problems

1. If X and Y are independent random variables with density functions f_X and f_Y, respectively, show that $U = XY$ and $V = X/Y$ have density functions

$$f_U(u) = \int_{-\infty}^{\infty} f_X(x) f_Y(u/x) \frac{1}{|x|}\, dx, \qquad f_V(v) = \int_{-\infty}^{\infty} f_X(vy) f_Y(y) |y|\, dy.$$

2. Is the function G, defined by

$$G(x, y) = \begin{cases} 1 & \text{if } x + y \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

the joint distribution function of some pair of random variables? Justify your answer.

3. Let (X, Y, Z) be a point chosen uniformly at random in the unit cube $(0, 1)^3$. Find the probability that the quadratic equation $Xt^2 + Yt + Z = 0$ has two distinct real roots.
4. Show that if X and Y are independent random variables having the exponential distribution with parameters λ and μ, respectively, then $\min\{X, Y\}$ has the exponential distribution with parameter $\lambda + \mu$.
5. *Lack-of-memory property.* If X has the exponential distribution, show that

$$\mathbb{P}\big(X > u + v \,\big|\, X > u\big) = \mathbb{P}(X > v) \qquad \text{for } u, v > 0.$$

   This is called the 'lack of memory' property, since it says that, if we are given that $X > u$, then the distribution of $X - u$ is the same as the original distribution of X. Show that if Y is a positive, continuous random variable with the lack-of-memory property above, then Y has the exponential distribution.
6. Let $X_1, X_2, \ldots, X_n$ be independent random variables, each having distribution function F and density function f. Find the distribution function of U and the density functions of U and V, where $U = \min\{X_1, X_2, \ldots, X_n\}$ and $V = \max\{X_1, X_2, \ldots, X_n\}$. Show that the joint density function of U and V is

$$f_{U,V}(u, v) = n(n-1)f(u)f(v)\big[F(v) - F(u)\big]^{n-2} \qquad \text{if } u < v.$$

7. Let $X_1, X_2, \ldots$ be independent, identically distributed, continuous random variables. Define N as the index such that

$$X_1 \geq X_2 \geq \cdots \geq X_{N-1} \quad \text{and} \quad X_{N-1} < X_N.$$

   Prove that $\mathbb{P}(N = k) = (k-1)/k!$ and that $\mathbb{E}(N) = e$.
8. Show that there exists a constant c such that the function

$$f(x, y) = \frac{c}{(1 + x^2 + y^2)^{3/2}} \qquad \text{for } x, y \in \mathbb{R}$$

   is a joint density function. Show that both marginal density functions of f are the density function of the Cauchy distribution.
9. Let X and Y have joint density function

$$f(x, y) = \begin{cases} \frac{1}{4}(x + 3y)e^{-(x+y)} & \text{if } x, y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

   Find the marginal density function of Y. Show that $\mathbb{P}(Y > X) = \frac{5}{8}$.
10. Let S_n be the sum of n independent, identically distributed random variables having the exponential distribution with parameter λ. Show that S_n has the gamma distribution with parameters n and λ.

    For given $t > 0$, show that $N_t = \max\{n : S_n \leq t\}$ has a Poisson distribution.
11. An aeroplane drops medical supplies to two duellists. With respect to Cartesian coordinates whose origin is at the target point, both the x and y coordinates of the landing point of the supplies have normal distributions which are independent. These two distributions have the same mean 0 and variance σ^2. Show that the expectation of the distance between the landing point and the target is $\sigma\sqrt{\pi/2}$. What is the variance of this distance? (Oxford 1976M)

12. X and Y are independent random variables normally distributed with mean zero and variance σ^2. Find the expectation of $\sqrt{X^2+Y^2}$. Find the probabilities of the following events, where a, b, c, and α are positive constants such that $b < c$ and $\alpha < \frac{1}{2}\pi$:

    (a) $\sqrt{X^2+Y^2} < a$,
    (b) $0 < \tan^{-1}(Y/X) < \alpha$ and $Y > 0$.

    (Consider various cases depending on the relative sizes of a, b, and c.) (Oxford 1981M)

13. The independent random variables X and Y are both exponentially distributed with parameter λ, that is, each has density function

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t > 0, \\ 0 & \text{otherwise.} \end{cases}$$

    (a) Find the (cumulative) distribution and density functions of the random variables $1-e^{-\lambda X}$, $\min\{X, Y\}$, and $X - Y$.
    (b) Find the probability that $\max\{X, Y\} \leq aX$, where a is a real constant.

    (Oxford 1982M)

14. The independent random variables X and Y are normally distributed with mean 0 and variance 1.

    (a) Show that $W = 2X - Y$ is normally distributed, and find its mean and variance.
    (b) Find the mean of $Z = X^2/(X^2+Y^2)$.
    (c) Find the mean of V/U, where $U = \max\{|X|, |Y|\}$ and $V = \min\{|X|, |Y|\}$.

    (Oxford 1985M)

15. Let X and Y be independent random variables, X having the normal distribution with mean 0 and variance 1, and Y having the χ^2 distribution with n degrees of freedom. Show that

$$T = \frac{X}{\sqrt{Y/n}}$$

    has density function

$$f(t) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma(\frac{1}{2}(n+1))}{\Gamma(\frac{1}{2}n)} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}(n+1)} \qquad \text{for } t \in \mathbb{R}.$$

    T is said to have the *t-distribution* with n degrees of freedom.

16. Let X and Y be independent random variables with the χ^2 distribution, X having m degrees of freedom and Y having n degrees of freedom. Show that

$$U = \frac{X/m}{Y/n}$$

    has density function

$$f(u) = \frac{m\Gamma(\frac{1}{2}(m+n))}{n\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)} \cdot \frac{(mu/n)^{\frac{1}{2}m-1}}{\left[1+(mu/n)\right]^{\frac{1}{2}(m+n)}} \qquad \text{for } u > 0.$$

    U is said to have the *F-distribution* with m and n degrees of freedom.

17. In a sequence of dependent Bernoulli trials, the conditional probability of success at the ith trial, given that all preceding trials have resulted in failure, is p_i $(i = 1, 2, \dots)$. Give an expression in terms of the p_i for the probability that the first success occurs at the nth trial.

    Suppose that $p_i = 1/(i+1)$ and that the time intervals between successive trials are independent random variables, the interval between the $(n-1)$th and the nth trials being exponentially distributed with density $n^\alpha \exp(-n^\alpha x)$, where α is a given constant. Show that the expected time to achieve the first success is finite if and only if $\alpha > 0$. (Oxford 1975F)

18. Let $a, b > 0$. Independent positive random variables X and Y have probability densities

$$\frac{1}{\Gamma(a)} x^{a-1} e^{-x}, \qquad \frac{1}{\Gamma(b)} y^{b-1} e^{-y}, \qquad \text{for } x, y \geq 0,$$

    respectively, and U and V are defined by

$$U = X + Y, \qquad V = \frac{X}{X+Y}.$$

    Prove that U and V are independent, and find their distributions.

    Deduce that

$$\mathbb{E}\left(\frac{X}{X+Y}\right) = \frac{\mathbb{E}(X)}{\mathbb{E}(X) + \mathbb{E}(Y)}.$$

    (Oxford 1971F)

19. Let X_1, X_2, X_3 be independent χ^2 random variables with r_1, r_2, r_3 degrees of freedom.

    (a) Show that $Y_1 = X_1/X_2$ and $Y_2 = X_1 + X_2$ are independent and that Y_2 is a χ^2 random variable with $r_1 + r_2$ degrees of freedom.
    (b) Deduce that the following random variables are independent:

$$\frac{X_1/r_1}{X_2/r_2} \qquad \text{and} \qquad \frac{X_3/r_3}{(X_1+X_2)/(r_1+r_2)}.$$

    (Oxford 1982F)

20. Let X and Y be random variables with the vector (X, Y) uniformly distributed on the region $R = \{(x, y) : 0 < y < x < 1\}$. Write down the joint probability density function of (X, Y). Find $\mathbb{P}(X + Y < 1)$.

    Find the probability density function $f_X(x)$ of X, and find also $\mathbb{E}(X)$. Find the conditional probability density function $f_{Y|X}(y \mid x)$ of Y given that $X = x$, and find also $\mathbb{E}(Y \mid X = x)$. (Oxford 2005)

21. Let X and Y be independent random variables, each uniformly distributed on $[0, 1]$. Let $U = \min\{U, V\}$ and $V = \max\{U, V\}$. Show that $\mathbb{E}(U) = \frac{1}{3}$, and hence find the covariance of U and V. (Cambridge 2007)

* 22. Three crew members of Dr Who's spacecraft Tardis are teleported to the surface of the spherical planet Zog. Their positions X, Y, Z are independent and uniformly distributed on the surface. Find the probability density function of the angle $\widehat{XCY}$, where C is the centre of Zog.

    Two people positioned on the surface at A and B are in direct radio communication if and only if $\widehat{ACB} < \frac{1}{2}\pi$.

    (a) Find the probability that Z is in direct radio communication with either X or Y, conditional on the event that $\phi := \widehat{XCY}$ satisfies $\phi < \frac{1}{2}\pi$.

(b) Find the probability that Z is in direct radio communication with both X and Y, conditional on the event that $\phi > \frac{1}{2}\pi$.

Deduce that the probability that all three crew members can keep in touch is $(\pi + 2)/(4\pi)$.

* 23. *Zog continued.* This time, n members of Dr Who's crew are transported to Zog, their positions being independent and uniformly distributed on the surface. In addition, Dr Who is required to choose a place W on the surface for his own transportation. Find the probability that, for every W, he is able to communicate with some member of his crew.

24. Let X and Y be independent non-negative random variables with densities f and g, respectively. Find the joint density function of $U = X$ and $V = X + aY$, where a is a positive constant.

Let X and Y be independent and exponentially distributed random variables, each with density

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

Find the density of $X + \frac{1}{2}Y$. Is it the same as the density of $\max\{X, Y\}$? (Cambridge 2007)

25. Let X and Y have the bivariate normal density function

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\} \qquad \text{for } x, y \in \mathbb{R},$$

for fixed $\rho \in (-1, 1)$. Let $Z = (Y - \rho X)/\sqrt{1-\rho^2}$. Show that X and Z are independent N(0, 1) variables. Hence or otherwise determine $\mathbb{P}(X > 0,\ Y > 0)$. (Cambridge 2008)

26. Let X and Y be random variables with the joint probability density function

$$f_{X,Y}(x, y) = \tfrac{1}{4}e^{-\frac{1}{2}(x+y)} \qquad \text{for } x, y > 0.$$

Show that the joint probability density function of $U = \frac{1}{2}(X - Y)$ and $V = Y$ is

$$f_{U,V}(u, v) = \begin{cases} \frac{1}{2}e^{-u-v} & \text{if } (u, v) \in A, \\ 0 & \text{otherwise,} \end{cases}$$

where A is a region of the (u, v) plane to be determined. Deduce that U has probability density function

$$f_U(u) = \tfrac{1}{2}e^{-|u|}, \qquad -\infty < u < \infty.$$

(Oxford 2008)

27. (a) Suppose that the continuous random variables X and Y are independent with probability density functions f and g, both of which are symmetric about zero.
(i) Find the joint probability density function of (U, V), where $U = X$ and $V = Y/X$.
(ii) Show that the marginal density function of V is

$$f_V(v) = 2\int_0^\infty x f(x) g(xv)\, dx.$$

(iii) Let X and Y be independent normal random variables, each with mean 0, and with non-zero variances a^2 and b^2, respectively. Show that $V = Y/X$ has probability density function

$$f_V(v) = \frac{c}{\pi(c^2 + v^2)} \qquad \text{for } -\infty < v < \infty,$$

where $c = b/a$. Hence find $\mathbb{P}(|Y| < |X|)$.

(b) Now let X and Y be independent random variables, each uniformly distributed on the interval $(0, 1)$. By considering the random variables $U = Y$ and $V = XY^2$, or otherwise, find the probability density function of V.

(Oxford 2010)

28. (a) Define the distribution function F of a random variable, and also its density function f, assuming F is differentiable. Show that

$$f(x) = -\frac{d}{dx}\mathbb{P}(X > x).$$

(b) Let U, V be independent random variables, each with the uniform distribution on $[0, 1]$. Show that

$$\mathbb{P}(V^2 > U > x) = \tfrac{1}{3} - x + \tfrac{2}{3}x^{3/2} \qquad \text{for } x \in (0, 1).$$

(c) What is the probability that the random quadratic equation $x^2 + 2Vx + U = 0$ has real roots?

(d) Given that the two roots R_1, R_2 of the above quadratic are real, what is the probability that both $|R_1| \leq 1$ and $|R_2| \leq 1$?

(Cambridge 2012)

7

Moments, and moment generating functions

Summary. Following a discussion of general random variables, the moments of a general distribution are defined. Covariance and correlation are introduced, and the Cauchy–Schwarz inequality is proved. The theory of moment generating functions may be viewed as an extension of the theory of probability generating functions. Special attention is given to the Markov and Jensen inequalities, and the chapter terminates with an account of characteristic functions.

7.1 A general note

Up to now, we have treated discrete and continuous random variables separately, and have hardly broached the existence of random variables which are neither discrete nor continuous. A brief overview of the material so far is depicted in Figure 7.1.

We cannot continue to treat the discrete case and the continuous case separately and the other cases not at all. The correct thing to be done at this point is to study random variables in their generality. Unfortunately, such a proper treatment is too advanced for this basic text since it involves defining the expectation of an arbitrary random variable, using ideas and techniques of abstract measure and integration theory. We are therefore forced to adopt another strategy. We shall try to state and prove theorems in ways which do not explicitly mention the type (discrete or continuous or ...) of the random variables involved; generally speaking, such arguments may be assumed to hold in the wide sense. Sometimes we will have to consider special cases, and then we shall normally treat *continuous* random variables. The discrete case is usually similar and easier, and a rule of thumb for converting an argument about continuous random variables into an argument about discrete random variables is to replace $f_X(x)\,dx$ by $p_X(x)$ and $\int$ by $\sum$.

Finally, curious readers may care to see the standard example of a random variable whose type is neither discrete nor continuous. Less enthusiastic readers should go directly to the next section.

Example 7.1 (The Cantor distribution) The celebrated Cantor set C is the often quoted example of an uncountable subset of the real line which is very sparse, in the sense that for any $\epsilon > 0$, there exist intervals $I_1, I_2, \ldots$ with total length less than ϵ such that $C \subseteq \bigcup_n I_n$. We construct this set as follows. Let $C_1 = [0, 1]$. Delete the middle third $(\frac{1}{3}, \frac{2}{3})$ of C_1 and

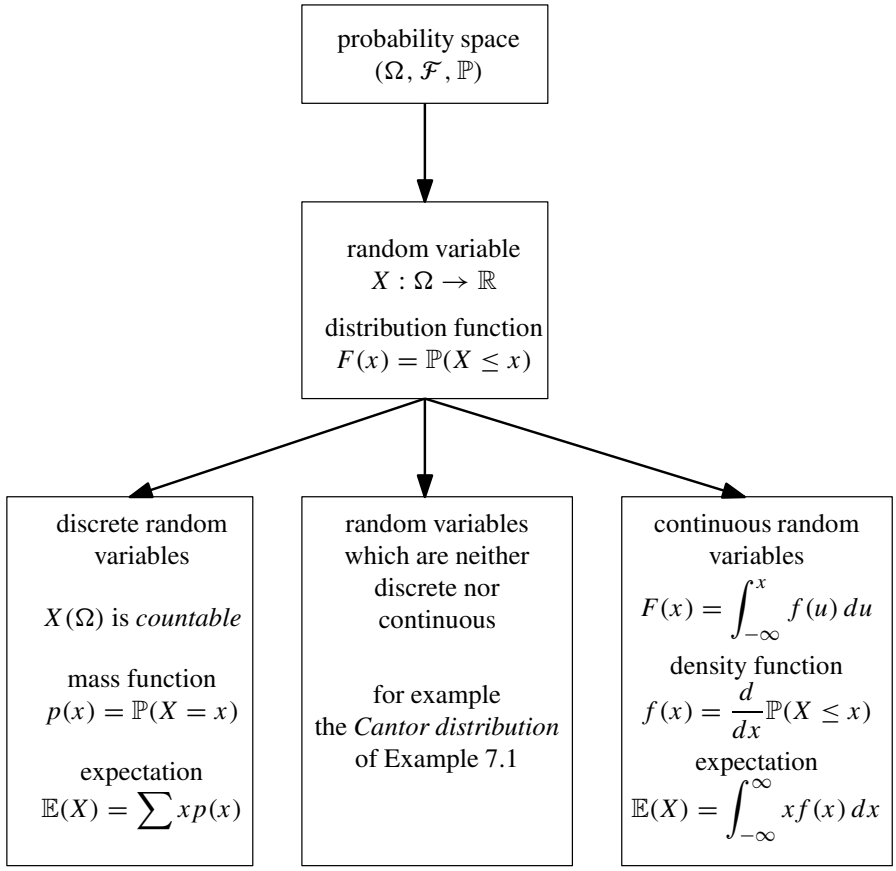


Fig. 7.1 An overview of probability spaces and random variables so far.

let $C_2 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$ be the remaining set. Next, delete the middle third in each of the two intervals comprising C_2 to obtain $C_3 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$, and continue similarly to obtain an infinite nested sequence $C_1 \supseteq C_2 \supseteq C_3 \supseteq \cdots$ of subsets of $[0, 1]$. The Cantor set C is defined to be the limit

$$C = \lim_{n\to\infty} C_n = \bigcap_{i=1}^{\infty} C_i.$$

There is another way of thinking about the Cantor set, and this is useful for us. Just as each number in $[0, 1]$ has an expansion in the base-10 system (namely its decimal expansion) so it has an expansion in the base-3 system. That is to say, any $x \in [0, 1]$ may be written in the form

$$x = \sum_{i=1}^{\infty} \frac{a_i}{3^i}, \tag{7.2}$$

where each of the a_i equals 0, 1, or 2. The Cantor set C is the set of all points $x \in [0, 1]$ for which the a_i above take the values 0 and 2 only.

We obtain the *Cantor distribution* as follows. Take $x \in C$ and express x in the form (7.2) with $a_i \in \{0, 2\}$ for all i. We define $F(x)$ by

$$F(x) = \sum_{i=1}^{\infty} \frac{a_i/2}{2^i}.$$

It is clear that

$$F(0) = 0, \quad F(1) = 1,$$

and F is non-decreasing in that

$$F(x) \leq F(y) \qquad \text{if } x \leq y.$$

Note that F is not a distribution function since it is defined on C only. However, we may extend the domain of F to the whole real line in the following natural way. If $x \in [0, 1] \setminus C$, x belongs to one of the intervals which were deleted in the construction of C. We define $F(x)$ to be the supremum of the set $\{F(y) : y \in C,\ y < x\}$. Finally, we set $F(x) = 0$ if $x < 0$ and $F(x) = 1$ if $x > 1$. It is fairly easy to see that F is a continuous non-decreasing function from $\mathbb{R}$ onto $[0, 1]$, and thus F is a distribution function.

Let X be a random variable with distribution function F. Clearly, X is not a discrete random variable, since F is continuous. It is not quite so easy to see that X cannot be continuous. Roughly speaking, this is because F is constant on each interval $(\frac{1}{3}, \frac{2}{3})$, $(\frac{1}{9}, \frac{2}{9})$, $(\frac{7}{9}, \frac{8}{9})$, ... that was deleted in constructing C. The total length of these intervals is

$$\tfrac{1}{3} + 2 \cdot \tfrac{1}{9} + 4 \cdot \tfrac{1}{27} + \cdots = \tfrac{1}{3} \sum_{i=0}^{\infty} (\tfrac{2}{3})^i = 1,$$

so that $F'(x) = 0$ for 'almost all' of $[0, 1]$. Thus, if F were to have density function f, then $f(x) = 0$ for 'almost all' x, giving that

$$\mathbb{P}(-\infty < X < \infty) = \int_{-\infty}^{\infty} f(x)\, dx = 0,$$

which is clearly absurd. Hence, X is neither discrete nor continuous. It turns out that the distribution function F is in an entirely new category, called the set of 'singular' distribution functions. Do not be too disturbed by this novelty; there are basically only three classes of distribution functions: those which are singular, those which arise from discrete random variables, and those which arise from continuous random variables. There is a theorem of Lebesgue called the 'decomposition theorem' that implies that every distribution function F may be expressed in the form $F = \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_3$ for non-negative α_i summing to 1, such that: F_1 is the distribution function of a discrete random variable, F_2 that of a continuous random variable, and F_3 is singular.[1] △

[1] See, for example, Taylor (1973, Sect. 9.3).

Exercise 7.3 On the kth toss of a fair coin, a gambler receives 0 if it is a tail and $2/3^k$ if it is a head. Let X be the total gain of the gambler after an infinite sequence of tosses of the coin. Show that X has the Cantor distribution.

Exercise 7.4 Show that the Cantor set is uncountable.

7.2 Moments

The main purpose of this chapter is to study the 'moments' of a random variable: what are moments, and how can we use them? For any random variable X, the kth *moment* of X is defined for $k = 1, 2, \ldots$ to be the number $\mathbb{E}(X^k)$, that is, the expectation of the kth power of X, whenever this expectation exists. We shall see that the sequence $\mathbb{E}(X), \mathbb{E}(X^2), \ldots$ contains a lot of information about X, but first we give some examples of calculations of moments.[2]

Example 7.5 If X has the *exponential distribution* with parameter λ, then

$$\begin{aligned}\mathbb{E}(X^k) &= \int_0^\infty x^k \lambda e^{-\lambda x}\, dx \qquad \text{by Theorem 5.58} \\ &= \left[-x^k e^{-\lambda x}\right]_0^\infty + \int_0^\infty k x^{k-1} e^{-\lambda x}\, dx \\ &= \frac{k}{\lambda}\mathbb{E}(X^{k-1})\end{aligned}$$

if $k \geq 1$, giving that

$$\begin{aligned}\mathbb{E}(X^k) &= \frac{k}{\lambda}\mathbb{E}(X^{k-1}) = \frac{k(k-1)}{\lambda^2}\mathbb{E}(X^{k-2}) = \cdots \\ &= \frac{k!}{\lambda^k}\mathbb{E}(X^0) = \frac{k!}{\lambda^k}\mathbb{E}(1) = \frac{k!}{\lambda^k}.\end{aligned}$$

In particular, the exponential distribution has moments of all orders. △

Example 7.6 If X has the *Cauchy distribution*, then

$$\mathbb{E}(X^k) = \int_{-\infty}^\infty \frac{x^k}{\pi(1+x^2)}\, dx$$

for values of k for which this integral converges absolutely. It is however an elementary exercise (remember Example 5.66) to see that

$$\int_{-\infty}^\infty \left|\frac{x^k}{\pi(1+x^2)}\right| dx = \infty$$

if $k \geq 1$, and so the Cauchy distribution possesses *no* moments.

[2] Strictly speaking, these moments are associated with the *distribution* of X rather than with the random variable X itself. Thus we shall speak of the *moments* of a distribution or of a density function.

You may see how to adapt this example to find a density function with some moments but not all. Consider the density function

$$f(x) = \frac{c}{1+|x|^m} \qquad \text{for } x \in \mathbb{R},$$

where m (≥ 2) is an integer, and c is chosen so that f is indeed a density function:

$$c = \left(\int_{-\infty}^{\infty} \frac{dx}{1+|x|^m} \right)^{-1}.$$

You may check that this density function has a kth moment for those values of k satisfying $1 \leq k \leq m-2$ only. △

Given the distribution function F_X of the random variable X, we may calculate its moments whenever they exist (at least, if X is discrete or continuous). It is interesting to ask whether or not the converse is true: given the sequence $\mathbb{E}(X), \mathbb{E}(X^2), \ldots$ of (finite) moments of X, is it possible to reconstruct the distribution of X? The general answer to this question is *no*, but is *yes* if we have some extra information about the moment sequence.

Theorem 7.7 (Uniqueness theorem for moments) *Suppose that all moments $\mathbb{E}(X)$, $\mathbb{E}(X^2), \ldots$ of the random variable X exist, and that the series*

$$\sum_{k=0}^{\infty} \frac{1}{k!} t^k \mathbb{E}(X^k) \tag{7.8}$$

is absolutely convergent for some $t > 0$. Then the sequence of moments uniquely determines the distribution of X.

Thus the absolute convergence of (7.8) for some $t > 0$ is sufficient (but not necessary) for the moments to determine the underlying distribution. We omit the proof of this, since it is not primarily a theorem about probability theory; a proof may be found in textbooks on real and complex analysis. The theorem is closely related to the uniqueness theorem, Theorem 4.13, for probability generating functions; the series in (7.8) is the exponential generating function of the sequences of moments.

Here is an example of a distribution which is not determined uniquely by its moments.

Example 7.9 (Log-normal distribution) If X has the normal distribution with mean 0 and variance 1, then $Y = e^X$ has the *log-normal distribution* with density function

$$f(x) = \begin{cases} \dfrac{1}{x\sqrt{2\pi}} \exp\left[-\frac{1}{2}(\log x)^2\right] & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases}$$

Suppose that $-1 \leq a \leq 1$ and define

$$f_a(x) = \left[1 + a \sin(2\pi \log x)\right] f(x).$$

It is not difficult to check that

(a) f_a is a density function,
(b) f has finite moments of all orders,
(c) f_a and f have equal moments of all orders, in that

$$\int_{-\infty}^{\infty} x^k f(x)\,dx = \int_{-\infty}^{\infty} x^k f_a(x)\,dx \qquad \text{for } k = 1, 2, \ldots.$$

Thus, $\{f_a : -1 \le a \le 1\}$ is a collection of distinct density functions having the same moments. △

Exercise 7.10 If X is uniformly distributed on (a, b), show that

$$\mathbb{E}(X^k) = \frac{b^{k+1} - a^{k+1}}{(b-a)(k+1)} \qquad \text{for } k = 1, 2, \ldots.$$

Exercise 7.11 If X has the gamma distribution with parameters w and λ, show that

$$\mathbb{E}(X^k) = \frac{\Gamma(w+k)}{\lambda^k \Gamma(w)} \qquad \text{for } k = 1, 2, \ldots.$$

Exercise 7.12 If X has the χ^2 distribution with n degrees of freedom, show that

$$\mathbb{E}(X^k) = 2^k \frac{\Gamma(k + \frac{1}{2}n)}{\Gamma(\frac{1}{2}n)} \qquad \text{for } k = 1, 2, \ldots.$$

7.3 Variance and covariance

We recall that the *variance* of a random variable X is defined to be

$$\operatorname{var}(X) = \mathbb{E}\big([X - \mu]^2\big), \tag{7.13}$$

where $\mu = \mathbb{E}(X)$ (see (2.33) and (5.61) for discrete and continuous random variables). The variance of X is a measure of its dispersion about its expectation μ, in the sense that if X often takes values which differ considerably from μ, then $|X - \mu|$ is often large and so $\mathbb{E}([X-\mu]^2)$ will be large, whereas if X is usually near to μ, then $|X - \mu|$ is usually small and $\mathbb{E}([X-\mu]^2)$ is small also. An extreme case arises when X is *concentrated* at some point. It is the case that, for a random variable Y,

$$\mathbb{E}(Y^2) = 0 \qquad \text{if and only if} \qquad \mathbb{P}(Y = 0) = 1. \tag{7.14}$$

Obviously, $\mathbb{E}(Y^2) = 0$ if $\mathbb{P}(Y = 0) = 1$, and the converse holds since (for discrete random variables, anyway)

$$\mathbb{E}(Y^2) = \sum_y y^2 \mathbb{P}(Y = y) \ge 0$$

with equality if and only if $\mathbb{P}(Y = y) = 0$ for all $y \ne 0$. Applying (7.14) to $Y = X - \mu$ gives

$$\operatorname{var}(X) = 0 \qquad \text{if and only if} \qquad \mathbb{P}(X = \mu) = 1, \tag{7.15}$$

so that 'zero variance' means 'no dispersion at all'.

There are many other possible measures of dispersion, such as $\mathbb{E}(|X-\mu|)$ and $\mathbb{E}(|X-\mu|^3)$ and so on, but it is easiest to work with variances.

As noted before, when calculating the variance of X, it is often simpler to work with the moments of X rather than with (7.13) directly. That is to say, it may be easier to make use of the formula

$$\begin{aligned}\operatorname{var}(X) &= \mathbb{E}\big([X-\mu]^2\big) \\ &= \mathbb{E}(X^2) - \mu^2 \end{aligned} \tag{7.16}$$

by (2.35) and (5.62), where $\mu = \mathbb{E}(X)$.

There is also a simple formula for calculating the variance of a linear function $aX+b$ of a random variable X, namely

$$\operatorname{var}(aX+b) = a^2 \operatorname{var}(X). \tag{7.17}$$

To see this, note by (6.63) that

$$\begin{aligned}\operatorname{var}(aX+b) &= \mathbb{E}\big(\big[aX+b-\mathbb{E}(aX+b)\big]^2\big) \\ &= \mathbb{E}\big(\big[aX+b-a\mathbb{E}(X)-b\big]^2\big) \\ &= \mathbb{E}\big(a^2[X-\mu]^2\big) = a^2\mathbb{E}\big([X-\mu]^2\big) \\ &= a^2 \operatorname{var}(X).\end{aligned}$$

As a measure of dispersion, the variance of X has an undesirable property: it is non-linear in the sense that the variance of aX is a^2 times the variance of X. For this reason, statisticians often prefer to work with the *standard deviation* of X, defined to be $\sqrt{\operatorname{var}(X)}$.

What can be said about $\operatorname{var}(X+Y)$ in terms of $\operatorname{var}(X)$ and $\operatorname{var}(Y)$? It is simple to see that

$$\begin{aligned}\operatorname{var}(X+Y) &= \mathbb{E}\big(\big\{(X+Y)-\mathbb{E}(X+Y)\big\}^2\big) \\ &= \mathbb{E}\big(\big\{[X-\mathbb{E}(X)]+[Y-\mathbb{E}(Y)]\big\}^2\big) \\ &= \mathbb{E}\big([X-\mathbb{E}(X)]^2 + 2[X-\mathbb{E}(X)][Y-\mathbb{E}(Y)] + [Y-\mathbb{E}(Y)]^2\big) \\ &= \operatorname{var}(X) + 2\mathbb{E}\big([X-\mathbb{E}(X)][Y-\mathbb{E}(Y)]\big) + \operatorname{var}(Y).\end{aligned} \tag{7.18}$$

It is convenient to have a special word for the middle term in the last expression, and to this end we define the 'covariance' of the pair X, Y.

Definition 7.19 *The* ***covariance*** *of the random variables X and Y is the quantity denoted* $\operatorname{cov}(X,Y)$ *and given by*

$$\operatorname{cov}(X,Y) = \mathbb{E}\big([X-\mathbb{E}(X)][Y-\mathbb{E}(Y)]\big), \tag{7.20}$$

whenever these expectations exist.

Note that $\text{cov}(X, Y)$ may be written in a simpler form: expand $[X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]$ in (7.20) and use the linearity of $\mathbb{E}$ to find that

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{7.21}$$

Equation (7.18) may be rewritten as

$$\text{var}(X + Y) = \text{var}(X) + 2\text{cov}(X, Y) + \text{var}(Y), \tag{7.22}$$

valid for all random variables X and Y. If X and Y are independent, then

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0 \tag{7.23}$$

by (6.64), giving that the sum of independent random variables has variance

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) \tag{7.24}$$

whenever the latter variances exist.

The converse of the last remark is false in general: we recall from (3.22) that there exist dependent random variables X and Y for which $\text{cov}(X, Y) = 0$. Despite this, $\text{cov}(X, Y)$ is often used as a measure of the dependence of X and Y, and the reason for this is that $\text{cov}(X, Y)$ is a single number (rather than a complicated object such as a joint density function) which contains some useful information about the *joint* behaviour of X and Y. For example, if $\text{cov}(X, Y) > 0$, then $X - \mathbb{E}(X)$ and $Y - \mathbb{E}(Y)$ may have a good chance (in some sense) of having the same sign. A principal disadvantage of covariance as a measure of dependence is that it is not 'scale-invariant': if X and Y are random measurements (in inches, say) and U and V are the same random measurements in centimetres (so that $U = \alpha X$ and $V = \alpha Y$, where $\alpha \approx 2.54$), then $\text{cov}(U, V) \approx 6\text{cov}(X, Y)$, despite the fact that the two pairs, (X, Y) and (U, V), measure essentially the same quantities. To deal with this, we 're-scale' covariance as follows.

Definition 7.25 *The **correlation (coefficient)** of the random variables X and Y is the quantity $\rho(X, Y)$ given by*

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}}, \tag{7.26}$$

whenever the latter quantities exist and $\text{var}(X)\,\text{var}(Y) \neq 0$.

It is a simple exercise to show that

$$\rho(aX + b, cY + d) = \rho(X, Y) \tag{7.27}$$

for all $a, b, c, d \in \mathbb{R}$ such that $ac \neq 0$, and so correlation is scale invariant. Correlation has another attractive property as a measure of dependence. It turns out that $-1 \leq \rho(X, Y) \leq 1$ always, and moreover there are specific interpretations in terms of the joint behaviour of X and Y of the cases when $\rho(X, Y) = \pm 1$.

Theorem 7.28 *If X and Y are random variables, then*

$$-1 \leq \rho(X, Y) \leq 1, \tag{7.29}$$

whenever this correlation exists.

The proof of this is a direct application of the next inequality.

Theorem 7.30 (Cauchy–Schwarz inequality) *If U and V are random variables, then*

$$[\mathbb{E}(UV)]^2 \leq \mathbb{E}(U^2)\mathbb{E}(V^2), \tag{7.31}$$

whenever these expectations exist.

Proof Let $s \in \mathbb{R}$ and define a new random variable $W = sU + V$. Clearly, $W^2 \geq 0$ always, and so

$$\begin{aligned} 0 \leq \mathbb{E}(W^2) &= \mathbb{E}(s^2U^2 + 2sUV + V^2) \\ &= as^2 + bs + c, \end{aligned} \tag{7.32}$$

where $a = \mathbb{E}(U^2)$, $b = 2\mathbb{E}(UV)$, $c = \mathbb{E}(V^2)$. Clearly, $a \geq 0$, and we may suppose that $a > 0$, since otherwise $\mathbb{P}(U = 0) = 1$ by (7.14) and the result holds trivially. Equation (7.32) implies that the quadratic function $g(s) = as^2 + bs + c$ intersects the line $t = 0$ (in the usual (s, t) plane) at most once (since if $g(s) = 0$ for distinct values $s = s_1$ and $s = s_2$, then $g(s) < 0$ for all values of s strictly between s_1 and s_2). Thus, the quadratic equation '$g(s) = 0$' has at most one real root, giving that its discriminant $b^2 - 4ac$ satisfies $b^2 - 4ac \leq 0$. Hence

$$[2\mathbb{E}(UV)]^2 - 4\mathbb{E}(U^2)\mathbb{E}(V^2) \leq 0$$

and the result is proved. □

Proof of Theorem 7.28 Set $U = X - \mathbb{E}(X)$ and $V = Y - \mathbb{E}(Y)$ in the Cauchy–Schwarz inequality to find that

$$\operatorname{cov}(X, Y)^2 \leq \operatorname{var}(X) \operatorname{var}(Y),$$

yielding (7.29) immediately. □

Only under very special circumstances can it be the case that $\rho(X, Y) = \pm 1$, and these circumstances are explored by considering the proof of (7.31) more carefully. Let $a = \operatorname{var}(X)$, $b = 2\operatorname{cov}(X, Y)$, $c = \operatorname{var}(Y)$ and suppose that $\rho(X, Y) = \pm 1$. Then $\operatorname{var}(X) \operatorname{var}(Y) \neq 0$ and

$$b^2 - 4ac = 4 \operatorname{var}(X) \operatorname{var}(Y)\big[\rho(X, Y)^2 - 1\big] = 0,$$

and so the quadratic equation

$$as^2 + bs + c = 0$$

has two equal real roots, at $s = \alpha$, say. Therefore, $W = \alpha[X - \mathbb{E}(X)] + [Y - \mathbb{E}(Y)]$ satisfies

$$\mathbb{E}(W^2) = a\alpha^2 + b\alpha + c = 0,$$

giving that $\mathbb{P}(W = 0) = 1$, by (7.14), and showing that (essentially) $Y = -\alpha X + \beta$, where $\beta = \alpha\mathbb{E}(X) + \mathbb{E}(Y)$. A slightly more careful treatment discriminates between the values $+1$ and -1 for $\rho(X, Y)$:

$$\rho(X, Y) = 1 \quad \text{if and only if } \mathbb{P}(Y = \alpha X + \beta) = 1 \text{ for some real } \alpha \text{ and } \beta \text{ with } \alpha > 0, \tag{7.33}$$

$$\rho(X, Y) = -1 \quad \text{if and only if } \mathbb{P}(Y = \alpha X + \beta) = 1 \text{ for some real } \alpha \text{ and } \beta \text{ with } \alpha < 0. \tag{7.34}$$

To recap, we may use $\rho(X, Y)$ as a measure of the dependence of X and Y. If X and Y have non-zero variances, then $\rho(X, Y)$ takes some value in the interval $[-1, 1]$, and this value should be interpreted in the light of the ways in which the values $-1, 0, 1$ may arise:

(a) if X and Y are independent, then $\rho(X, Y) = 0$,
(b) Y is a linear *increasing* function of X if and only if $\rho(X, Y) = 1$,
(c) Y is a linear *decreasing* function of X if and only if $\rho(X, Y) = -1$.

If $\rho(X, Y) = 0$, we say that X and Y are *uncorrelated*.

Exercise 7.35 If X and Y have the bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$ (see (6.76)), show that

$$\operatorname{cov}(X, Y) = \rho\sigma_1\sigma_2 \qquad \text{and} \qquad \rho(X, Y) = \rho.$$

Exercise 7.36 Let $X_1, X_2, \ldots$ be a sequence of uncorrelated random variables, each having variance σ^2. If $S_n = X_1 + X_2 + \cdots + X_n$, show that

$$\operatorname{cov}(S_m, S_n) = \operatorname{var}(S_m) = m\sigma^2 \qquad \text{if } m < n.$$

Exercise 7.37 Show that $\operatorname{cov}(X, Y) = 1$ in the case when X and Y have joint density function

$$f(x, y) = \begin{cases} \dfrac{1}{y} e^{-y - x/y} & \text{if } x, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

7.4 Moment generating functions

If X is a discrete random variable taking values in $\{0, 1, 2, \ldots\}$, its probability generating function is defined by

$$G_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} s^k \mathbb{P}(X = k). \tag{7.38}$$

Probability generating functions are very useful, but only when the random variables in question take non-negative integral values. For more general random variables, it is customary to consider a modification of (7.38).

Definition 7.39 *The **moment generating function** (or **mgf**) of the random variable X is the function M_X defined by*

$$M_X(t) = \mathbb{E}(e^{tX}), \tag{7.40}$$

for all $t \in \mathbb{R}$ for which this expectation exists.

This is a modification of (7.38) in the sense that, if X takes values in $\{0, 1, 2, \dots\}$, then

$$M_X(t) = \mathbb{E}(e^{tX}) = G_X(e^t), \tag{7.41}$$

by the substitution $s = e^t$. In general,

$$M_X(t) = \mathbb{E}(e^{tX}) = \begin{cases} \displaystyle\sum_x e^{tx}\mathbb{P}(X = x) & \text{if } X \text{ is discrete,} \\ \displaystyle\int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx & \text{if } X \text{ is continuous,} \end{cases} \tag{7.42}$$

whenever this sum or integral converges absolutely. In some cases, the existence of $M_X(t)$ can pose a problem for non-zero values of t.

Example 7.43 If X has the *normal distribution* with mean 0 and variance 1, then

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,dx \\ &= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2}\,dx \\ &= e^{\frac{1}{2}t^2}, \end{aligned} \tag{7.44}$$

since the integrand in the latter integral is the density function of the normal distribution with mean t and variance 1, and thus has integral 1. The moment generating function $M_X(t)$ exists for all $t \in \mathbb{R}$. △

Example 7.45 If X has the *exponential distribution* with parameter λ, then

$$M_X(t) = \int_0^{\infty} e^{tx}\lambda e^{-\lambda x}\,dx = \begin{cases} \dfrac{\lambda}{\lambda - t} & \text{if } t < \lambda, \\ \infty & \text{if } t \geq \lambda, \end{cases} \tag{7.46}$$

so that $M_X(t)$ exists only for values of t satisfying $t < \lambda$. △

Example 7.47 If X has the *Cauchy distribution*, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\pi(1 + x^2)}\,dx = \begin{cases} 1 & \text{if } t = 0, \\ \infty & \text{if } t \neq 0, \end{cases}$$

so that $M_X(t)$ exists only at $t = 0$. △

This difficulty over the existence of $\mathbb{E}(e^{tX})$ may be avoided by studying the complex-valued *characteristic function* $\phi_X(t) = \mathbb{E}(e^{itX})$ of X instead—this function can be shown to exist for all $t \in \mathbb{R}$. However, we want to avoid $i = \sqrt{-1}$ at this stage, and so we must accustom ourselves to the difficulty, although we shall return to characteristic functions in Section 7.6. It turns out to be important only that $\mathbb{E}(e^{tX})$ exists in some neighbourhood $(-\delta, \delta)$ of the origin, and the reason for this is contained in the uniqueness theorem for moment generating functions (see the forthcoming Theorem 7.55). We shall generally use moment generating functions freely, but always subject to the implicit assumption of existence in a neighbourhood of the origin.

The reason for the name 'moment generating function' is the following intuitively attractive expansion:

$$\begin{aligned} M_X(t) = \mathbb{E}(e^{tX}) &= \mathbb{E}\left(1 + tX + \frac{1}{2!}(tX)^2 + \cdots\right) \\ &= 1 + t\mathbb{E}(X) + \frac{1}{2!}t^2\mathbb{E}(X^2) + \cdots . \end{aligned} \tag{7.48}$$

That is to say, subject to a rigorous derivation of (7.48) which does not interchange the two operations $\mathbb{E}$ and $\sum$ so light-heartedly, $M_X(t)$ is the exponential generating function of the moments of X.

Theorem 7.49 *If $M_X(t)$ exists in a neighbourhood of 0, then, for $k = 1, 2, \ldots$,*

$$\mathbb{E}(X^k) = M_X^{(k)}(0), \tag{7.50}$$

the kth derivative of $M_X(t)$ evaluated at $t = 0$.

Sketch proof Cross your fingers for the sake of rigour to obtain

$$\begin{aligned} \frac{d^k}{dt^k} M_X(t) &= \frac{d^k}{dt^k}\mathbb{E}(e^{tX}) \\ &= \mathbb{E}\left(\frac{d^k}{dt^k} e^{tX}\right) = \mathbb{E}(X^k e^{tX}), \end{aligned}$$

and finish by setting $t = 0$. It is the interchange of the expectation operator and the differential operator which requires justification here. □

As noted before, much of probability theory is concerned with sums of random variables. It can be difficult in practice to calculate the distribution of a sum from knowledge of the distributions of the summands, and it is here that moment generating functions are extremely useful. Consider first the linear function $aX + b$ of the random variable X. If $a, b \in \mathbb{R}$,

$$\begin{aligned} M_{aX+b}(t) &= \mathbb{E}(e^{t(aX+b)}) = \mathbb{E}(e^{atX}e^{tb}) \\ &= e^{tb}\mathbb{E}(e^{(at)X}) \qquad \text{by (6.63)} \end{aligned}$$

giving that

$$M_{aX+b}(t) = e^{tb} M_X(at). \tag{7.51}$$

A similar argument enables us to find the moment generating function of the sum of independent random variables.

Theorem 7.52 *If X and Y are independent random variables, then $X + Y$ has moment generating function*

$$M_{X+Y}(t) = M_X(t) M_Y(t). \tag{7.53}$$

Proof We have that

$$\begin{aligned} M_{X+Y}(t) &= \mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX} e^{tY}) \\ &= \mathbb{E}(e^{tX}) \mathbb{E}(e^{tY}), \end{aligned}$$

by independence and Theorem 6.66. □

By Theorem 7.52, the sum $S = X_1 + \cdots + X_n$ of n independent random variables has moment generating function

$$M_S(t) = M_{X_1}(t) \cdots M_{X_n}(t). \tag{7.54}$$

Finally, we state the uniqueness theorem for moment generating functions.

Theorem 7.55 (Uniqueness theorem for moment generating functions) *If the moment generating function M_X satisfies $M_X(t) = \mathbb{E}(e^{tX}) < \infty$ for all t satisfying $-\delta < t < \delta$ and some $\delta > 0$, there is a unique distribution with moment generating function M_X. Furthermore, under this condition, we have that $\mathbb{E}(X^k) < \infty$ for $k = 1, 2, \ldots$ and*

$$M_X(t) = \sum_{k=0}^{\infty} \frac{1}{k!} t^k \mathbb{E}(X^k) \qquad \text{for } |t| < \delta. \tag{7.56}$$

We do not prove this here. This theorem is basically the Laplace inverse theorem since, by (7.42), $M_X(t)$ is essentially the Laplace transform of the density function $f_X(x)$. The Laplace inverse theorem says that if the Laplace transform of f_X exists in a suitable manner, then f_X may be found from this transform by using the inversion formula. Equation (7.56) is the same as (7.48), but some care is needed to justify the interchange of $\mathbb{E}$ and $\sum$ noted after (7.48). Clearly, there is a close relationship between Theorems 7.55 and 7.7, but we do not explore this here.

Finally, we give an example of the use of moment generating functions in which the uniqueness part of Theorem 7.55 is essential.

Example 7.57 Let X and Y be independent random variables, X having the normal distribution with parameters μ_1 and σ_1^2 and Y having the normal distribution with parameters μ_2 and σ_2^2. Show that their sum $Z = X + Y$ has the normal distribution with parameters $\mu_1 + \mu_2$ and $\sigma_1^2 + \sigma_2^2$.

Solution Let U be a random variable having the normal distribution with parameters μ and σ^2. The moment generating function of U is

$$\begin{aligned} M_U(t) &= \int_{-\infty}^{\infty} e^{tu} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(u-\mu)^2\right) du \\ &= e^{\mu t} \int_{-\infty}^{\infty} e^{x\sigma t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx \qquad \text{on substituting } x = \frac{u-\mu}{\sigma} \\ &= \exp\bigl(\mu t + \tfrac{1}{2}\sigma^2 t^2\bigr) \qquad \text{by (7.44).} \end{aligned} \tag{7.58}$$

By Theorem 7.52,

$$\begin{aligned} M_Z(t) &= M_X(t) M_Y(t) \\ &= \exp\bigl(\mu_1 t + \tfrac{1}{2}\sigma_1^2 t^2\bigr) \exp\bigl(\mu_2 t + \tfrac{1}{2}\sigma_2^2 t^2\bigr) \qquad \text{by (7.58)} \\ &= \exp\Bigl[(\mu_1+\mu_2)t + \tfrac{1}{2}(\sigma_1^2+\sigma_2^2)t^2\Bigr], \end{aligned}$$

which we recognize by (7.58) as the moment generating function of the normal distribution with parameters $\mu_1 + \mu_2$ and $\sigma_1^2 + \sigma_2^2$. We deduce that Z has this distribution by appealing to Theorem 7.55. △

Exercise 7.59 Find the moment generating function of a random variable having

(a) the gamma distribution with parameters w and λ,

(b) the Poisson distribution with parameter λ.

Exercise 7.60 If X has the normal distribution with mean μ and variance σ^2, find $\mathbb{E}(X^3)$.

Exercise 7.61 Show that, if X has a normal distribution, then so does $aX + b$, for any $a, b \in \mathbb{R}$ with $a \neq 0$. You may use Theorem 7.55 together with (7.51) and (7.58).

Exercise 7.62 Let $X_1, X_2, \ldots$ be identically distributed random variables with common moment generating function M. Let N be a random variable taking non-negative integer values with probability generating function G, and suppose N is independent of the sequence (X_i). Show that the random sum $S = X_1 + X_2 + \cdots + X_N$ has moment generating function $M_S(t) = G(M(t))$.

7.5 Two inequalities

The purpose of this section is to state and prove two very useful inequalities, attributed to Markov and Jensen, and involving the moments of a random variable.

Markov's inequality is concerned with the following question: if you know that the mean $\mathbb{E}(X)$ of a non-negative random variable X is finite, what does this tell you about the distribution of X? The so-called right and left tails of X are the probabilities that X is large and positive (respectively, large and negative). More specifically, the (*right*) *tail* is the function $\mathbb{P}(X \geq t)$ for large t, with a corresponding definition for the left tail. If X is positive and $\mathbb{E}(X) < \infty$, then the right tail of X cannot be too 'fat'.

Theorem 7.63 (Markov's inequality) *For any non-negative random variable X,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t} \qquad \text{for } t > 0. \tag{7.64}$$

Proof This is often proved in elementary texts by expressing $\mathbb{E}(X)$ as either a sum or an integral, as appropriate. It is simpler to argue as follows. Let X be a non-negative random variable, and $t > 0$. Recalling that X is a function from Ω to $[0, \infty)$, we have the trivial inequality

$$X(\omega) \geq \begin{cases} t & \text{if } X(\omega) \geq t, \\ 0 & \text{if } X(\omega) < t, \end{cases}$$

for $\omega \in \Omega$. We write this as an inequality between the function X and the indicator function of the event $A = \{X \geq t\}$:

$$X \geq t 1_A.$$

Now take expectations, and remember that $\mathbb{E}(1_A) = \mathbb{P}(A)$. The result is the required inequality $\mathbb{E}(X) \geq t\mathbb{P}(X \geq t)$. □

Example 7.65 Let X be a random variable. The real number m is called a *median* of X if

$$\mathbb{P}(X < m) \leq \tfrac{1}{2} \leq \mathbb{P}(X \leq m).$$

Exercise 5.14 was to show that every random variable possesses at least one median m. If X is non-negative, then

$$\tfrac{1}{2} \leq \mathbb{P}(X \geq m) \leq \frac{\mathbb{E}(X)}{m},$$

by Markov's inequality. Therefore, any median m satisfies $m \leq 2\mathbb{E}(X)$. It is left to Exercise 7.72 to determine whether or not equality can hold. △

Jensen's inequality is of a different type, and concerns convexity. Let $-\infty \leq a < b \leq \infty$. A function $g : (a, b) \to \mathbb{R}$ is called *convex* if

$$g\big([1-t]u + tv\big) \geq (1-t)g(u) + tg(v) \tag{7.66}$$

for every $t \in [0, 1]$ and $u, v \in (a, b)$. Condition (7.66) may be expressed geometrically as follows. Let $u, v \in (a, b)$ and consider the straight line joining the two points $(u, g(u))$ and $(v, g(v))$ on the curve $y = g(x)$. Then (7.66) requires that this chord lies always above the curve itself (see Figure 7.2).

Theorem 7.67 (Jensen's inequality) *Let X be a random variable taking values in the (possibly infinite) interval (a, b) such that $\mathbb{E}(X)$ exists, and let $g : (a, b) \to \mathbb{R}$ be a convex function such that $\mathbb{E}|g(X)| < \infty$. Then*

$$\mathbb{E}(g(X)) \geq g(E(X)).$$

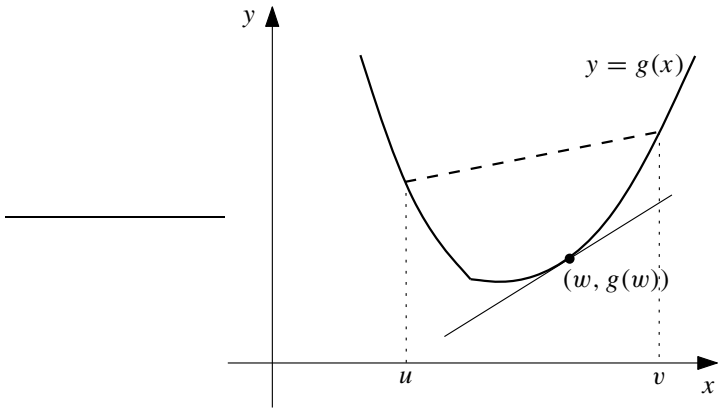


Fig. 7.2 A function g is convex if every chord of the curve $y = g(x)$ lies above the curve. At any point on the curve, there exists a tangent that 'supports' the curve. Note that a convex function is not necessarily everywhere differentiable.

This may be interpreted as follows. Inequality (7.66) requires that a weighted average of two values of g lies above the value of g at their average value. Taking expectations is itself a type of averaging operation, and Jensen's inequality extends the two-point average of (7.66) to this more general average.

In preparation for the proof, we present next a result known in a more general form as the 'supporting hyperplane theorem'. It is illustrated in Figure 7.2.

Theorem 7.68 (Supporting tangent theorem) *Let $g : (a, b) \to \mathbb{R}$ be convex, and let $w \in (u, v)$. There exists $\alpha \in \mathbb{R}$ such that*

$$g(x) \geq g(w) + \alpha(x - w) \qquad \text{for } x \in (a, b). \tag{7.69}$$

Proof Let $a < w < b$. The theorem says there exists a straight line that touches the curve at the point $(w, g(w))$ and such that the curve never passes below the line: the line 'supports' the curve. Some readers will be content with the 'proof by picture' of Figure 7.2. Others may prefer the following proof.

Let $u < w < v$. We may express w as a linear combination of u and v thus:

$$w = (1 - t)u + tv, \quad \text{where} \quad t = \frac{w - u}{v - u}.$$

By convexity,

$$g(w) \leq (1 - t)g(u) + tg(v),$$

which we reorganize as

$$\frac{g(w) - g(u)}{w - u} \leq \frac{g(v) - g(w)}{v - w}.$$

By maximizing the left side and minimizing the right side, we obtain that $L_w \leq R_w$, where

$$L_w = \sup\left\{\frac{g(w)-g(u)}{w-u} : u < w\right\}, \quad R_w = \inf\left\{\frac{g(v)-g(w)}{v-w} : v > w\right\}.$$

Pick $\alpha \in [L_w, R_w]$, so that

$$\frac{g(w)-g(u)}{w-u} \le \alpha \le \frac{g(v)-g(w)}{v-w}, \qquad \text{for } u < w < v.$$

On multiplying up the left inequality, we deduce (7.69) for $x = u < w$. Similarly, the right inequality yields (7.69) for $x = v > w$. □

Proof of Theorem 7.67 Let X take values in (a, b) with mean $\mu = \mathbb{E}(X)$. Let g be a convex function on this interval satisfying $\mathbb{E}|g(X)| < \infty$. By Theorem 7.68 with $w = \mu$, there exists $\alpha \in \mathbb{R}$ such that $g(x) \ge g(\mu) + \alpha(x - \mu)$. Therefore, $g(X) \ge g(\mu) + \alpha(X - \mu)$. Now take expectations to obtain $\mathbb{E}(g(X)) \ge g(\mu)$. □

Example 7.70 (Arithmetic/geometric mean inequality) The function $g(x) = -\log x$ is convex on the interval $(0, \infty)$. By Jensen's inequality applied to a positive random variable X with finite mean,

$$\mathbb{E}(\log X) \le \log \mathbb{E}(X). \tag{7.71}$$

Suppose X is a discrete random variable which is equally likely to take any of the positive values $x_1, x_2, \ldots, x_n$. Then

$$\mathbb{E}(\log X) = \frac{1}{n}\sum_{i=1}^{n} \log x_i = \log \gamma, \qquad \mathbb{E}(X) = \overline{x},$$

where

$$\gamma = \left(\prod_{i=1}^{n} x_i\right)^{1/n}, \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

are the geometric and arithmetic means of the x_i, respectively. By (7.71), $\log \gamma \le \log \overline{x}$, and we deduce that $\gamma \le \overline{x}$. In summary, the geometric mean of a set of positive numbers cannot exceed its arithmetic mean. This may be proved by a more direct method. △

Exercise 7.72 Determine which distributions on the non-negative reals, if any, with mean μ are such that 2μ is a median.

Exercise 7.73 Let I be an interval of the real line, and let $f : I \to \mathbb{R}$ be twice differentiable with $f''(x) > 0$ for $x \in I$. Show that f is convex on I.

Exercise 7.74 Show by Jensen's inequality that $\mathbb{E}(X^2) \ge \mathbb{E}(X)^2$.

Exercise 7.75 The harmonic mean η of the positive reals $x_1, x_2, \ldots, x_n$ is given by

$$\frac{1}{\eta} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}.$$

Show that η is no greater than the geometric mean of the x_i.

7.6 Characteristic functions

The Cauchy distribution is not the only distribution for which the moment generating function does not exist, and this problem of existence is a serious handicap to the use of moment generating functions. However, by a slight modification of the definition, we may obtain another type of generating function whose existence is guaranteed and which has broadly the same properties as before.[3]

Definition 7.76 *The **characteristic function** of the random variable X is defined to be the function ϕ_X given by*

$$\phi_X(t) = \mathbb{E}(e^{itX}) \qquad \text{for } t \in \mathbb{R}, \tag{7.77}$$

where $i = \sqrt{-1}$.

You may doubt the legitimacy of the expectation of the complex-valued random variable e^{itX}, but we recall that $e^{itX} = \cos tx + i \sin tx$ for $t, x \in \mathbb{R}$, so that (7.77) may be replaced by

$$\phi_X(t) = \mathbb{E}(\cos tX) + i\mathbb{E}(\sin tX)$$

if this is preferred.

Compare the characteristic function of X with its moment generating function $M_X(t) = \mathbb{E}(e^{tX})$. The finiteness of the latter is questionable since the exponential function is unbounded, so that e^{tX} may be very large indeed. On the other hand, e^{itX} lies on the unit circle in the complex plane, so that $|e^{itX}| = 1$ and giving that $|\phi_X(t)| \leq 1$ for all $t \in \mathbb{R}$.

Example 7.78 Suppose that the random variable X may take either the value a, with probability p, or the value b, with probability $1 - p$. Then

$$\phi_X(t) = \mathbb{E}(e^{itX}) = pe^{ita} + (1 - p)e^{itb}. \qquad \triangle$$

Example 7.79 If X has the *exponential distribution* with parameter λ, then

$$\phi_X(t) = \int_0^\infty e^{itx} \lambda e^{-\lambda x}\, dx = \frac{\lambda}{\lambda - it} \qquad \text{for } t \in \mathbb{R}.$$

This integral may be found either by splitting e^{itx} into real and imaginary parts or by using the calculus of residues. $\triangle$

Example 7.80 If X has the *Cauchy distribution*, then

$$\phi_X(t) = \int_{-\infty}^\infty e^{itx} \frac{1}{\pi(1 + x^2)}\, dx = e^{-|t|} \qquad \text{for } t \in \mathbb{R},$$

a result obtainable by the calculus of residues. $\triangle$

[3] Beginners to probability theory may wish to omit this section.

Some readers may prefer to avoid using the calculus of residues in calculating characteristic functions, arguing instead as in the following example. The moment generating function of a random variable X having the normal distribution with mean 0 and variance 1 is

$$M_X(t) = \mathbb{E}(e^{tX}) = e^{\frac{1}{2}t^2} \qquad \text{for } t \in \mathbb{R}.$$

It is therefore clear that the characteristic function of X must be

$$\phi_X(t) = \mathbb{E}(e^{itX}) = M_X(it) = e^{-\frac{1}{2}t^2} \qquad \text{for } t \in \mathbb{R}.$$

It is important to realize that this argument is not rigorous unless justified. It produces the correct answer for the normal and exponential distributions, as well as many others, but it will not succeed with the Cauchy distribution, since that distribution has no moment generating function in the first place. The argument may be shown to be valid whenever the moment generating function exists near the origin, the proof being an exercise in complex analysis. Thus, the following formal procedure is acceptable for calculating the characteristic function of a random variable X. If the moment generating function M_X is finite in a non-trivial neighbourhood of the origin, the characteristic function of X may be found by substituting $s = it$ in the formula for $M_X(s)$:

$$\phi_X(t) = M_X(it) \qquad \text{for } t \in \mathbb{R}. \tag{7.81}$$

Example 7.82 If X has the *normal distribution* with mean μ and variance σ^2, then the moment generating function of X is

$$M_X(s) = \exp\bigl(\mu s + \tfrac{1}{2}\sigma^2 s^2\bigr)$$

by (7.58), an expression valid for all $s \in \mathbb{R}$. We substitute $s = it$ here, to obtain

$$\phi_X(t) = \exp\bigl(i\mu t - \tfrac{1}{2}\sigma^2 t^2\bigr). \qquad \triangle$$

In broad terms, characteristic functions have the same useful properties as moment generating functions, and we finish this chapter with a brief account of these.

First, we consider the question of moments. Setting rigour to one side for the moment, the following expansion is interesting and informative:

$$\begin{aligned}\phi_X(t) = \mathbb{E}(e^{itX}) &= \mathbb{E}\left(1 + itX + \frac{1}{2!}(itX)^2 + \cdots\right) \\ &= 1 + it\mathbb{E}(X) + \frac{1}{2!}(it)^2\mathbb{E}(X^2) + \cdots,\end{aligned} \tag{7.83}$$

which is to say that ϕ_X is the exponential generating function of the sequence 1, $i\mathbb{E}(X)$, $i^2\mathbb{E}(X^2), \ldots$. There are technical difficulties in expressing this more rigorously, but we note that (7.83) is valid so long as $\mathbb{E}|X^k| < \infty$ for $k = 1, 2, \ldots$. Under this condition, it follows that the moments of X may be obtained in terms of the derivatives of ϕ_X:

$$i^k\mathbb{E}(X^k) = \phi_X^{(k)}(0), \tag{7.84}$$

the kth derivative of ϕ_X at 0.

If the moments of X are not all finite, then only a truncated form of the infinite series in (7.83) is valid.

Theorem 7.85 *If $\mathbb{E}|X^N| < \infty$ for some positive integer N, then*

$$\phi_X(t) = \sum_{k=0}^{N} \frac{1}{k!}(it)^k \mathbb{E}(X^k) + \mathrm{o}(t^N) \qquad \text{as } t \to 0. \tag{7.86}$$

We do not prove this here, but we remind the reader briefly about the meaning of the term $\mathrm{o}(t^N)$. The expression $\mathrm{o}(h)$ denotes some function of h which is of a smaller order of magnitude than h as $h \to 0$. More precisely, we write $f(h) = \mathrm{o}(h)$ if $f(h)/h \to 0$ as $h \to 0$. The term $\mathrm{o}(h)$ generally represents a different function of h at each appearance. Thus, for example, $\mathrm{o}(h) + \mathrm{o}(h) = \mathrm{o}(h)$.[4] The conclusion of Theorem 7.85 is that the remainder in (7.86) is negligible compared with the terms involving $1, t, t^2, \ldots, t^N$, when t is small. For a proof of Theorem 7.85, see Feller (1971, p. 487) or Chung (2001, p. 168).

When adding together independent random variables, characteristic functions are just as useful as moment generating functions.

Theorem 7.87 *Let X and Y be independent random variables with characteristic functions ϕ_X and ϕ_Y, respectively.*

(a) *If $a, b \in \mathbb{R}$ and $Z = aX + b$, then $\phi_Z(t) = e^{itb}\phi_X(at)$.*
(b) *The characteristic function of $X + Y$ is $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.*

Proof (a) We have that

$$\begin{aligned}\phi_Z(t) &= \mathbb{E}(e^{it(aX+b)}) = \mathbb{E}(e^{itb}e^{it(at)X}) \\ &= e^{itb}\phi_X(at).\end{aligned}$$

If you are in doubt about treating these complex-valued quantities as if they were real, simply expand the complex exponential function in terms of the cosine and sine functions, and collect the terms back together at the end.

(b) Similarly,

$$\begin{aligned}\phi_{X+Y} &= \mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itX}e^{itY}) \\ &= \mathbb{E}(e^{itX})\mathbb{E}(e^{itY}) \qquad \text{by independence.}\end{aligned}$$

□

Finally, we discuss the uniqueness of characteristic functions.

Theorem 7.88 (Uniqueness theorem for characteristic functions) *Let X and Y have characteristic functions ϕ_X and ϕ_Y, respectively. Then X and Y have the same distributions if and only if $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$.*

[4] This notation is sometimes termed Landau's notation.

That is to say, any given characteristic function ϕ corresponds to a unique distribution function. However, it is not always a simple matter to find this distribution function in terms of ϕ. There is a general 'inversion formula', but this is rather complicated and is omitted (see Grimmett and Stirzaker (2001, p. 189)). For distributions with density functions, the inversion formula takes on a relatively simple form.

Theorem 7.89 (Inversion theorem) *Let X have characteristic function ϕ and density function f. Then*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t)\, dt \tag{7.90}$$

at every point x at which f is differentiable.

This formula is often useful, but there is an obstacle in the way of its application. If we are given a characteristic function ϕ, we may only apply formula (7.90) once we know that ϕ comes from a continuous random variable, but how may we check that this is the case? There is no attractive necessary and sufficient condition on ϕ for this to hold, but a sufficient condition is that

$$\int_{-\infty}^{\infty} |\phi(t)|\, dt < \infty. \tag{7.91}$$

This condition is only of limited value: although it holds for the characteristic function of the normal distribution (7.82) for example, it fails for that of the exponential distribution (7.79).

Example 7.92 Those in the know will have spotted that characteristic functions are simply Fourier transforms in disguise, and that Theorem 7.89 is a version of the Fourier inversion theorem. The relationship between characteristic functions and Fourier analysis may easily be made more concrete in the case of integer-valued random variables. Suppose that X is a random variable taking values in the set $\{0, 1, 2, \dots\}$ of non-negative integers, with probability mass function $p_j = \mathbb{P}(X = j)$ for $j = 0, 1, 2, \dots$. The characteristic function of X is

$$\phi(t) = \sum_{k=0}^{\infty} p_k e^{itk}. \tag{7.93}$$

Suppose now that we know ϕ, but we wish to recover the probabilities p_j. We multiply through (7.93) by e^{-itj} to obtain

$$e^{-itj}\phi(t) = \sum_{k=0}^{\infty} p_k e^{it(k-j)}.$$

Next, we integrate with respect to t over the interval $[0, 2\pi]$, remembering that for integers m

$$\int_0^{2\pi} e^{imt}\, dt = \begin{cases} 2\pi & \text{if } m = 0, \\ 0 & \text{if } m \neq 0, \end{cases}$$

thereby obtaining

$$\int_0^{2\pi} e^{-itj}\phi(t)\,dt = 2\pi p_j.$$

Therefore,

$$p_j = \frac{1}{2\pi}\int_0^{2\pi} e^{-itj}\phi(t)\,dt \qquad \text{for } j = 0, 1, 2, \ldots. \tag{7.94}$$

We are merely calculating the Fourier series for ϕ. Notice the close resemblance between (7.94) and the inversion formula (7.90) for density functions. △

Exercise 7.95 Show that the characteristic function of a random variable having the binomial distribution with parameters n and p is

$$\phi(t) = (q + pe^{it})^n,$$

where $q = 1 - p$.

Exercise 7.96 Let X be uniformly distributed on (a, b). Show that

$$\phi_X(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}.$$

If X is uniformly distributed on $(-b, b)$, show that

$$\phi_X(t) = \frac{1}{bt}\sin bt.$$

Exercise 7.97 Find the characteristic function of a random variable having

(a) the gamma distribution with parameters w and λ,

(b) the Poisson distribution with parameter λ.

Exercise 7.98 If X and Y are independent and identically distributed random variables, show that

$$\phi_{X-Y}(t) = |\phi_X(t)|^2.$$

7.7 Problems

1. Let X and Y be random variables with equal variance. Show that $U = X - Y$ and $V = X + Y$ are uncorrelated. Give an example to show that U and V need not be independent even if, further, X and Y are independent.

2. Let $X_1, X_2, \ldots$ be uncorrelated random variables, each having mean μ and variance σ^2. If $\overline{X} = n^{-1}(X_1 + X_2 + \cdots + X_n)$, show that

$$\mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \sigma^2.$$

This fact is of importance in statistics and is used when estimating the population variance from knowledge of a random sample.

3. Let $X_1, X_2, \ldots$ be identically distributed, independent random variables and let $S_n = X_1 + X_2 + \cdots + X_n$. Show that

$$\mathbb{E}\left(\frac{S_m}{S_n}\right) = \frac{m}{n} \qquad \text{for } m \leq n,$$

provided that all the necessary expectations exist. Is the same true if $m > n$?

4. Show that every distribution function has only a countable set of points of discontinuity.

5. Let X and Y be independent random variables, X having the gamma distribution with parameters s and λ, and Y having the gamma distribution with parameters t and λ. Use moment generating functions to show that $X + Y$ has the gamma distribution with parameters $s + t$ and λ.

6. Let $X_1, X_2, \ldots, X_n$ be independent random variables with the exponential distribution, parameter λ. Show that $X_1 + X_2 + \cdots + X_n$ has the gamma distribution with parameters n and λ.

7. Show from the result of Problem 7.7.5 that the χ^2 distribution with n degrees of freedom has moment generating function

$$M(t) = (1 - 2t)^{-\frac{1}{2}n} \qquad \text{if } t < \tfrac{1}{2}.$$

Deduce that, if $X_1, X_2, \ldots, X_n$ are independent random variables having the normal distribution with mean 0 and variance 1, then

$$Z = X_1^2 + X_2^2 + \cdots + X_n^2$$

has the χ^2 distribution with n degrees of freedom. Hence or otherwise show that the sum of two independent random variables, having the χ^2 distribution with m and n degrees of freedom, respectively, has the χ^2 distribution with $m + n$ degrees of freedom.

8. Let $X_1, X_2, \ldots$ be independent, identically distributed random variables and let N be a random variable which takes values in the positive integers and is independent of the X_i. Find the moment generating function of

$$S = X_1 + X_2 + \cdots + X_N$$

in terms of the moment generating functions of N and X_1, when these exist.

9. Random variables $X_1, X_2, \ldots, X_N$ have zero expectations, and

$$\mathbb{E}(X_m X_n) = v_{mn} \qquad \text{for } m, n = 1, 2, \ldots, N.$$

Calculate the variance of the random variable

$$Z = \sum_{n=1}^{N} a_n X_n,$$

and deduce that the symmetric matrix $V = (v_{mn})$ is non-negative definite. It is desired to find an $N \times N$ matrix A such that the random variables

$$Y_n = \sum_{r=1}^{N} a_{nr} X_r \qquad \text{for } n = 1, 2, \ldots, N$$

are uncorrelated and have unit variance. Show that this will be the case if and only if

$$AVA' = I,$$

and show that A can be chosen to satisfy this equation if and only if V is non-singular. (Any standard results from matrix theory may, if clearly stated, be used without proof. A' denotes the transpose of A.) (Oxford 1971F).

10. Prove that if $X = X_1 + \cdots + X_n$ and $Y = Y_1 + \cdots + Y_n$, where X_i and Y_j are independent whenever $i \neq j$, then $\operatorname{cov}(X, Y) = \sum_{i=1}^{n} \operatorname{cov}(X_i, Y_i)$. (Assume that all series involved are absolutely convergent.)

    Two players A and B play a series of independent games. The probability that A wins any particular game is p^2, that B wins is q^2, and that the game is a draw is $2pq$, where $p + q = 1$. The winner of a game scores 2 points, the loser none; if a game is drawn, each player scores 1 point. Let X and Y be the number of points scored by A and B, respectively, in a series of n games. Prove that $\operatorname{cov}(X, Y) = -2npq$. (Oxford 1982M)

11. The *joint moment generating function* of two random variables X and Y is defined to be the function $M(s, t)$ of two real variables defined by

$$M(s, t) = \mathbb{E}(e^{sX+tY})$$

for all values of s and t for which this expectation exists. Show that the joint moment generating function of a pair of random variables having the standard bivariate normal distribution (6.73) is

$$M(s, t) = \exp\left[\tfrac{1}{2}(s^2 + 2\rho st + t^2)\right].$$

Deduce the joint moment generating function of a pair of random variables having the bivariate normal distribution (6.76) with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$.

* 12. Let X and Y be independent random variables, each having mean 0, variance 1, and finite moment generating function $M(t)$. If $X + Y$ and $X - Y$ are independent, show that

$$M(2t) = M(t)^3 M(-t)$$

and deduce that X and Y have the normal distribution with mean 0 and variance 1.

13. Let X have moment generating function $M(t)$.

    (a) Show that $M(t)M(-t)$ is the moment generating function of $X - Y$, where Y is independent of X but has the same distribution.

    (b) In a similar way, describe random variables which have moment generating functions

$$\frac{1}{2 - M(t)}, \qquad \int_0^\infty M(ut)e^{-u}\,du.$$

14. *Coupon-collecting problem*. There are c different types of coupon, and each coupon obtained is equally likely to be any one of the c types. Find the moment generating function of the total number N of coupons which you must collect in order to obtain a complete set.

15. Prove that if ϕ_1 and ϕ_2 are characteristic functions, then so is $\phi = \alpha\phi_1 + (1 - \alpha)\phi_2$ for any $\alpha \in \mathbb{R}$ satisfying $0 \leq \alpha \leq 1$.

16. Show that X and $-X$ have the same distribution if and only if ϕ_X is a purely real-valued function.

17. Find the characteristic function of a random variable with density function

$$f(x) = \tfrac{1}{2}e^{-|x|} \qquad \text{for } x \in \mathbb{R}.$$

18. Let $X_1, X_2, \ldots$ be independent random variables each having the Cauchy distribution, and let

$$A_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

Show that A_n has the Cauchy distribution regardless of the value of n.

19. Show that $\phi(t) = \exp(-|t|^\alpha)$ can be the characteristic function of a distribution with finite variance if and only if $\alpha = 2$.

20. Let X be a random variable whose moment generating function $M(t)$ exists for $|t| < h$, where $h > 0$. Let N be a random variable taking positive integer values such that

$$\mathbb{P}(N = k) > 0 \qquad \text{for } k = 1, 2, \ldots.$$

Show that

$$M(t) = \sum_{k=1}^{\infty} \mathbb{P}(N = k)\mathbb{E}(e^{tX} \mid N = k) \qquad \text{for } |t| < h.$$

Let $X = \max\{U_1, U_2, \ldots, U_N\}$, where the U_i are independent random variables uniformly distributed on $(0, 1)$ and N is an independent random variable whose distribution is given by

$$\mathbb{P}(N = k) = \frac{1}{(e-1)k!} \qquad \text{for } k = 1, 2, \ldots.$$

Obtain the moment generating function of X and hence show that if R is another independent random variable with

$$\mathbb{P}(R = r) = (e-1)e^{-r} \qquad \text{for } r = 1, 2, \ldots,$$

then $R - X$ is exponentially distributed. (Oxford 1981F)

21. Let $X_1, X_2, \ldots, X_n$ be independent random variables, each with characteristic function $\phi(t)$. Obtain the characteristic function of

$$Y_n = a_n + b_n(X_1 + X_2 + \cdots + X_n),$$

where a_n and b_n are arbitrary real numbers.

Suppose that $\phi(t) = e^{-|t|^\alpha}$, where $0 < \alpha \leq 2$. Determine a_n and b_n such that Y_n has the same distribution as X_1 for $n = 1, 2, \ldots$. Find the probability density functions of X_1 when $\alpha = 1$ and when $\alpha = 2$. (Oxford 1980F)

22. (a) Suppose that $f(x) = x^m$, where m is a positive integer, and X is a random variable taking values $x_1, x_2, \ldots, x_N \geq 0$ with equal probabilities, and where the sum $x_1 + x_2 + \cdots + x_N = 1$. Deduce from Jensen's inequality that

$$\sum_{i=1}^{N} f(x_i) \geq Nf(1/N).$$

(b) There are N horses that compete in m races. The results of different races are independent. The probability of horse i winning any given race is $p_i \geq 0$, with $p_1 + p_2 + \cdots + p_N = 1$. Let Q be the probability that the same horse wins all m races. Express Q as a polynomial of degree m in the variables $p_1, p_2, \ldots, p_N$.
Prove that $Q \geq N^{1-m}$.

(Cambridge 2010)

23. Define the moment generating function of a random variable X.

    If X and Y are independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$, respectively, find the moment generating function of $X + Y$.

    For $n = 1, 2, \ldots$, let X_n have probability density function

    $$f_n(x) = \frac{1}{(n-1)!} x^{n-1} e^{-x} \qquad \text{for } x > 0.$$

    Find the moment generating function of X_n.

    Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables, each having the same distribution as X_1. Find the moment generating function of $\sum_{i=1}^{n} Y_i$, and deduce its distribution. (Oxford 2005)

24. (a) Let X be an exponential random variable with parameter λ. Find the moment generating function of X, and hence find $\mathbb{E}(X^3)$.
    (b) Let X_1 and X_2 be independent random variables with moment generating functions $M_1(t)$ and $M_2(t)$. Find random variables with the following moment generating functions:
        (i) $e^{bt} M_1(at)$,
        (ii) $M_1(t) M_2(t)$,
        (iii) $[M_1(t)]^2$,
        (iv) $\int_0^1 M_1(ut)\, du$.
    (c) Suppose Y has moment generating function $M_Y(t)$, where

    $$M_Y(t) = \frac{1}{2(1-t)} + \frac{1}{2-t}.$$

    Find $\mathbb{P}(Y \leq 1)$.

    (Oxford 2010)

25. Let $p \geq 1$. By Jensen's inequality or otherwise, find the smallest value of the constant c_p such that $(a+b)^p \leq c_p(a^p + b^p)$ for all $a, b \geq 0$. (Cambridge 2006)

26. *Lyapunov's inequality.* Let Z be a positive random variable. By Jensen's inequality or otherwise, show that $\mathbb{E}(Z^r)^{1/r} \geq \mathbb{E}(Z^s)^{1/s}$ when $r \geq s > 0$. Thus, if Z has finite rth moment, then it has finite sth moment, for $r \geq s > 0$.

8

The main limit theorems

Summary. The law of large numbers and the central limit theorem are two of the principal results of probability theory. The weak law of large numbers is derived from the mean-square law via Chebyshev's inequality. The central limit theorem is proved using the continuity theorem for moment generating functions. A short account is presented of Cramér's large deviation theorem for sums of random variables. Convergence in distribution (or 'weak convergence') is introduced, and the continuity theorem for characteristic functions stated.

8.1 The law of averages

We aim in this chapter to describe the two main limit theorems of probability theory, namely the 'law of large numbers' and the 'central limit theorem'. We begin with the law of large numbers.

Here is an example of the type of phenomenon which we are thinking about. Before writing this sentence, we threw a fair die one million times (with the aid of a computer, actually) and kept a record of the results. The average of the numbers which we threw was 3.500867. Since the mean outcome of each throw is $\frac{1}{6}(1 + 2 + \cdots + 6) = 3\frac{1}{2}$, this number is not too surprising. If x_i is the result of the ith throw, most people would accept that the running average

$$a_n = \frac{1}{n}(x_i + x_2 + \cdots + x_n) \tag{8.1}$$

approaches the mean value $3\frac{1}{2}$ as n gets larger and larger. Indeed, the foundations of probability theory are based upon our belief that sums of the form (8.1) converge to some limit as $n \to \infty$. It is upon the ideas of 'repeated experimentation' and 'the law of averages' that many of our notions of chance are founded. Accordingly, we should like to find a theorem of probability theory which says something like 'if we repeat an experiment many times, then the average of the results approaches the underlying mean value'.

With the above example about throwing a die in the backs of our minds, we suppose that we have a sequence $X_1, X_2, \ldots$ of independent and identically distributed random variables, each having mean value μ. We should like to prove that the average

$$\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \tag{8.2}$$

converges as $n \to \infty$ to the underlying mean value μ. There are various ways in which random variables can be said to converge (advanced textbooks generally list four to six such ways). One simple way is as follows.

Definition 8.3 *We say that the sequence $Z_1, Z_2, \ldots$ of random variables* ***converges in mean square to the (limit) random variable*** *Z if*

$$\mathbb{E}\big([Z_n - Z]^2\big) \to 0 \qquad \text{as } n \to \infty. \tag{8.4}$$

If this holds, we write '$Z_n \to Z$ in mean square as $n \to \infty$'.

Here is a word of motivation for this definition. Remember that if Y is a random variable and $\mathbb{E}(Y^2) = 0$, then Y equals 0 with probability 1. If $\mathbb{E}([Z_n - Z]^2) \to 0$, then it follows that $Z_n - Z$ tends to 0 (in some sense) as $n \to \infty$.

Example 8.5 Let Z_n be a discrete random variable with mass function

$$\mathbb{P}(Z_n = 1) = \frac{1}{n}, \qquad \mathbb{P}(Z_n = 2) = 1 - \frac{1}{n}.$$

Then Z_n converges to the constant random variable 2 in mean square as $n \to \infty$, since

$$\begin{aligned}\mathbb{E}\big([Z_n - 2]^2\big) &= (1-2)^2\frac{1}{n} + (2-2)^2\left(1 - \frac{1}{n}\right) \\ &= \frac{1}{n} \to 0 \qquad \text{as } n \to \infty.\end{aligned}$$

△

It is often quite simple to show convergence in mean square: just calculate a certain expectation and take the limit as $n \to \infty$. It is this type of convergence which appears in our first law of large numbers.

Theorem 8.6 (Mean-square law of large numbers) *Let $X_1, X_2, \ldots$ be a sequence of independent random variables, each with mean μ and variance σ^2. The average of the first n of the X_i satisfies, as $n \to \infty$,*

$$\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \to \mu \qquad \textit{in mean square.} \tag{8.7}$$

Proof This is a straightforward calculation. We write

$$S_n = X_1 + X_2 + \cdots + X_n$$

for the nth partial sum of the X_i. Then

$$\mathbb{E}\left(\frac{1}{n}S_n\right) = \frac{1}{n}\mathbb{E}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}n\mu = \mu,$$

and so

$$\begin{aligned}\mathbb{E}\left(\left[\frac{1}{n}S_n - \mu\right]^2\right) &= \operatorname{var}\left(\frac{1}{n}S_n\right) \\ &= \frac{1}{n^2}\operatorname{var}\big(X_1 + X_2 + \cdots + X_n\big) \quad \text{by (7.17)} \\ &= \frac{1}{n^2}\big(\operatorname{var} X_1 + \cdots + \operatorname{var} X_n\big) \quad \text{by independence and (7.24)} \\ &= \frac{1}{n^2}n\sigma^2 = \frac{1}{n}\sigma^2 \to 0 \quad \text{as } n \to \infty.\end{aligned}$$

Hence, $n^{-1}S_n \to \mu$ in mean square as $n \to \infty$. □

It is customary to assume that the random variables in the law of large numbers are identically distributed as well as independent. We demand here only that the X_i have the same mean and variance.

Exercise 8.8 Let Z_n be a discrete random variable with mass function

$$\mathbb{P}(Z_n = n^\alpha) = \frac{1}{n}, \qquad \mathbb{P}(Z_n = 0) = 1 - \frac{1}{n}.$$

Show that Z_n converges to 0 in mean square if and only if $\alpha < \frac{1}{2}$.

Exercise 8.9 Let $Z_1, Z_2, \ldots$ be a sequence of random variables which converges to the random variable Z in mean square. Show that $aZ_n + b \to aZ + b$ in mean square as $n \to \infty$, for any real numbers a and b.

Exercise 8.10 Let N_n be the number of occurrences of 5 or 6 in n throws of a fair die. Use Theorem 8.6 to show that, as $n \to \infty$,

$$\frac{1}{n}N_n \to \frac{1}{3} \qquad \text{in mean square.}$$

Exercise 8.11 Show that the conclusion of the mean-square law of large numbers, Theorem 8.6, remains valid if the assumption that the X_i are independent is replaced by the weaker assumption that they are uncorrelated.

8.2 Chebyshev's inequality and the weak law

The earliest versions of the law of large numbers were found in the eighteenth century and dealt with a form of convergence different from convergence in mean square. This other mode of convergence also has intuitive appeal and is defined in the following way.

Definition 8.12 *We say that the sequence $Z_1, Z_2, \ldots$ of random variables* ***converges in probability*** *to Z as $n \to \infty$ if*

$$\text{for all } \epsilon > 0, \qquad \mathbb{P}(|Z_n - Z| > \epsilon) \to 0 \quad \text{as } n \to \infty. \tag{8.13}$$

If this holds, we write '$Z_n \to Z$ in probability as $n \to \infty$'.

Condition (8.13) requires that for all small positive δ and ϵ and all sufficiently large n, it is the case that $|Z_n - Z| \leq \epsilon$ with probability at least $1 - \delta$.

It is not clear at first sight how the two types of convergence (in mean square and in probability) are related to one another. It turns out that convergence in mean square is a more powerful property than convergence in probability, and we make this more precise in the next theorem.

Theorem 8.14 *If $Z_1, Z_2, \ldots$ is a sequence of random variables and $Z_n \to Z$ in mean square as $n \to \infty$, then $Z_n \to Z$ in probability also.*

The proof of this follows immediately from a famous inequality which is usually ascribed to Chebyshev but which was discovered independently by Bienaymé and others, and is closely related to Markov's inequality, Theorem 7.63. There are many forms of this inequality in the probability literature, and we feel that the following is the simplest.

Theorem 8.15 (Chebyshev's inequality) *If Y is a random variable and $\mathbb{E}(Y^2) < \infty$, then*

$$\mathbb{P}(|Y| \geq t) \leq \frac{1}{t^2}\mathbb{E}(Y^2) \qquad \text{for } t > 0. \tag{8.16}$$

Proof By Markov's inequality, Theorem 7.63, applied to the positive random variable Y^2,

$$\mathbb{P}(|Y| \geq t) = \mathbb{P}(Y^2 \geq t^2) \leq \frac{\mathbb{E}(Y^2)}{t^2},$$

as required. □

Proof of Theorem 8.14 We apply Chebyshev's inequality to the random variable $Y = Z_n - Z$ to find that

$$\mathbb{P}(|Z_n - Z| > \epsilon) \leq \frac{1}{\epsilon^2}\mathbb{E}\big([Z_n - Z]^2\big) \qquad \text{for } \epsilon > 0.$$

If $Z_n \to Z$ in mean square as $n \to \infty$, the right-hand side tends to 0 as $n \to \infty$, and so the left-hand side tends to 0 for all $\epsilon > 0$ as required. □

The converse of Theorem 8.14 is false: there exist sequences of random variables which converge in probability but not in mean square (see Example 8.19).

The mean-square law of large numbers, Theorem 8.6, combines with Theorem 8.14 to produce what is commonly called the 'weak law of large numbers'.

Theorem 8.17 (Weak law of large numbers) *Let $X_1, X_2, \ldots$ be a sequence of independent random variables, each with mean μ and variance σ^2. The average of the first n of the X_i satisfies, as $n \to \infty$,*

$$\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \to \mu \qquad \text{in probability.} \tag{8.18}$$

The principal reason for stating both the mean-square law and the weak law are historical and traditional—the first laws of large numbers to be proved were in terms of convergence in probability. There is also a good mathematical reason for stating the weak law separately—unlike the mean-square law, the conclusion of the weak law is valid without the assumption that the X_i have finite variance so long as they all have the same distribution. This is harder to prove than the form of the weak law presented above, and we defer its proof until Section 8.5 and the treatment of characteristic functions therein.

There are many forms of the laws of large numbers in the literature, and each has a set of assumptions and a set of conclusions. Some are difficult to prove (with weak assumptions and strong conclusions) and others can be quite easy to prove (such as those above). Our selection is simple but contains a number of the vital ideas. Incidentally, the weak law is called 'weak' because it may be formulated in terms of distributions alone. There is a more powerful 'strong law' which concerns intrinsically the convergence of random variables themselves.

Example 8.19 Here is an example of a sequence of random variables which converges in probability but not in mean square. Suppose that Z_n is a random variable with mass function

$$\mathbb{P}(Z_n = 0) = 1 - \frac{1}{n}, \qquad \mathbb{P}(Z_n = n) = \frac{1}{n}.$$

Then, for $\epsilon > 0$ and all large n,

$$\mathbb{P}(|Z_n| > \epsilon) = \mathbb{P}(Z_n = n) = \frac{1}{n} \to 0 \qquad \text{as } n \to \infty,$$

giving that $Z_n \to \infty$ in probability. On the other hand

$$\begin{aligned}\mathbb{E}\big([Z_n - 0]^2\big) = \mathbb{E}(Z_n^2) &= 0 \cdot \left(1 - \frac{1}{n}\right) + n^2 \frac{1}{n} \\ &= n \to \infty \qquad \text{as } n \to \infty,\end{aligned}$$

so Z_n does not converge to 0 in mean square. △

Exercise 8.20 Prove the following alternative form of Chebyshev's inequality: if X is a random variable with finite variance and $a > 0$, then

$$\mathbb{P}\big(|X - \mathbb{E}(X)| > a\big) \leq \frac{1}{a^2} \operatorname{var}(X).$$

Exercise 8.21 Use Chebyshev's inequality to show that the probability that in n throws of a fair die the number of sixes lies between $\frac{1}{6}n - \sqrt{n}$ and $\frac{1}{6}n + \sqrt{n}$ is at least $\frac{31}{36}$.

Exercise 8.22 Show that if $Z_n \to Z$ in probability then, as $n \to \infty$,

$$aZ_n + b \to aZ + b \qquad \text{in probability,}$$

for any real numbers a and b.

8.3 The central limit theorem

Our second main result is the central limit theorem. This also concerns sums of independent random variables. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables, each with mean μ and non-zero variance σ^2. We know from the law of large numbers that the sum $S_n = X_1 + X_2 + \cdots + X_n$ is about as large as $n\mu$ for large n, and the next natural problem is to determine the order of the difference $S_n - n\mu$. It turns out that this difference has order $\sqrt{n}$.

Rather than work with the sum S_n directly, we work with the so-called *standardized* version of S_n,

$$Z_n = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathrm{var}(S_n)}}. \tag{8.23}$$

This is a linear function $Z_n = a_n S_n + b_n$ of S_n, where a_n and b_n have been chosen in such a way that $\mathbb{E}(Z_n) = 0$ and $\mathrm{var}(Z_n) = 1$. Note that

$$\begin{aligned} \mathbb{E}(S_n) &= \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n) \qquad \text{by (6.63)} \\ &= n\mu. \end{aligned}$$

Also,

$$\begin{aligned} \mathrm{var}(S_n) &= \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n) \qquad \text{by independence and (7.24)} \\ &= n\sigma^2, \end{aligned}$$

and so

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}. \tag{8.24}$$

It is a remarkable fact that the distribution of Z_n settles down to a limit as $n \to \infty$. Even more remarkable is the fact that the limiting distribution of Z_n is the normal distribution with mean 0 and variance 1, irrespective of the original distribution of the X_i. This theorem is one of the most beautiful in mathematics and is known as the 'central limit theorem'.

Theorem 8.25 (Central limit theorem) *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables, each with mean μ and non-zero variance σ^2. The standardized version*

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

of the sum $S_n = X_1 + X_2 + \cdots + X_n$ satisfies, as $n \to \infty$,

$$\mathbb{P}(Z_n \le x) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du \qquad \text{for } x \in \mathbb{R}. \tag{8.26}$$

The right-hand side of (8.26) is just the distribution function of the normal distribution with mean 0 and variance 1, and thus (8.26) may be written as

$$\mathbb{P}(Z_n \le x) \to \mathbb{P}(Y \le x) \qquad \text{for } x \in \mathbb{R},$$

where Y is a random variable with this standard normal distribution.

Special cases of the central limit theorem were proved by de Moivre (in about 1733) and Laplace, who considered the case when the X_i have the Bernoulli distribution. Lyapunov proved the first general version in about 1901, but the details of his proof were very complicated. Here we shall give an elegant and short proof based on the method of moment generating functions. As one of our tools, we shall use a special case of a fundamental theorem of analysis, and we present this next without proof. There is therefore a sense in which our 'short and elegant' proof does not live up to that description: it is only a partial proof, since some of the analytical details are packaged elsewhere.

Theorem 8.27 (Continuity theorem) *Let $Z_1, Z_2, \ldots$ be a sequence of random variables with moment generating functions $M_1, M_2, \ldots$ and suppose that, as $n \to \infty$,*

$$M_n(t) \to e^{\frac{1}{2}t^2} \qquad \text{for } t \in \mathbb{R}.$$

Then

$$\mathbb{P}(Z_n \leq x) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du \qquad \text{for } x \in \mathbb{R}.$$

In other words, the distribution function of Z_n converges to the distribution function of the normal distribution if the moment generating function of Z_n converges to the moment generating function of the normal distribution. We shall use this to prove the central limit theorem in the case when the X_i have a common moment generating function

$$M_X(t) = \mathbb{E}(\exp(tX_i)) \qquad \text{for } i = 1, 2, \ldots,$$

although we stress that the central limit theorem is valid even when this expectation does not exist so long as both the mean and the variance of the X_i are finite.

Proof of Theorem 8.25 Let $U_i = X_i - \mu$. Then $U_1, U_2, \ldots$ are independent and identically distributed random variables with mean and variance given by

$$\mathbb{E}(U_i) = 0, \qquad \mathbb{E}(U_i^2) = \text{var}(U_i) = \sigma^2, \tag{8.28}$$

and moment generating function

$$M_U(t) = M_X(t)e^{-\mu t}.$$

Now,

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} U_i,$$

giving that Z_n has moment generating function

$$\begin{aligned} M_n(t) &= \mathbb{E}(\exp(tZ_n)) = \mathbb{E}\left(\exp\left(\frac{t}{\sigma\sqrt{n}} \sum_{i=1}^{n} U_i\right)\right) \\ &= \left[M_U\left(\frac{1}{\sigma\sqrt{n}}\right)\right]^n \qquad \text{by (7.51) and (7.53).} \end{aligned} \tag{8.29}$$

We need to know the behaviour of $M_U(t/(\sigma\sqrt{n}))$ for large n, and to this end we use Theorem 7.55 to expand $M_U(x)$ as a power series about $x = 0$:

$$\begin{aligned} M_U(x) &= 1 + x\mathbb{E}(U_1) + \tfrac{1}{2}x^2\mathbb{E}(U_1^2) + \mathrm{o}(x^2) \\ &= 1 + \tfrac{1}{2}\sigma^2 x^2 + \mathrm{o}(x^2) \qquad \text{by (8.28).} \end{aligned}$$

Substitute this into (8.29) with $x = t/(\sigma\sqrt{n})$ and t fixed to obtain

$$M_n(t) = \left[1 + \frac{t^2}{2n} + \mathrm{o}\left(\frac{1}{n}\right)\right]^n \to e^{\frac{1}{2}t^2} \qquad \text{as } n \to \infty,$$

and the result follows from Theorem 8.27. This proof requires the existence of $M_X(t)$ for values of t near 0 only, and this is consistent with the discussion before Theorem 7.49. We shall see in Example 8.54 how to adapt the proof without this assumption. □

Example 8.30 (Statistical sampling) The central limit theorem has many applications in statistics, and here is one such. An unknown fraction p of the population are jedi knights. It is desired to estimate p with error not exceeding 0.005 by asking a sample of individuals (it is assumed they answer truthfully). How large a sample is needed?

Solution Suppose a sample of n individuals is chosen. Let X_i be the indicator function of the event that the ith such person admits to being a jedi knight, and assume the X_i are independent, Bernoulli random variables with parameter p. Write

$$S_n = \sum_{i=1}^{n} X_i. \tag{8.31}$$

We choose to estimate p with the 'sample mean' $n^{-1}S_n$, which, following statistical notation, we denote as $\widehat{p}$.

We wish to choose n sufficiently large that $|\widehat{p} - p| \le 0.005$. This cannot be done, since $|\widehat{p} - p|$ is a random variable which may (albeit with only small probability) take a value larger than 0.005 for any given n. The accepted approach is to set a maximal level of probability at which an error is permitted to occur. By convention, we take this to be 0.05, and we are thus led to the following problem: find n such that

$$\mathbb{P}\big(|\widehat{p} - p| \le 0.005\big) \ge 0.95.$$

By (8.31), S_n is the sum of independent, identically distributed random variables with mean p and variance $p(1-p)$. The above probability may be written as

$$\begin{aligned} \mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \le 0.005\right) &= \mathbb{P}\left(\frac{|S_n - np|}{\sqrt{np(1-p)}} \le 0.005\sqrt{\frac{n}{p(1-p)}}\right) \\ &= \mathbb{P}\left(\frac{|S_n - \mathbb{E}(S_n)|}{\sqrt{\mathrm{var}(S_n)}} \le 0.005\sqrt{\frac{n}{p(1-p)}}\right). \end{aligned}$$

By the central limit theorem, $(S_n - \mathbb{E}(S_n))/\sqrt{\mathrm{var}(S_n)}$ converges in distribution to the normal distribution, and hence the final probability may be approximated by an integral of the normal density function. Unfortunately, the range of this integral depends on p, which is unknown.

Since $p(1-p) \le \frac{1}{4}$ for $p \in [0,1]$,

$$\mathbb{P}\left(\frac{|S_n - \mathbb{E}(S_n)|}{\sqrt{\operatorname{var}(S_n)}} \le 0.005\sqrt{\frac{n}{p(1-p)}}\right) \ge \mathbb{P}\left(\frac{|S_n - \mathbb{E}(S_n)|}{\sqrt{\operatorname{var}(S_n)}} \le 0.005\sqrt{4n}\right),$$

and the right-hand side is approximately $\mathbb{P}\big(|N| \le 0.005\sqrt{4n}\big)$, where N is normal with mean 0 and variance 1. Therefore,

$$\begin{aligned}\mathbb{P}\big(|\widehat{p} - p| \le 0.005\big) &\gtrapprox \int_{-0.005\sqrt{4n}}^{0.005\sqrt{4n}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\,du \\ &= 2\Phi(0.005\sqrt{4n}) - 1,\end{aligned}$$

where Φ is the distribution function of N. On consulting statistical tables, we find this to be greater than 0.95 if $0.005\sqrt{4n} \ge 1.96$, which is to say that $n \gtrapprox 40{,}000$. △

Exercise 8.32 A fair die is thrown 12,000 times. Use the central limit theorem to find values of a and b such that

$$\mathbb{P}(1900 < S < 2200) \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\,dx,$$

where S is the total number of sixes thrown.

Exercise 8.33 For $n = 1, 2, \ldots$, let X_n be a random variable having the gamma distribution with parameters n and 1. Show that the moment generating function of $Z_n = (X_n - n)/\sqrt{n}$ is

$$M_n(t) = e^{-t\sqrt{n}}\left(1 - \frac{t}{\sqrt{n}}\right)^{-n},$$

and deduce that, as $n \to \infty$,

$$\mathbb{P}(Z_n \le x) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\,du \qquad \text{for } x \in \mathbb{R}.$$

8.4 Large deviations and Cramér's theorem

Let S_n be the sum of n independent, identically distributed random variables with mean μ and variance σ^2. The weak law of large numbers asserts that S_n has approximate order $n\mu$. By the central limit theorem, the deviations of S_n are typically of the order $\sigma\sqrt{n}$. It is unlikely that S_n will deviate from its mean $n\mu$ by more than order n^α with $\alpha > \frac{1}{2}$. The study of such unlikely events has proved extremely fruitful in recent decades. The following theorem, proved in its original form by Cramér in 1938, is of enormous practical use within the modern theory of 'large deviations', despite the low probability of the events under study.

Let $X_1, X_2, \ldots$ be independent, identically distributed random variables, and $S_n = X_1 + X_2 + \cdots + X_n$. For simplicity, we assume that the X_i have common mean 0; if this does not hold, we replace X_i by $X_i - \mu$. We shall assume quite a lot of regularity on the distribution

of the X_i, namely that the common moment generating function $M(t) = \mathbb{E}(e^{tX})$ satisfies $M(t) < \infty$ for values of t in some neighbourhood $(-\delta, \delta)$ of the origin. Let $t > 0$. The function $g(x) = e^{tx}$ is strictly increasing on $\mathbb{R}$, so that $S_n > an$ if and only if $g(S_n) > g(an)$. By Markov's inequality, Theorem 7.63,

$$\begin{aligned}\mathbb{P}(S_n > an) &= \mathbb{P}\big(g(S_n) > g(an)\big) \\ &\leq \frac{\mathbb{E}(g(S_n))}{g(an)} = \frac{\mathbb{E}(e^{tS_n})}{e^{tan}}.\end{aligned}$$

By Theorem 7.52, $\mathbb{E}(e^{tS_n}) = M(t)^n$, and so

$$\mathbb{P}(S_n > an) \leq \left(\frac{M(t)}{e^{at}}\right)^n \qquad \text{for } t > 0.$$

This provides an upper bound for the chance of a 'large deviation' of S_n from its mean 0, in terms of the arbitrary constant $t > 0$. We minimize the right-hand side over t to obtain

$$\mathbb{P}(S_n > an) \leq \Big[\inf\big\{e^{-at}M(t) : t > 0\big\}\Big]^n. \tag{8.34}$$

This is an exponentially decaying bound for the probability of a large deviation.

It turns out that, neglecting sub-exponential corrections, the bound (8.34) is an equality, and this is the content of Cramér's theorem, Theorem 8.36. The precise result is usually stated in logarithmic form. Let $\Lambda(t) = \log M(t)$, and define the so-called *Fenchel–Legendre transform* of Λ by

$$\Lambda^*(a) = \sup\big\{at - \Lambda(t) : t \in \mathbb{R}\big\}, \qquad a \in \mathbb{R}. \tag{8.35}$$

The function Λ^* is illustrated in Figure 8.1.

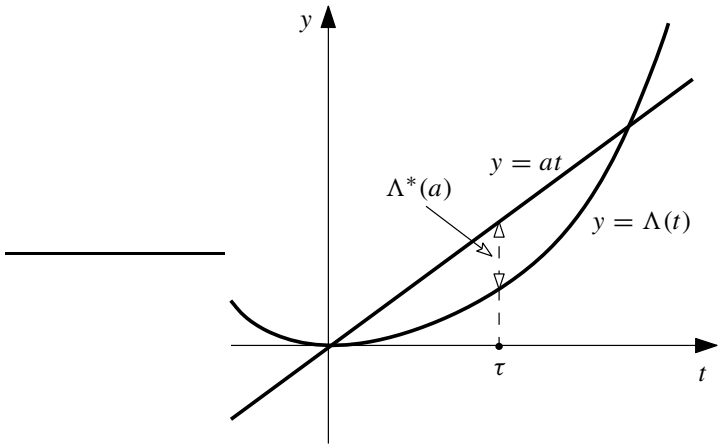


Fig. 8.1 The function $\Lambda(t)$ plotted against the line $y = at$, in the case when $\Lambda'(0) = 0$. The point τ marks the value of t at which $at - \Lambda(t)$ is a maximum, and this maximum is denoted $\Lambda^*(a)$.

Theorem 8.36 (Large deviation theorem) *Let $X_1, X_2, \dots$ be independent, identically distributed random variables with mean 0, whose common moment generating function $M(t) = \mathbb{E}(e^{tX})$ is finite in some neighbourhood of the origin. Let $a > 0$ be such that $\mathbb{P}(X > a) > 0$. Then $\Lambda^*(a) > 0$ and*

$$\frac{1}{n} \log \mathbb{P}(S_n > an) \to -\Lambda^*(a) \qquad \text{as } n \to \infty. \tag{8.37}$$

That is to say, $\mathbb{P}(S_n > na)$ decays to 0 in the manner of $\exp\{-n\Lambda^*(a)\}$. If $\mathbb{P}(X > a) = 0$, then $\mathbb{P}(S_n > na) = 0$ for all n. Theorem 8.36 accounts for deviations *above* the mean. For deviations *below* the mean, the theorem may be applied to the sequence $-X_i$.

Partial proof We begin with some properties of the function $\Lambda = \log M$. First,

$$\Lambda(0) = \log M(0) = 0, \qquad \Lambda'(0) = \frac{M'(0)}{M(0)} = \mathbb{E}(X) = 0.$$

Next,

$$\Lambda''(t) = \frac{M(t)M''(t) - M'(t)^2}{M(t)^2} = \frac{\mathbb{E}(e^{tX})\mathbb{E}(X^2e^{tX}) - \mathbb{E}(Xe^{tX})^2}{M(t)^2}.$$

By the Cauchy–Schwarz inequality, Theorem 7.30, applied to the random variables $Y = Xe^{\frac{1}{2}tX}$ and $Z = e^{\frac{1}{2}tX}$, the numerator is positive. Therefore, Λ is a convex function wherever it is finite (see Exercise 7.73).

We turn now to Figure 8.1. Since Λ is convex with $\Lambda'(0) = 0$, and since $a > 0$, the supremum of $at - \Lambda(t)$ over $t \in \mathbb{R}$ is unchanged by the restriction $t > 0$. That is,

$$\Lambda^*(a) = \sup\bigl\{at - \Lambda(t) : t > 0\bigr\}, \qquad a > 0. \tag{8.38}$$

Next, we show that $\Lambda^*(a) > 0$ under the conditions of the theorem. By Theorem 7.55,

$$at - \Lambda(t) = \log\left(\frac{e^{at}}{M(t)}\right) = \log\left(\frac{1 + at + \mathrm{o}(t)}{1 + \frac{1}{2}\sigma^2t^2 + \mathrm{o}(t^2)}\right)$$

for small positive t, where $\sigma^2 = \mathrm{var}(X)$. This is where we have used the assumption that $M(t) < \infty$ on a neighbourhood of the origin. For sufficiently small positive t,

$$1 + at + \mathrm{o}(t) > 1 + \tfrac{1}{2}\sigma^2t^2 + \mathrm{o}(t^2),$$

whence $\Lambda^*(a) > 0$ by (8.38).

It is immediate from (8.34) and (8.38) that

$$\frac{1}{n} \log \mathbb{P}(S_n > an) \le -\Lambda^*(a), \qquad n \ge 1. \tag{8.39}$$

The proof of the sharpness of the limit in (8.37) is more complicated, and is omitted. A full proof may be found in Grimmett and Stirzaker (2001, Sect. 5.11). □

Example 8.40 Let X be a random variable with distribution

$$\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = \tfrac{1}{2},$$

and moment generating function

$$M(t) = \mathbb{E}(e^{tX}) = \tfrac{1}{2}(e^t + e^{-t}).$$

Let $a \in (0, 1)$. By (8.35), the Fenchel–Legendre transformation of $\Lambda(t) = \log M(t)$ is obtained by maximizing $at - \Lambda(t)$ over the variable t. The function Λ is differentiable, and therefore the maximum may be found by calculus. We have that

$$\frac{d}{dt}[at - \Lambda(t)] = a - \Lambda'(t) = a - \frac{M'(t)}{M(t)} = a - \frac{e^t - e^{-t}}{e^t + e^{-t}}.$$

Setting this equal to 0, we find that

$$e^t = \sqrt{\frac{1+a}{1-a}},$$

and hence

$$\Lambda^*(a) = \log\left(\sqrt{(1-a)^{1-a}(1+a)^{1+a}}\right), \qquad 0 < a < 1.$$

Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of n independent copies of X. By the large deviation theorem, Theorem 8.36,

$$\left[\mathbb{P}(S_n > an)\right]^{1/n} \to e^{-\Lambda^*(a)} = \frac{1}{\sqrt{(1-a)^{1-a}(1+a)^{1+a}}} \qquad \text{as } n \to \infty, \tag{8.41}$$

for $a \in (0, 1)$. △

Exercise 8.42 Find the Fenchel–Legendre transform Λ^* in the case of the normal distribution with mean 0 and variance 1.

Exercise 8.43 Show that the moment generating function of a random variable X is finite on a neighbourhood of the origin if and only if there exist $a, b > 0$ such that $\mathbb{P}(|X| \geq x) \leq ae^{-bx}$ for $x > 0$.

Exercise 8.44 Let $X_1, X_2, \ldots$ be independent random variables with the Cauchy distribution, and let $S_n = X_1 + X_2 + \cdots + X_n$. Find $\mathbb{P}(S_n \geq an)$ for $a > 0$.

8.5 Convergence in distribution, and characteristic functions

We have now encountered the ideas of convergence in mean square and convergence in probability, and we have seen that the former implies the latter. To these two types of convergence we are about to add a third. We motivate this by recalling the conclusion of the central limit theorem, Theorem 8.25: the distribution function of the standardized sum Z_n converges as $n \to \infty$ to the distribution function of the normal distribution. This notion of the convergence of distribution functions may be set in a more general context as follows.

Definition 8.45 *The sequence $Z_1, Z_2, \ldots$ is said to* ***converge in distribution****, or to* ***converge weakly****, to Z as $n \to \infty$ if*

$$\mathbb{P}(Z_n \leq x) \to \mathbb{P}(Z \leq x) \qquad \text{for } x \in C,$$

where C is the set of reals at which the distribution function $F_Z(z) = \mathbb{P}(Z \leq z)$ is continuous. If this holds, we write $Z_n \Rightarrow Z$.

The condition involving points of continuity is an unfortunate complication of the definition, but turns out to be desirable (see Exercise 8.56).

Convergence in distribution is a property of the *distributions* of random variables rather than a property of the random variables themselves, and for this reason, explicit reference to the limit random variable Z is often omitted. For example, the conclusion of the central limit theorem may be expressed as 'Z_n converges in distribution to the normal distribution with mean 0 and variance 1'.

Theorem 8.14 asserts that convergence in mean square implies convergence in probability. It turns out that convergence in distribution is weaker than both of these.

Theorem 8.46 *If $Z_1, Z_2, \ldots$ is a sequence of random variables and $Z_n \to Z$ in probability as $n \to \infty$, then $Z_n \Rightarrow Z$.*

The converse assertion is generally false; see the forthcoming Example 8.49 for a sequence of random variables which converges in distribution but not in probability. The next theorem describes a partial converse.

Theorem 8.47 *Let $Z_1, Z_2, \ldots$ be a sequence of random variables which converges in distribution to the constant c. Then Z_n converges to c in probability also.*

Proof of Theorem 8.46 Suppose $Z_n \to Z$ in probability, and write

$$F_n(z) = \mathbb{P}(Z_n \leq z), \qquad F(z) = \mathbb{P}(Z \leq z)$$

for the distribution functions of Z_n and Z. Let $\epsilon > 0$, and suppose that F is continuous at the point z. Then

$$\begin{aligned} F_n(z) &= \mathbb{P}(Z_n \leq z) \\ &= \mathbb{P}(Z_n \leq z,\ Z \leq z + \epsilon) + \mathbb{P}(Z_n \leq z,\ Z > z + \epsilon) \\ &\leq \mathbb{P}(Z \leq z + \epsilon) + \mathbb{P}(Z - Z_n > \epsilon) \\ &\leq F(z + e) + \mathbb{P}(|Z_n - Z| > \epsilon). \end{aligned}$$

Similarly,

$$\begin{aligned} F(z-\epsilon) &= \mathbb{P}(Z \le z-\epsilon) \\ &= \mathbb{P}(Z \le z-\epsilon,\ Z_n \le z) + \mathbb{P}(Z \le z-\epsilon,\ Z_n > z) \\ &\le \mathbb{P}(Z_n \le z) + \mathbb{P}(Z_n - Z > \epsilon) \\ &\le F_n(z) + \mathbb{P}(|Z_n - Z| > \epsilon). \end{aligned}$$

Thus

$$F(z-\epsilon) - \mathbb{P}(|Z_n - Z| > \epsilon) \le F_n(z) \le F(z+\epsilon) + \mathbb{P}(|Z_n - Z| > \epsilon). \tag{8.48}$$

We let $n \to \infty$ and $\epsilon \downarrow 0$ throughout these inequalities. The left-hand side of (8.48) behaves as follows:

$$\begin{aligned} F(z-\epsilon) - \mathbb{P}(|Z_n - Z| > \epsilon) &\to F(z-\epsilon) && \text{as } n \to \infty \\ &\to F(z) && \text{as } \epsilon \downarrow 0, \end{aligned}$$

where we have used the facts that $Z_n \to Z$ in probability and that F is continuous at z, respectively. Similarly, the right-hand side of (8.48) satisfies

$$\begin{aligned} F(z+\epsilon) + \mathbb{P}(|Z_n - Z| > \epsilon) &\to F(z+\epsilon) && \text{as } n \to \infty \\ &\to F(z) && \text{as } \epsilon \downarrow 0. \end{aligned}$$

Thus, the left- and right-hand sides of (8.48) have the same limit $F(z)$, implying that the central term $F_n(z)$ satisfies $F_n(z) \to F(z)$ as $n \to \infty$. Hence $Z_n \Rightarrow Z$. □

Proof of Theorem 8.47 Suppose that $Z_n \Rightarrow c$. It follows that the distribution function F_n of Z_n satisfies

$$F_n(z) \to \begin{cases} 0 & \text{if } z < c, \\ 1 & \text{if } z > c. \end{cases}$$

Thus, for $\epsilon > 0$,

$$\begin{aligned} \mathbb{P}(|Z_n - c| > \epsilon) &= \mathbb{P}(Z_n < c-\epsilon) + \mathbb{P}(Z_n > c+\epsilon) \\ &\le F_n(c-\epsilon) + 1 - F_n(c+\epsilon) \\ &\to 0 + 1 - 1 = 0 \qquad \text{as } n \to \infty. \end{aligned}$$

□

The following is an example of a sequence of random variables which converges in distribution but not in probability.

Example 8.49 Let U be a random variable which takes the values -1 and 1, each with probability $\frac{1}{2}$. We define the sequence $Z_1, Z_2, \ldots$ by

$$Z_n = \begin{cases} U & \text{if } n \text{ is odd}, \\ -U & \text{if } n \text{ is even}. \end{cases}$$

It is clear that $Z_n \Rightarrow U$, since each Z_n has the same distribution. On the other hand

$$Z_n - U = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ -2U & \text{if } n \text{ is even,} \end{cases}$$

so that $\mathbb{P}(|Z_{2m} - U| > 1) = \mathbb{P}(|U| > \frac{1}{2}) = 1$ for all m. Hence, Z_n does not converge to U in probability. △

Finally, we return to characteristic functions. In proving the central limit theorem we employed a result (Theorem 8.27) linking the convergence of moment generating functions to convergence in distribution. This result is a weak form of the so-called continuity theorem, a more powerful version of which we present next (the proof is omitted).

Theorem 8.50 (Continuity theorem) *Let $Z, Z_1, Z_2, \ldots$ be random variables with characteristic functions $\phi, \phi_1, \phi_2, \ldots$. Then $Z_n \Rightarrow Z$ if and only if*

$$\phi_n(t) \to \phi(t) \qquad \text{for } t \in \mathbb{R}.$$

This is a difficult theorem to prove—see Feller (1971, p. 481). We close the section with several examples of this theorem in action.

Example 8.51 Suppose that $Z_n \Rightarrow Z$ and $a, b \in \mathbb{R}$. Prove that $aZ_n + b \Rightarrow aZ + b$.

Solution Let ϕ_n be the characteristic function of Z_n and ϕ the characteristic function of Z. By the continuity theorem, Theorem 8.50, $\phi_n(t) \to \phi(t)$ as $n \to \infty$. The characteristic function of $aZ_n + b$ is

$$\begin{aligned} \phi_{aZ_n+b}(t) &= e^{itb}\phi_n(at) && \text{by Theorem 7.87} \\ &\to e^{itb}\phi(at) && \text{as } n \to \infty \\ &= \phi_{aZ+b}(t), \end{aligned}$$

and the result follows by another appeal to Theorem 8.50. A direct proof of this fact using distribution functions is messy when a is negative. △

Example 8.52 (The weak law) Here is another proof of the weak law of large numbers, Theorem 8.17, for the case of identically distributed random variables. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean μ, and let

$$U_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

By Theorem 7.87, the characteristic function ψ_n of U_n is given by

$$\psi_n(t) = \phi_X(t/n)^n, \tag{8.53}$$

where ϕ_X is the common characteristic function of the X_i. By Theorem 7.85,

$$\phi_X(t) = 1 + it\mu + \mathrm{o}(t) \qquad \text{as } t \to 0.$$

Substitute this into (8.53) to obtain

$$\psi_n(t) = \left[1 + \frac{it\mu}{n} + \mathrm{o}\left(\frac{t}{n}\right)\right]^n \to e^{i\mu t} \qquad \text{as } n \to \infty.$$

The limit here is the characteristic function of the constant μ, and thus the continuity theorem, Theorem 8.50, implies that $U_n \Rightarrow \mu$. A glance at Theorem 8.47 confirms that the convergence takes place in probability also, and we have proved a version of the weak law of large numbers. This version differs from the earlier one in two regards—we have assumed that the X_i are identically distributed, but we have made no assumption that they have finite variance. △

Example 8.54 Central limit theorem. Our proof of the central limit theorem in Section 8.3 was valid only for random variables which possess finite moment generating functions. Very much the same arguments go through using characteristic functions, and thus Theorem 8.25 is true as it is stated. △

Exercise 8.55 Let $X_1, X_2, \ldots$ be independent random variables, each having the Cauchy distribution. Show that $A_n = n^{-1}(X_1 + X_2 + \cdots + X_n)$ converges in distribution to the Cauchy distribution as $n \to \infty$. Compare this with the conclusion of the weak law of large numbers.

Exercise 8.56 Let X_n, Y_n, Z be 'constant' random variables with distributions

$$\mathbb{P}\left(X_n = -\frac{1}{n}\right) = 1, \quad \mathbb{P}\left(Y_n = \frac{1}{n}\right) = 1, \quad \mathbb{P}(Z = 0) = 1.$$

Show that

$$\mathbb{P}(X_n \le x) \to \mathbb{P}(Z \le x) \qquad \text{for } x \in \mathbb{R},$$

but $\mathbb{P}(Y_n \le 0) \not\to \mathbb{P}(Z \le 0)$.

This motivates the condition of continuity in Definition 8.45. Without this condition, it would be the case that $X_n \Rightarrow Z$ but $Y_n \not\Rightarrow Z$.

8.6 Problems

1. Let $X_1, X_2 \ldots$ be independent random variables, each having the uniform distribution on the interval $(0, a)$, and let $Z_n = \max\{X_1, X_2, \ldots, X_n\}$. Show that
   (a) $Z_n \to a$ in probability as $n \to \infty$,
   (b) $\sqrt{Z_n} \to \sqrt{a}$ in probability as $n \to \infty$,
   (c) if $U_n = n(1 - Z_n)$ and $a = 1$, then

$$\mathbb{P}(U_n \le x) \to \begin{cases} 1 - e^{-x} & \text{if } x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

   so that U_n converges in distribution to the exponential distribution as $n \to \infty$.

2. By applying the central limit theorem to a sequence of random variables with the Bernoulli distribution, or otherwise, prove the following result in analysis. If $0 < p = 1 - q < 1$ and $x > 0$, then

$$\sum \binom{n}{k} p^k q^{n-k} \to 2 \int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du \qquad \text{as } n \to \infty,$$

   where the summation is over all values of k satisfying $np - x\sqrt{npq} \le k \le np + x\sqrt{npq}$.

3. Let X_n be a discrete random variable with the binomial distribution, parameters n and p. Show that $n^{-1}X_n$ converges to p in probability as $n \to \infty$.
4. *Binomial–Poisson limit.* Let Z_n have the binomial distribution with parameters n and λ/n, where λ is fixed. Use characteristic functions to show that Z_n converges in distribution to the Poisson distribution, parameter λ, as $n \to \infty$.
5. By applying the central limit theorem to a sequence of random variables with the Poisson distribution, or otherwise, prove that
$$e^{-n}\left(1 + n + \frac{n^2}{2!} + \cdots + \frac{n^n}{n!}\right) \to \frac{1}{2} \qquad \text{as } n \to \infty.$$
6. (a) Let $0 < a < 1$ and
$$T_n = \sum_{k:\,|k-\frac{1}{2}n|>\frac{1}{2}an} \binom{n}{k}.$$
By considering the binomial distribution or otherwise, show that
$$T_n^{1/n} \to \frac{2}{\sqrt{(1+a)^{1+a}(1-a)^{1-a}}}.$$
(b) Find the asymptotic behaviour of $T_n^{1/n}$, where $a > 0$ and
$$T_n = \sum_{k:\,k>n(1+a)} \frac{n^k}{k!}.$$
7. Use the Cauchy–Schwarz inequality to prove that if $X_n \to X$ in mean square and $Y_n \to Y$ in mean square, then $X_n + Y_n \to X + Y$ in mean square.
8. Use the Cauchy–Schwarz inequality to prove that if $X_n \to X$ in mean square, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$. Give an example of a sequence $X_1, X_2, \ldots$ such that $X_n \to X$ in probability but $\mathbb{E}(X_n)$ does not converge to $\mathbb{E}(X)$.
9. If $X_n \to X$ in probability and $Y_n \to Y$ in probability, show that $X_n + Y_n \to X + Y$ in probability.
10. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be independent random variables each having mean μ and non-zero variance σ^2. Show that
$$U_n = \frac{1}{\sqrt{2n\sigma^2}}\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right)$$
satisfies, as $n \to \infty$,
$$\mathbb{P}(U_n \le x) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\,du \qquad \text{for } x \in \mathbb{R}.$$
11. Adapt the proof of Chebyshev's inequality to show that, if X is a random variable and $a > 0$, then
$$\mathbb{P}(|X| \ge a) \le \frac{1}{g(a)}\mathbb{E}(g(X)),$$
for any function $g : \mathbb{R} \to \mathbb{R}$ which satisfies

(a) $g(x) = g(-x)$ for $x \in \mathbb{R}$,
(b) $g(0) > 0$ for $x \neq 0$,
(c) g is increasing on $[0, \infty)$.

12. Let X be a random variable which takes values in the interval $[-M, M]$ only. Show that

$$\mathbb{P}(|X| \geq a) \geq \frac{\mathbb{E}|X| - a}{M - a}$$

if $0 \leq a < M$.

13. Show that $X_n \to 0$ in probability if and only if

$$\mathbb{E}\left(\frac{|X_n|}{1 + |X_n|}\right) \to 0 \qquad \text{as } n \to \infty.$$

14. Let $(X_n : n \geq 1)$ be a sequence of random variables which converges in mean square. Show that $\mathbb{E}\big([X_n - X_m]^2\big) \to 0$ as $m, n \to \infty$.

If $\mathbb{E}(X_n) = \mu$ and $\operatorname{var}(X_n) = \sigma^2$ for all n, show that the correlation between X_n and X_m converges to 1 as $m, n \to \infty$.

15. Let Z have the normal distribution with mean 0 and variance 1. Find $\mathbb{E}(Z^2)$ and $\mathbb{E}(Z^4)$, and find the probability density function of $Y = Z^2$.

* 16. Let $X_1, X_2, \ldots$ be independent random variables each having distribution function F and density function f. The *order statistics* $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ of the subsequence $X_1, X_2, \ldots, X_n$ are obtained by rearranging the values of the X_i in non-decreasing order. That is to say, $X_{(1)}$ is set to the smallest observed value of the X_i, $X_{(2)}$ is set to the second smallest value, and so on, so that $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$. The *sample median* Y_n of the sequence $X_1, X_2, \ldots, X_n$ is the 'middle value', so that Y_n is defined to be

$$Y_n = \begin{cases} X_{(r+1)} & \text{if } n = 2r + 1 \text{ is odd,} \\ \frac{1}{2}(X_{(r)} + X_{(r+1)}) & \text{if } n = 2r \text{ is even.} \end{cases}$$

Assume that $n = 2r + 1$ is odd, and show that Y_n has density function

$$f_n(y) = (r + 1)\binom{n}{r} F(y)^r [1 - F(y)]^r f(y).$$

Deduce that, if F has a unique median m, then

$$\mathbb{P}(Z_n \leq x) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du \qquad \text{for } u \in \mathbb{R},$$

where $Z_n = (Y_n - m)\sqrt{4nf(m)^2}$.

17. The sequence (X_i) of independent, identically distributed random variables is such that

$$\mathbb{P}(X_i = 0) = 1 - p, \qquad \mathbb{P}(X_i = 1) = p.$$

If f is a continuous function on $[0, 1]$, prove that

$$B_n(p) = \mathbb{E}\left(f\left(\frac{X_1 + \cdots + X_n}{n}\right)\right)$$

is a polynomial in p of degree at most n. Use Chebyshev's inequality to prove that for all p with $0 \le p \le 1$, and any $\epsilon > 0$,

$$\sum_{k \in K} \binom{n}{k} p^k (1-p)^{n-k} \le \frac{1}{4n\epsilon^2},$$

where $K = \{k : 0 \le k \le n,\ |k/n - p| > \epsilon\}$. Using this and the fact that f is bounded and uniformly continuous in $[0, 1]$, prove the following version of the Weierstrass approximation theorem:

$$\lim_{n \to \infty} \sup_{0 \le p \le 1} |f(p) - B_n(p)| = 0.$$

(Oxford 1976F)

18. Let Z_n have the geometric distribution with parameter λ/n, where λ is fixed. Show that Z_n/n converges in distribution as $n \to \infty$, and find the limiting distribution.

* 19. Let $(X_k : k = 1, 2, \dots)$ and $(Y_k : k = 1, 2, \dots)$ be two sequences of independent random variables with

$$\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = \tfrac{1}{2}$$

and

$$\mathbb{P}(Y_k = 1) = \mathbb{P}(Y_k = -1) = \frac{1}{2}\left(1 - \frac{1}{k^2}\right),$$
$$\mathbb{P}(Y_k = k+1) = \mathbb{P}(Y_k = -k-1) = \frac{1}{2k^2},$$

for $k = 1, 2, \dots$. Let

$$S_n = \sum_{k=1}^{n} \frac{X_k}{\sqrt{n}}, \qquad T_n = \sum_{k=1}^{n} \frac{Y_k}{\sqrt{n}},$$

and let Z denote a normally distributed random variable with mean 0 and variance 1.

Prove or disprove the following:

(a) S_n converges in distribution to Z,
(b) the mean and variance of T_n converge to the mean and variance of Z,
(c) T_n converges in distribution to Z.

State carefully any theorems which you use. (Oxford 1980F)

* 20. Let X_j, $j = 1, 2, \dots, n$, be independent identically distributed random variables with probability density function $e^{-\frac{1}{2}x^2}/\sqrt{2\pi}$, $-\infty < x < \infty$. Show that the characteristic function of $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ is $(1 - 2i\theta)^{-\frac{1}{2}n}$. Consider a sequence of independent trials where the probability of success is p for each trial. Let N be the number of trials required to obtain a fixed number of k successes. Show that, as p tends to zero, the distribution of $2Np$ tends to the distribution of Y with $n = 2k$. (Oxford 1979F)

21. Let $X_1, X_2, \dots, X_n$ be independent and identically distributed random variables such that

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = \tfrac{1}{2}.$$

Derive the moment generating function of the random variable $Y_n = \sum_{j=1}^{n} a_j X_j$, where $a_1, a_2, \dots, a_n$ are constants.

In the special case $a_j = 2^{-j}$ for $j \ge 1$, show that Y_n converges in distribution as $n \to \infty$ to the uniform distribution on the interval $(-1, 1)$.

* 22. X and Y are independent, identically distributed random variables with mean 0, variance 1, and characteristic function ϕ. If $X + Y$ and $X - Y$ are independent, prove that

$$\phi(2t) = \phi(t)^3\phi(-t).$$

By making the substitution $\gamma(t) = \phi(t)/\phi(-t)$ or otherwise, show that, for any positive integer n,

$$\phi(t) = \left\{1 - \frac{1}{2}\left(\frac{t}{2^n}\right)^2 + \mathrm{o}\left(\left[t/2^n\right]^2\right)\right\}^{4^n}.$$

Hence, find the common distribution of X and Y. (Oxford 1976F)

23. Let $u(t)$ and $v(t)$ be the real and imaginary parts, respectively, of the characteristic function of the random variable X. Prove that

(a) $\mathbb{E}(\cos^2 tX) = \frac{1}{2}[1 + u(2t)]$,

(b) $\mathbb{E}(\cos sX \cos tX) = \frac{1}{2}[u(s + t) + u(s - t)]$.

Hence, find the variance of $\cos tX$ and the covariance of $\cos tX$ and $\cos sX$ in terms of u and v.

Consider the special case when X is uniformly distributed on $[0, 1]$. Are the random variables $\{\cos j\pi X : j = 1, 2, \dots\}$ (i) uncorrelated, (ii) independent? Justify your answers. (Oxford 1975F)

24. State the central limit theorem.

The cumulative distribution function F of the random variable X is continuous and strictly increasing. Show that $Y = F(X)$ is uniformly distributed. Find the probability density function of the random variable $-\log(1 - Y)$, and calculate its mean and variance.

Let $\{X_k\}$ be a sequence of independent random variables whose corresponding cumulative distribution functions $\{F_k\}$ are continuous and strictly increasing. Let

$$Z_n = -\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\big(1 + \log[1 - F_k(X_k)]\big), \qquad n = 1, 2, \dots.$$

Show that, as $n \to \infty$, $\{Z_n\}$ converges in distribution to a normal distribution with mean zero and variance one. (Oxford 2007)

Part C

Random Processes

9

Branching processes

Summary. The branching process is a fundamental model for the random growth of populations. The method of generating functions, and in particular the random sum formula, provides the key to the study of this process. The criterion for the ultimate extinction of a branching process is stated and proved.

9.1 Random processes

Until now, we have been developing the basic terminology and results of probability theory, next, we turn our attention to simple applications. The passing of time plays an essential part in the world which we inhabit, and consequently many applications of probability involve quantities which develop randomly as time passes. Such randomly evolving processes are called *random processes* or *stochastic processes*, and there are many different types of these. Most real processes in nature, such as the pollen count in Phoenix or the position of Swansea City in the football league, develop according to rules which are too complicated to describe exactly, and good probabilistic models for these processes can be very complicated indeed. We shall stick to some of the simplest random processes, and the specific processes which we shall consider in some depth are

(a) *branching processes*: modelling the growth of a self-reproducing population (such as mankind),
(b) *random walks*: modelling the movement of a particle which moves erratically within a medium (a dust particle in the atmosphere, say),
(c) *Poisson processes and related processes*: modelling processes such as the emission of radioactive particles from a slowly decaying source, or the length of the queue at the supermarket cash register.

There is a fairly complete theory of each of these three types of process, of which the main features are described in Chapters 9–11, respectively. In contrast, the *general* theory of stochastic processes is much more challenging and is outside the range of this book. At one extreme, probabilists study 'concrete' processes such as those above, often designed to meet the needs of a particular application area, and at the other extreme there is an abstract theory of 'general' stochastic processes. Any tension between these two extremes is resolved through the identification of key properties which are shared by large families of processes and yet are sufficiently specific to allow the development of a useful theory. Probably the most important

such property is the so-called 'Markov property'. We do not discuss this here, but refer the reader to Chapter 12 for an account of Markov processes in discrete time.

9.2 A model for population growth

We define the term *nomad* to be a type of hypothetical object which is able to reproduce itself according to the following rules. At time $n = 0$, there exists a single nomad. This nomad lives for a unit of time and then, at time $n = 1$, it dies in the act of childbirth and is replaced by a family of offspring nomads. These nomads have similar biographies, each surviving only until time $n = 2$ and then each dying and being replaced by a family of offspring. This death–birth process continues at all subsequent time points $n = 3, 4, \ldots$. If we know the sizes of all the individual nomad families, then we know everything about the development of the nomad population, and we might represent this in the usual way as a family tree (see Figure 9.1). The problem is that different nomads may have different numbers of offspring, and these numbers may not be entirely predictable in advance. We shall assume here that the family sizes are random variables which satisfy the following two conditions:

I. the family sizes are *independent* random variables each taking values in $\{0, 1, 2, \ldots\}$,
II. the family sizes are *identically distributed* random variables with known mass function p, so that the number C of children of a typical nomad has mass function $\mathbb{P}(C = k) = p_k$ for $k = 0, 1, 2, \ldots$.

Such a process is called a *branching process* and may be used as a simple model for bacterial growth or the spread of a family name (to give but two examples). Having established the model, the basic problem is to say something about how the development of the process depends on the family-size mass function p. In order to avoid trivialities, *we shall suppose throughout this chapter that*

$$p_k \neq 1 \qquad \text{for } k = 0, 1, 2, \ldots. \tag{9.1}$$

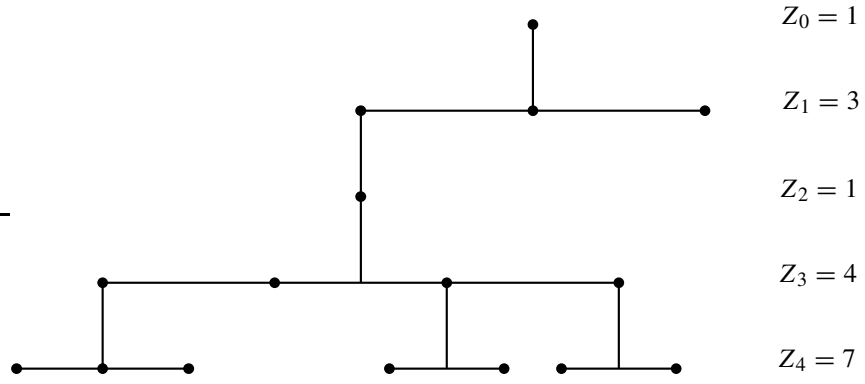


Fig. 9.1 A typical nomad family tree, with generation sizes 1, 3, 1, 4, 7, ...

We introduce some notation. The set of nomads born at time n is called the nth *generation* of the branching process, and we write Z_n for the number of such nomads. The evolution of

the process is described by the sequence $Z_0, Z_1, \ldots$ of random variables, and it is with this sequence that we work. Specific properties of the Z_n are given in the next section, and we close this section with a list of interesting questions.

(a) What is the mean and variance of Z_n?
(b) What is the mass function of Z_n?
(c) What is the probability that nomadkind is extinct by time n?
(d) What is the probability that nomadkind ultimately becomes extinct?

9.3 The generating-function method

The first step in the study of this branching process is to explain how to find the distributions of the Z_i in terms of the family-size mass function p. Clearly, $Z_0 = 1$ and

$$\mathbb{P}(Z_1 = k) = p_k \qquad \text{for } k = 0, 1, 2, \ldots, \tag{9.2}$$

since Z_1 is the number of children of the founding nomad. It is not easy to give the mass function of Z_2 directly, since Z_2 is the sum of a random number Z_1 of random family sizes: writing C_i for the number of children of the ith nomad in the first generation, we have that

$$Z_2 = C_1 + C_2 + \cdots + C_{Z_1},$$

that is, Z_2 is the sum of the family sizes of the Z_1 nomads in the first generation. More generally, for $n = 1, 2, \ldots$,

$$Z_n = C_1' + C_2' + \cdots + C_{Z_{n-1}}', \tag{9.3}$$

where $C_1', C_2', \ldots$ are the numbers of children of the nomads in the $(n-1)$th generation. The sum of a random number of random variables is treated better by using probability generating functions than by using mass functions. We write

$$G_n(s) = \mathbb{E}(s^{Z_n}) = \sum_{k=0}^{\infty} s^k \mathbb{P}(Z_n = k)$$

for the probability generating function of Z_n, and

$$G(s) = \sum_{k=0}^{\infty} s^k p_k$$

for the probability generating function of a typical family size. We wish to express G_n in terms of G, and we do this in the following theorem.

Theorem 9.4 *The probability generating functions G, G_0, $G_1, \ldots$ satisfy*

$$G_0(s) = s, \ G_n(s) = G_{n-1}(G(s)), \qquad \text{for } n = 1, 2, \ldots, \tag{9.5}$$

and hence G_n is the nth iterate of G,

$$G_n(s) = G(G(\cdots G(s)\cdots)) \qquad \text{for } n = 0, 1, 2, \ldots. \tag{9.6}$$

Proof We have $Z_0 = 1$, and so $G_0(s) = s$. Equation (9.3) expresses Z_n as the sum of Z_{n-1} independent random variables, each having probability generating function G, and so the random sum formula, Theorem 4.36, may be applied with $X_i = C_i'$ and $N = Z_{n-1}$ to deduce that

$$G_n(s) = G_{n-1}(G(s)). \tag{9.7}$$

By iteration,

$$\begin{aligned} G_n(s) &= G_{n-1}(G(s)) = G_{n-2}(G(G(s))) = \cdots \\ &= G_1(G(G(\cdots(s)\cdots))), \end{aligned}$$

where $G_1 = G$ by (9.2). □

Theorem 9.4 contains the information necessary for studying the development of the process. The next result is an immediate corollary.

Theorem 9.8 *The mean value of Z_n is*

$$\mathbb{E}(Z_n) = \mu^n, \tag{9.9}$$

where $\mu = \sum_k kp_k$ is the mean of the family-size distribution.

Proof By the theory of probability generating functions,

$$\begin{aligned} \mathbb{E}(Z_n) &= G_n'(1) && \text{by (4.26)} \\ &= G_{n-1}'(G(1))G'(1) && \text{by (9.5)} \\ &= G_{n-1}'(1)G'(1) && \text{since } G(1) = 1, \text{ by (4.9)} \\ &= \mathbb{E}(Z_{n-1})\mu. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}(Z_n) &= \mu\mathbb{E}(Z_{n-1}) = \mu^2\mathbb{E}(Z_{n-2}) = \cdots \\ &= \mu^n\mathbb{E}(Z_0) = \mu^n. \end{aligned}$$ □

The variance of Z_n may be derived similarly in terms of the mean μ and the variance σ^2 of the family-size distribution. See Exercise 9.12.

It follows by Theorem 9.8 that

$$\mathbb{E}(Z_n) \to \begin{cases} 0 & \text{if } \mu < 1, \\ 1 & \text{if } \mu = 1, \\ \infty & \text{if } \mu > 1, \end{cases}$$

indicating that the behaviour of the process depends substantially on which of the three cases $\mu < 1$, $\mu = 1$, $\mu > 1$ holds. We shall see this in more detail in the next two sections, where it

is shown that if $\mu \leq 1$, the nomad population is bound to become extinct, whereas if $\mu > 1$, there is a strictly positive probability that the line of descent of nomads will continue forever. This dependence on the mean family-size μ is quite natural since '$\mu < 1$' means that each nomad gives birth to (on average) strictly fewer nomads than are necessary to fill the gap caused by its death, whereas '$\mu > 1$' means that each death results (on average) in an increase in the population. The case when $\mu = 1$ is called *critical* since then the mean population-size equals 1 for all time; in this case, random fluctuations ensure that the population size will take the value 0 sooner or later, and henceforth nomadkind will be extinct.

Exercise 9.10 Show that, in the above branching process,

$$G_n(s) = G_r(G_{n-r}(s))$$

for any $r = 0, 1, 2, \ldots, n$. This may be proved either directly from the conclusion of Theorem 9.4 or by adapting the method of proof of (9.7).

Exercise 9.11 Suppose that each family size of a branching process contains either one member only (with probability p) or is empty (with probability $1 - p$). Find the probability that the process becomes extinct at or before the nth generation.

Exercise 9.12 Let μ and σ^2 be the mean and variance of the family-size distribution. Adapt the proof of Theorem 9.8 to show that the variance of Z_n, the size of the nth generation of the branching process, is given by

$$\operatorname{var}(Z_n) = \begin{cases} n\sigma^2 & \text{if } \mu = 1, \\ \sigma^2 \mu^{n-1} \dfrac{\mu^n - 1}{\mu - 1} & \text{if } \mu \neq 1. \end{cases}$$

9.4 An example

The key to the analysis of branching processes is the functional equation

$$G_n(s) = G_{n-1}(G(s)), \tag{9.13}$$

relating the probability generating functions of Z_n and Z_{n-1} and derived in Theorem 9.4. There are a few instances in which this equation may be solved in closed form, and we consider one of these cases here. Specifically, we suppose that the mass function of each family size is given by

$$p_k = (\tfrac{1}{2})^{k+1} \qquad \text{for } k = 0, 1, 2, \ldots,$$

so that each family size is one member smaller than a geometrically distributed random variable with parameter $\frac{1}{2}$ (remember (2.16)) and has probability generating function

$$G(s) = \sum_{k=0}^{\infty} s^k (\tfrac{1}{2})^{k+1} = \frac{1}{2 - s} \qquad \text{for } |s| < 2.$$

We proceed as follows in order to solve (9.13). First, if $|s| \leq 1$,

$$G_1(s) = G(s) = \frac{1}{2-s}.$$

Secondly, we apply (9.13) with $n = 2$ to find that

$$\begin{aligned} G_2(s) &= G(G(s)) \\ &= \frac{1}{2-(2-s)^{-1}} = \frac{2-s}{3-2s} \qquad \text{if } |s| \le 1. \end{aligned}$$

The next step gives

$$G_3(s) = G_2(G(s)) = \frac{3-2s}{4-3s} \qquad \text{if } |s| \le 1.$$

It is natural to guess that

$$G_n(s) = \frac{n-(n-1)s}{n+1-ns} \qquad \text{if } |s| \le 1, \tag{9.14}$$

for any $n \ge 1$, and this is proved easily from (9.13), by the method of induction.

The mass function of Z_n follows by expanding the right-hand side of (9.14) as a power series in s, to find that the coefficient of s^k is

$$\mathbb{P}(Z_n = k) = \begin{cases} \dfrac{n}{n+1} & \text{if } k = 0, \\ \dfrac{n^{k-1}}{(n+1)^{k+1}} & \text{if } k = 1, 2, \ldots. \end{cases} \tag{9.15}$$

In particular,

$$\mathbb{P}(Z_n = 0) = \frac{n}{n+1} \to 1 \qquad \text{as } n \to \infty,$$

so that this branching process becomes extinct sooner or later, with probability 1.

There is a more general case of greater interest. Suppose that the mass function of each family size is given by

$$p_k = pq^k \qquad \text{for } k = 0, 1, 2, \ldots,$$

where $0 < p = 1 - q < 1$. The previous example is the case when $p = q = \frac{1}{2}$, but we suppose here that $p \ne \frac{1}{2}$ so that $p \ne q$. In this case,

$$G(s) = \frac{p}{1-qs} \qquad \text{if } |s| < q^{-1},$$

and the solution to (9.13) is

$$G_n(s) = p\frac{(q^n - p^n) - qs(q^{n-1} - p^{n-1})}{(q^{n+1} - p^{n+1}) - qs(q^n - p^n)}, \tag{9.16}$$

valid for $n = 1, 2, \ldots$ and $|s| \le 1$; again, this can be proved from (9.13) by induction on n. The mass function of Z_n is rather more complicated than (9.15) but may be expressed in very much the same way. The probability of extinction is found to be

$$\begin{aligned}\mathbb{P}(Z_n = 0) &= G_n(0) \qquad \text{by (4.9)} \\ &= p\frac{q^n - p^n}{q^{n+1} - p^{n+1}} = \frac{\mu^n - 1}{\mu^{n+1} - 1},\end{aligned}$$

where $\mu = q/p$ is the mean family-size. Hence

$$\mathbb{P}(Z_n = 0) \to \begin{cases} 1 & \text{if } \mu < 1, \\ \mu^{-1} & \text{if } \mu > 1, \end{cases}$$

giving that ultimate extinction is certain if $\mu < 1$ and less than certain if $\mu > 1$. Combined with the result when $p = q = \frac{1}{2}$ and $\mu = q/p = 1$, this shows that ultimate extinction is certain if and only if $\mu \leq 1$. We shall see in the next section that this is a special case of a general result.

Exercise 9.17 Find the mean and variance of Z_n when the family-size distribution is given by $p_k = pq^k$ for $k = 0, 1, 2, \ldots$, and $0 < p = 1 - q < 1$. Deduce that $\text{var}(Z_n) \to 0$ if and only if $p > \frac{1}{2}$.

9.5 The probability of extinction

In the previous example, ultimate extinction of the branching process is certain if and only if the mean family-size μ satisfies $\mu \leq 1$. This conclusion is valid for *all* branching processes (except for the trivial branching process in which every family size equals 1 always), and we shall prove this. First, we define the *extinction probability*

$$e = \mathbb{P}\big(Z_n = 0 \text{ for some } n \geq 0\big).$$

Next, we show how to find e. Let $E_n = \{Z_n = 0\}$ be the event that the branching process is extinct (in that nomadkind has died out) by the nth generation, and let $e_n = \mathbb{P}(E_n)$. Now,

$$\big\{Z_n = 0 \text{ for some } n \geq 0\big\} = \bigcup_{n=0}^{\infty} E_n.$$

If $Z_n = 0$, then necessarily $Z_{n+1} = 0$, so that $E_n \subseteq E_{n+1}$, and in particular $e_n \leq e_{n+1}$. Since the sequence (E_n) of events is increasing, we may use the continuity of probability measures, Theorem 1.54, to obtain that

$$e = \lim_{n\to\infty} e_n. \tag{9.18}$$

How do we calculate e in practice? Clearly, if $p_0 = 0$, then $e = 0$, since all families are non-empty. The next theorem deals with the general case.

Theorem 9.19 (Extinction probability theorem) *The probability e of ultimate extinction is the smallest non-negative root of the equation*

$$x = G(x). \tag{9.20}$$

Proof Since $e_n = \mathbb{P}(Z_n = 0)$, we have by (4.9) that $e_n = G_n(0)$. By (9.5) and (9.6),

$$\begin{aligned} G_n(s) &= G_{n-1}(G(s)) = G(G(\cdots(s)\cdots)) \\ &= G(G_{n-1}(s)). \end{aligned}$$

Set $s = 0$ to find that $e_n = G_n(0)$ satisfies

$$e_n = G(e_{n-1}) \qquad \text{for } n = 1, 2, \ldots \tag{9.21}$$

with the boundary condition $e_0 = 0$. Now take the limit as $n \to \infty$. By (9.18), $e_n \to e$. Furthermore, G is a power series with radius of convergence at least 1, giving that G is continuous on $[0, 1]$. It follows that e is a root of the equation $e = G(e)$, as required.

In order to show that e is the *smallest* non-negative root of (9.20), suppose that η is any non-negative root of (9.20); we shall show that $e \leq \eta$. First, G is non-decreasing on $[0, 1]$ since it has non-negative coefficients, and hence

$$e_1 = G(0) \leq G(\eta) = \eta$$

by (9.21). Similarly,

$$e_2 = G(e_1) \leq G(\eta) = \eta$$

by (9.21), giving by induction that

$$e_n \leq \eta \qquad \text{for } n = 1, 2, \ldots.$$

Hence, $e = \lim_{n\to\infty} e_n \leq \eta$. □

The last theorem explains how to find the probability of ultimate extinction, the next tells us when extinction is *bound* to occur.

Theorem 9.22 (Extinction/survival theorem) *Assume that $p_1 \neq 1$. The probability e of ultimate extinction satisfies $e = 1$ if and only if the mean family-size μ satisfies $\mu \leq 1$.*

We have eliminated the special case with $p_1 = 1$, since in this trivial case all families have size 1, so that $\mu = 1$ and $e = 0$.

Proof We may suppose that $p_0 > 0$, since otherwise $e = 0$ and $\mu > 1$. We have seen that e is the smallest non-negative root of the equation $x = G(x)$. On the interval $[0, 1]$, G is

(a) continuous, since its radius of convergence is at least 1,
(b) non-decreasing, since $G'(x) = \sum_k k x^{k-1} p_k \geq 0$,
(c) convex, since $G''(x) = \sum_k k(k-1)x^{k-2} p_k \geq 0$,

and so a sketch of G looks something like the curves in Figure 9.2. Generally speaking, there are either one or two intersections between the curve $y = G(x)$ and the line $y = x$ in the interval $[0, 1]$. If the derivative $G'(1)$ satisfies $G'(1) > 1$, the geometry of Figure 9.2(a) indicates that there are two distinct intersections (and, in particular $e < 1$). On the other hand, if $G'(1) \leq 1$, Figure 9.2(b) indicates that there is a unique such intersection, and in this case $x = 1$ is the unique root in $[0, 1]$ of the equation $x = G(x)$. However, $G'(1) = \mu$, and in summary $e = 1$ if and only if $\mu \leq 1$. □

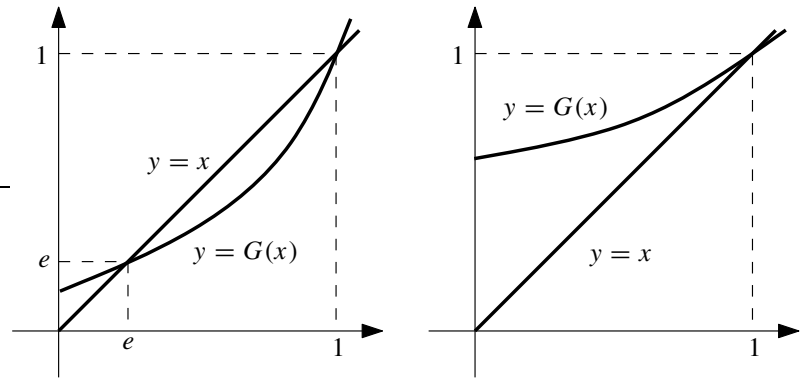


Fig. 9.2 The curve $y = G(x)$ and the line $y = x$, in the two cases (a) $G'(1) > 1$ and (b) $G'(1) \leq 1$.

Exercise 9.23 If the family-size distribution of a branching process has mass function $p_k = pq^k$ for $k = 0, 1, 2, \ldots$ and $0 < p = 1 - q < 1$, use Theorem 9.19 to show that the probability that the process becomes extinct ultimately is p/q if $p \leq \frac{1}{2}$.

Exercise 9.24 If each family size of a branching process has the binomial distribution with parameters 2 and p $(= 1 - q)$, show that the probability of ultimate extinction is

$$e = \begin{cases} 1 & \text{if } 0 \leq p \leq \frac{1}{2}, \\ (q/p)^2 & \text{if } \frac{1}{2} \leq p \leq 1. \end{cases}$$

9.6 Problems

1. Let $X_1, X_2, \ldots$ be independent random variables, each with mean μ and variance σ^2, and let N be a random variable which takes values in the positive integers $\{1, 2, \ldots\}$ and which is independent of the X_i. Show that the sum

$$S = X_1 + X_2 + \cdots + X_N$$

has variance given by

$$\operatorname{var}(S) = \sigma^2 \mathbb{E}(N) + \mu^2 \operatorname{var}(N).$$

If $Z_0, Z_1, \ldots$ are the generation sizes of a branching process in which each family size has mean μ and variance σ^2, use the above fact to show that

$$\begin{aligned} \operatorname{var}(Z_n) &= \sigma^2 \mu^{n-1} + \mu^2 \operatorname{var}(Z_{n-1}), \\ &= \mu \operatorname{var}(Z_{n-1}) + \sigma^2 \mu^{2n-2}. \end{aligned}$$

Deduce an expression for $\operatorname{var}(Z_n)$ in terms of μ, σ^2, and n.

2. Use the result of Problem 9.6.1 to show that, if $Z_0, Z_1, \ldots$ is a branching process whose family sizes have mean μ (> 1) and variance σ^2, then $\text{var}(Z_n/\mu^n) \to \sigma^2/[\mu(\mu-1)]$ as $n \to \infty$.
3. By using the partition theorem and conditioning on the value of Z_m, show that if $Z_0, Z_1, \ldots$ is a branching process with mean family-size μ, then

$$\mathbb{E}(Z_m Z_n) = \mu^{n-m}\mathbb{E}(Z_m^2) \qquad \text{if } m < n.$$

4. If $(Z_n : 0 \le n < \infty)$ is a branching process in which $Z_0 = 1$ and the size of the rth generation Z_r has the generating function $P_r(s)$, prove that

$$P_n(s) = P_r(P_{n-r}(s)) \qquad \text{for } 1 \le r \le n-1.$$

Suppose that the process is modified so that the initial generation Z_0 is Poisson with parameter λ, and the number of offspring of each individual is also Poisson with parameter λ. Find a function f such that if $H_n(s)$ is the generating function of the total number of individuals $Z_0 + Z_1 + \cdots + Z_n$, then

$$H_n(s) = f(sH_{n-1}(s)).$$

(Oxford 1977F)

5. A branching process $(X_n : n \ge 0)$ has $\mathbb{P}(X_0 = 1) = 1$. Let the total number of individuals in the first n generations of the process be Z_n, with probability generating function Q_n. Prove that, for $n \ge 2$,

$$Q_n(s) = sP_1(Q_{n-1}(s)),$$

where P_1 is the probability generating function of the family-size distribution. (Oxford 1974F)

6. (a) Explain what is meant by the term 'branching process'.
   (b) Let X_n be the size of the nth generation of a branching process in which each family size has probability generating function G, and assume that $X_0 = 1$. Show that the probability generating function G_n of X_n satisfies $G_{n+1}(s) = G_n(G(s))$ for $n \ge 1$.
   (c) Show that $G(s) = 1 - \alpha(1-s)^\beta$ is the probability generating function of a non-negative integer-valued random variable when $\alpha, \beta \in (0, 1)$, and find G_n explicitly when G is thus given.
   (d) Find the probability that $X_n = 0$, and show that it converges as $n \to \infty$ to $1 - \alpha^{1/(1-\beta)}$. Explain why this implies that the probability of ultimate extinction equals $1 - \alpha^{1/(1-\beta)}$.

(Cambridge 2001)

10
Random walks

Summary. After a general introduction to random walks, it is proved that a one-dimensional simple random walk is recurrent if and only if it is symmetric. The Gambler's Ruin Problem is discussed.

10.1 One-dimensional random walks

There are many practical instances of random walks. Many processes in physics involve atomic and sub-atomic particles which migrate about the space which they inhabit, and we may often model such motions by random-walk processes. In addition, random walks may often be detected in non-physical disguises, such as in models for gambling, epidemic spread, and stockmarket indices. We shall consider only the simplest type of random walk in this chapter, and we describe this in terms of a hypothetical particle which inhabits a one-dimensional set (that is to say, a line) and which moves randomly within this set as time passes.

For simplicity, we assume that both space and time are *discrete*. We shall observe the particle's position at each of the discrete time-points $0, 1, 2, \ldots$, and we assume that at each of these times the particle is located at one of the integer positions $\ldots, -2, -1, 0, 1, 2, \ldots$ of the real line. The particle moves in the following way. Let p be a real number satisfying $0 < p < 1$, and let $q = 1 - p$. During each unit of time, the particle moves its location either one unit leftwards (with probability q) or one unit rightwards (with probability p). More specifically, if S_n denotes the location of the particle at time n, then

$$S_{n+1} = \begin{cases} S_n + 1 & \text{with probability } p, \\ S_n - 1 & \text{with probability } q, \end{cases}$$

and we suppose that the random direction of each jump is independent of all earlier jumps. Therefore,

$$S_n = S_0 + X_1 + X_2 + \cdots + X_n \qquad \text{for } n = 0, 1, 2, \ldots, \tag{10.1}$$

where S_0 is the starting position and $X_1, X_2, \ldots$ are independent random variables, each taking either the value -1 with probability q or the value 1 with probability p. We call the process $S_0, S_1, \ldots$ a *simple random walk*. It is called *symmetric* random walk if $p = q = \frac{1}{2}$ and *asymmetric* random walk otherwise.

Random walks are examples of so-called Markov chains, and as such are studied in Chapter 12. In particular, the one-dimensional calculations of the current chapter are extended there to a general number d of dimensions (see Section 12.5).

Example 10.2 (Gambling) A gambler enters a casino with £1000 in his pocket and sits at a table, where he proceeds to play the following game. The croupier flips a coin repeatedly, and on each flip the coin shows heads with probability p and tails with probability $q = 1 - p$. Whenever the coin shows heads, the casino pays the gambler £1, and whenever the coin shows tails, the gambler pays the casino £1. That is, at each stage the gambler's capital increases by £1 with probability p and decreases by £1 with probability q. Writing S_n for the gambler's capital after n flips of the coin, we have that $S_0 = 1000$, and $S_0, S_1, \ldots$ is a simple random walk. If the casino refuses to extend credit, then the gambler becomes bankrupt at the first time the random walk visits the point 0, and he may be interested in the probability that he ultimately becomes bankrupt. It turns out that

$$\mathbb{P}\big(S_n = 0 \text{ for some } n \geq 1\big) \begin{cases} < 1 & \text{if } p > \frac{1}{2}, \\ = 1 & \text{if } p \leq \frac{1}{2}, \end{cases} \tag{10.3}$$

so that a compulsive gambler can avoid ultimate bankruptcy (with positive probability) if and only if the odds are stacked in his favour. We shall respect the time-honoured tradition of probability textbooks by returning later to this example. △

In the following exercises, $S_0, S_1, \ldots$ is a random walk on the integers in which p $(= 1 - q)$ is the probability that any given step is to the right.

Exercise 10.4 Find the mean and variance of S_n when $S_0 = 0$.

Exercise 10.5 Find the probability that $S_n = n + k$ given that $S_0 = k$.

10.2 Transition probabilities

Consider a simple random walk starting from the point $S_0 = 0$. For a given time point n and location k, what is the probability that $S_n = k$? The probabilities of such transitions of the random walker are calculated by primitive arguments involving counting, and the following result is typical.

Theorem 10.6 *Let $u_n = \mathbb{P}(S_n = S_0)$ be the probability that a simple random walk revisits its starting point at time n. Then $u_n = 0$ if n is odd, and*

$$u_{2m} = \binom{2m}{m} p^m q^m \tag{10.7}$$

if $n = 2m$ is even.

Proof We may suppose without loss of generality that $S_0 = 0$. If $S_0 = S_n = 0$, then the walk made equal numbers of leftward and rightward steps in its first n steps This is impossible if n is odd, giving that $u_n = 0$ if n is odd. Suppose now that $n = 2m$ for some integer m. From (10.1),

$$S_{2m} = X_1 + X_2 + \cdots + X_{2m},$$

so that $S_{2m} = 0$ if and only if exactly m of the X_i equal $+1$ and exactly m equal -1 (giving m rightward steps and m leftward steps). There are $\binom{2m}{m}$ way of dividing the X_i into two sets

with equal sizes, and the probability that each of the first set equals $+1$ and each of the second set equals -1 is $p^m q^m$. Equation (10.7) follows. □

More general transition probabilities may be calculated similarly. Perhaps the simplest argument proceeds as follows. For $i \geq 1$, the random variable $\frac{1}{2}(X_i + 1)$ has the Bernoulli distribution with parameter p, giving that $B_n = \frac{1}{2}(S_n + n)$ has the binomial distribution with parameters n and p. Hence,

$$\begin{aligned}\mathbb{P}(S_n = k \mid S_0 = 0) &= \mathbb{P}\big(B_n = \tfrac{1}{2}(n+k)\big) \\ &= \binom{n}{\frac{1}{2}(n+k)} p^{\frac{1}{2}(n+k)} q^{\frac{1}{2}(n-k)}. \end{aligned} \tag{10.8}$$

This is non-zero whenever k is such that $\frac{1}{2}(n+k)$ is an integer between 0 and n. The result of Theorem 10.6 is retrieved by setting $k = 0$.

Exercise 10.9 Find $\mathbb{P}(S_{2n+1} = 1 \mid S_0 = 0)$.

Exercise 10.10 Show that $u_n = \mathbb{P}(S_n = S_0)$ satisfies

$$\sum_{n=0}^{\infty} u_n \begin{cases} < \infty & \text{if } p \neq q, \\ = \infty & \text{if } p = q, \end{cases}$$

and deduce that an asymmetric random walk revisits its starting point only finitely often with probability 1. You will need Stirling's formula (see Theorem A.4): $n! \sim (n/e)^n \sqrt{2\pi n}$ as $n \to \infty$.

Exercise 10.11 Consider a two-dimensional random walk in which a particle moves between the points $\{(i, j) : i, j = \ldots, -1, 0, 1, \ldots\}$ with integer coordinates in the plane. Let p, q, r, s be numbers such that $0 < p, q, r, s < 1$ and $p + q + r + s = 1$. If the particle is at position (i, j) at time n, its position at time $n + 1$ is

$$\begin{aligned} &(i+1, j) \text{ with probability } p, \qquad &&(i, j+1) \text{ with probability } q, \\ &(i-1, j) \text{ with probability } r, \qquad &&(i, j-1) \text{ with probability } s, \end{aligned}$$

and successive moves are independent of each other. Writing $\mathbf{S}_n$ for the position of the particle after n moves, we have that

$$\mathbf{S}_{n+1} = \begin{cases} \mathbf{S}_n + (1,0) & \text{with probability } p, \\ \mathbf{S}_n + (0,1) & \text{with probability } q, \\ \mathbf{S}_n - (1,0) & \text{with probability } r, \\ \mathbf{S}_n - (0,1) & \text{with probability } s, \end{cases}$$

and we suppose that $\mathbf{S}_0 = (0, 0)$. Let $v_n = \mathbb{P}(\mathbf{S}_n = (0, 0))$ be the probability that the particle revisits its starting point at time n. Show that $v_n = 0$ if n is odd and

$$v_{2m} = \sum_{k=0}^{m} \frac{(2m)!}{k!^2 (m-k)!^2} (pr)^k (qs)^{m-k}$$

if $n = 2m$ is even.

10.3 Recurrence and transience of random walks

Consider a simple random walk starting from the point $S_0 = 0$. In the subsequent motion, the random walk may or may not revisit its starting point. If the walk is bound (with probability 1) to revisit its starting point, we call it *recurrent*, and otherwise we call it *transient*. We shall see that a simple random walk is recurrent if and only if it is symmetric (in that $p = q = \frac{1}{2}$), and there is a simple intuitive reason why this is the case. The position at time n is the sum $S_n = X_1 + X_2 + \cdots + X_n$ of independent random variables, each having mean value $\mathbb{E}(X) = p - q$ and finite variance, and hence

$$\frac{1}{n} S_n \to p - q \qquad \text{as } n \to \infty$$

in mean square, by the law of large numbers, Theorem 8.6. Thus, if $p > q$, the walk tends to drift linearly towards $+\infty$, whilst if $p < q$, the drift is linear towards $-\infty$. If $p = q$, then $n^{-1} S_n \to 0$ in mean square and the walk remains 'centred' at its starting point 0.

Theorem 10.12 *The probability that a simple random walk ever revisits its starting point is given by*

$$\mathbb{P}\big(S_n = 0 \text{ for some } n = 1, 2, \dots \,\big|\, S_0 = 0\big) = 1 - |p - q|. \tag{10.13}$$

Hence the walk is recurrent if and only if $p = q = \frac{1}{2}$.

Proof We use generating functions in this proof. The basic step is as follows. We suppose that $S_0 = 0$ and we write

$$A_n = \{S_n = 0\}$$

for the event that the walk revisits its starting point at time n, and

$$B_n = \big\{S_n = 0,\ S_k \neq 0 \text{ for } 1 \leq k \leq n-1\big\}$$

for the event that the first return of the walk to its starting point occurs at time n. If A_n occurs, then exactly one of $B_1, B_2, \dots, B_n$ occurs, giving by (1.14) that

$$\mathbb{P}(A_n) = \sum_{k=1}^{n} \mathbb{P}(A_n \cap B_k). \tag{10.14}$$

Now $A_n \cap B_k$ is the event that the walk returns for the first time at time k and then returns again after a subsequent time $n - k$. Hence,

$$\mathbb{P}(A_n \cap B_k) = \mathbb{P}(B_k)\mathbb{P}(A_{n-k}) \qquad \text{for } 1 \leq k \leq n, \tag{10.15}$$

since transitions in disjoint intervals of time are independent of each other. We write $f_n = \mathbb{P}(B_n)$ and $u_n = \mathbb{P}(A_n)$, and substitute (10.15) into (10.14) to obtain

$$u_n = \sum_{k=1}^{n} f_k u_{n-k} \qquad \text{for } n = 1, 2, \dots. \tag{10.16}$$

In this equation, we know the u_i from (10.7) and we want to find the f_k. The form of the summation as a convolution suggests the use of generating functions, and so we introduce the generating functions of the sequences of u_i and f_k,

$$U(s) = \sum_{n=0}^{\infty} u_n s^n, \qquad F(s) = \sum_{n=0}^{\infty} f_n s^n,$$

noting that $u_0 = 1$ and $f_0 = 0$. These sequences converge absolutely if $|s| < 1$, since $|u_n| \leq 1$ and $|f_n| \leq 1$ for each n. We multiply both sides of (10.16) by s^n and sum over n to find

$$\begin{aligned}
\sum_{n=1}^{\infty} u_n s^n &= \sum_{n=1}^{\infty} \sum_{k=1}^{n} f_k u_{n-k} s^n \\
&= \sum_{k=1}^{\infty} \sum_{n=k}^{\infty} f_k s^k u_{n-k} s^{n-k} && \text{by interchanging the order of summation} \\
&= \sum_{k=1}^{\infty} f_k s^k \sum_{m=0}^{\infty} u_m s^m && \text{by setting } m = n - k \\
&= F(s)U(s) && \text{if } |s| < 1. \qquad (10.17)
\end{aligned}$$

The left-hand side of the equation equals $U(s) - u_0 s^0 = U(s) - 1$, and so we have that

$$U(s) = U(s)F(s) + 1 \qquad \text{if } |s| < 1.$$

Hence,

$$F(s) = 1 - \frac{1}{U(s)} \qquad \text{if } |s| < 1. \qquad (10.18)$$

Finally, by Theorem 10.6,

$$\begin{aligned}
U(s) &= \sum_{n=0}^{\infty} u_n s^n \\
&= \sum_{m=0}^{\infty} u_{2m} s^{2m} && \text{since } u_n = 0 \text{ when } n \text{ is odd} \\
&= \sum_{m=0}^{\infty} \binom{2m}{m} (pqs^2)^m \\
&= (1 - 4pqs^2)^{-\frac{1}{2}} && \text{if } |s| < 1,
\end{aligned}$$

by the extended binomial theorem, Theorem A.3. This implies by (10.18) that

$$F(s) = 1 - \sqrt{1 - 4pqs^2} \qquad \text{if } |s| < 1, \qquad (10.19)$$

from which expression the f_k may be found explicitly. To prove the theorem, we note that

$$\begin{aligned}\mathbb{P}\big(S_n = 0 \text{ for some } n \geq 1 \,\big|\, S_0 = 0\big) &= \mathbb{P}(B_1 \cup B_2 \cup \cdots) \\ &= f_1 + f_2 + \cdots && \text{by (1.14)} \\ &= \lim_{s \uparrow 1} \sum_{n=1}^{\infty} f_n s^n && \text{by Abel's lemma}[1] \\ &= F(1) \\ &= 1 - \sqrt{1 - 4pq} && \text{by (10.19).}\end{aligned}$$

Finally, remember that $p + q = 1$ to see that

$$\begin{aligned}\sqrt{1 - 4pq} &= \sqrt{(p+q)^2 - 4pq} \\ &= \sqrt{(p-q)^2} = |p - q|.\end{aligned}$$

□

Thus, if $p = q = \frac{1}{2}$, the walk is bound to return to its starting point. Let

$$T = \min\{n \geq 1 : S_n = 0\}$$

be the (random) time until the first return. We have shown that $\mathbb{P}(T < \infty) = 1$ if $p = q = \frac{1}{2}$. Against this positive observation we must set the following negative one: although T is finite (with probability 1), it has infinite mean in that $\mathbb{E}(T) = \infty$. To see this, just note that

$$\begin{aligned}\mathbb{E}(T) &= \sum_{n=1}^{\infty} n f_n \\ &= \lim_{s \uparrow 1} \sum_{n=1}^{\infty} n f_n s^{n-1} && \text{by Abel's lemma} \\ &= \lim_{s \uparrow 1} F'(s).\end{aligned}$$

From (10.19), if $p = q = \frac{1}{2}$, then

$$F(s) = 1 - \sqrt{1 - s^2} \qquad \text{for } |s| < 1$$

and

$$F'(s) = \frac{s}{\sqrt{1 - s^2}} \to \infty \qquad \text{as } s \uparrow 1.$$

In other words, although a symmetric random walk is certain to return to its starting point, the expected value of the time which elapses before this happens is infinite.

[1] For a statement of Abel's lemma, see the footnote on p. 55.

Exercise 10.20 Consider a simple random walk with $p \neq q$. Show that, conditional on the walk returning to its starting point at some time, the expected number of steps taken before this occurs is

$$\frac{4pq}{|p-q|(1-|p-q|)}.$$

Exercise 10.21 Show that a symmetric random walk revisits its starting point infinitely often with probability 1.

Exercise 10.22 Show that a symmetric random walk starting from the origin visits the point 1 with probability 1.

10.4 The Gambler's Ruin Problem

The Gambler's Ruin Problem is an old chestnut of probability textbooks. It concerns a game between two players, A and B, say, who compete with each other as follows. Initially A possesses £a and B possesses £$(N-a)$, so that their total capital is £N, where $N \geq 1$. A coin is flipped repeatedly and comes up either heads with probability p or tails with probability q, where $0 < p = 1 - q < 1$. Each time heads turns up, player B gives £1 to player A, while each time tails turns up, player A gives £1 to player B. This game continues until either A or B runs out of money. We record the state of play by noting A's capital after each flip. Clearly, the sequence of such numbers is a random walk on the set $\{0, 1, \ldots, N\}$. This walk starts at the point a and follows a simple random walk until it reaches either 0 or N, at which time it stops. We speak of 0 and N as being *absorbing barriers* since the random walker sticks to whichever of these points it hits first. We shall say that A *wins the game* if the random walker is absorbed at N, and that B *wins the game* if the walker is absorbed at 0. It is fairly clear (see Exercises 10.41–10.42) that there is zero probability that the game will continue forever, so that either A or B (but not both) will win the game. What is the probability that A wins the game? The answer is given in the following theorem.

Theorem 10.23 (Gambler's Ruin) *Consider a simple random walk on* $\{0, 1, \ldots, N\}$ *with absorbing barriers at* 0 *and* N. *If the walk begins at the point* a, *where* $0 \leq a \leq N$, *then the probability* $v(a)$ *that the walk is absorbed at* N *is given by*

$$v(a) = \begin{cases} \dfrac{(q/p)^a - 1}{(q/p)^N - 1} & \text{if } p \neq q, \\ a/N & \text{if } p = q = \frac{1}{2}. \end{cases} \tag{10.24}$$

Thus, the probability that player A wins the game is given by equation (10.24). Our proof of this theorem uses the jargon of the Gambler's Ruin Problem.

Proof The first step of the argument is simple, and provides a difference equation for the numbers $v(0), v(1), \ldots, v(N)$. Let H be the event that the first flip of the coin shows heads. We use the partition theorem, Theorem 1.48, to see that

$$\mathbb{P}(\text{A wins}) = \mathbb{P}(\text{A wins} \mid H)\mathbb{P}(H) + \mathbb{P}(\text{A wins} \mid H^c)\mathbb{P}(H^c), \tag{10.25}$$

where, as usual, H^c is the complement of H. If H occurs, A's capital increases from £a to £$(a+1)$, giving that $\mathbb{P}(\text{A wins} \mid H) = v(a+1)$. Similarly, $\mathbb{P}(\text{A wins} \mid H^c) = v(a-1)$. We substitute these expressions into (10.25) to obtain

$$v(a) = pv(a+1) + qv(a-1) \qquad \text{for } 1 \leq a \leq N-1. \tag{10.26}$$

This is a second-order difference equation subject to the boundary conditions

$$v(0) = 0, \qquad v(N) = 1, \tag{10.27}$$

since if A starts with £0 (or £N), he or she has already lost (or won) the game. We solve (10.26) by the standard methods described in Appendix B, obtaining as general solution

$$v(a) = \begin{cases} \alpha + \beta(q/p)^a & \text{if } p \neq q, \\ \alpha + \beta a & \text{if } p = q = \frac{1}{2}, \end{cases}$$

where α and β are constants which are found from the boundary conditions (10.27) as required. □

There is another standard calculation which involves difference equations and arises from the Gambler's Ruin Problem. This deals with the expected length of the game. Once again, we formulate this in terms of the related random walk.

Theorem 10.28 (Recurrence/transience of random walk) *Consider a simple random walk on* $\{0, 1, \ldots, N\}$ *with absorbing barriers at* 0 *and* N. *If the walk begins at the point* a, *where* $0 \leq a \leq N$, *then the expected number* $e(a)$ *of steps before the walk is absorbed at either* 0 *or* N *is given by*

$$e(a) = \begin{cases} \dfrac{1}{p-q}\left(N\dfrac{(q/p)^a - 1}{(q/p)^N - 1} - a\right) & \text{if } p \neq q, \\ a(N-a) & \text{if } p = q = \frac{1}{2}. \end{cases} \tag{10.29}$$

Thus, the expected number of flips of the coin before either A or B becomes bankrupt in the Gambler's Ruin Problem is given by (10.29).

Proof Let F be the number of flips of the coin before the game ends, and let H be the event that the first flip shows heads as before. By the partition theorem, Theorem 2.42, we have that

$$\mathbb{E}(F) = \mathbb{E}(F \mid H)\mathbb{P}(H) + \mathbb{E}(F \mid H^c)\mathbb{P}(H^c). \tag{10.30}$$

Now, if H occurs then, after the first flip of the coin, A's capital increases from £a to £$(a+1)$, giving that $\mathbb{E}(F \mid H) = 1 + e(a+1)$, where the 1 arises from the first flip, and $e(a+1)$ is the mean number of subsequent flips. Similarly, $\mathbb{E}(F \mid H^c) = 1 + e(a-1)$, and (10.30) becomes

$$e(a) = \big[1 + e(a+1)\big]p + \big[1 + e(a-1)\big]q$$

or

$$pe(a+1) - e(a) + qe(a-1) = -1 \qquad \text{for } 1 \leq a \leq N-1. \tag{10.31}$$

The boundary conditions for this second-order difference equation are

$$e(0) = e(N) = 0,$$

since the game is finished already if it starts in locations 0 or N. We solve (10.31) in the standard manner (as shown in Appendix B) to obtain (10.29). □

Finally, what are A's fortunes if the opponent is infinitely rich? In practice, this situation cannot arise, but the hypothetical situation may help us to understand the consequences of a visit to the casino at Monte Carlo. In this case, A can never defeat the opponent, but A might at least hope to be spared ultimate bankruptcy in order to play the game forever. The probability that A is ultimately bankrupted is given by our final theorem about random walks in one dimension.

Theorem 10.32 *Consider a simple random walk on* $\{0, 1, 2, \dots\}$ *with an absorbing barrier at* 0*. If the walk begins at the point* a (≥ 0)*, the probability* $\pi(a)$ *that the walk is ultimately absorbed at* 0 *is given by*

$$\pi(a) = \begin{cases} (q/p)^a & \text{if } p > q, \\ 1 & \text{if } p \leq q. \end{cases} \tag{10.33}$$

Thus, the probability that player A is able to play forever is strictly positive if and only if the odds are stacked in his or her favour at each flip of the coin. This justifies equation (10.3). An intuitive approach to this theorem is to think of this new game as the limit of the previous game as the total capital N tends to infinity while A's initial capital remains fixed at a. Thus,

$$\mathbb{P}(\text{A is bankrupted}) = \lim_{N\to\infty} [1 - v(a)],$$

where $v(a)$ is given by (10.24), and it is easy to see that the value of this limit is given by (10.33). There is a limiting argument here which requires some justification, but we shall use a different approach which avoids this.

Proof The sequence $\pi(0), \pi(1), \dots$ satisfies the difference equation

$$\pi(a) = p\pi(a+1) + q\pi(a-1) \qquad \text{for } a = 1, 2, \dots, \tag{10.34}$$

derived in exactly the same way as (10.26). The general solution is

$$\pi(a) = \begin{cases} \alpha + \beta(q/p)^a & \text{if } p \neq q, \\ \alpha + \beta a & \text{if } p = q = \frac{1}{2}, \end{cases} \tag{10.35}$$

where α and β are constants. Unfortunately, we have only one boundary condition, namely $\pi(0) = 1$. Using this condition we find that

$$\begin{aligned} \alpha + \beta &= 1 \qquad \text{if } p \neq q, \\ \alpha &= 1 \qquad \text{if } p = q, \end{aligned}$$

and hence,

$$\pi(a) = \begin{cases} \beta(q/p)^a + 1 - \beta & \text{if } p \neq q, \\ 1 + \beta a & \text{if } p = q = \frac{1}{2}, \end{cases} \tag{10.36}$$

for some $\beta \in \mathbb{R}$. Now, $\pi(a)$ is a probability, and so $0 \leq \pi(a) \leq 1$ for all a. Hence, if $p = q$, then $\beta = 0$, giving $\pi(a) = 1$ for all a. On the other hand, if $p < q$, then $(q/p)^a \to \infty$ as $a \to \infty$. Thus, $\beta = 0$ if $p < q$, and we have proved that

$$\pi(a) = 1 \quad \text{if } p \leq q, \qquad \text{for } a = 0, 1, 2, \ldots. \tag{10.37}$$

It is not quite so easy to find the correct value of β in (10.36) for the case when $p > q$, and we shall do this by calculating $\pi(1)$ explicitly for this case. For the remaining part of the proof we write $\pi(a) = \pi_{p,q}(a)$ in order to emphasize the roles of p and q; thus, (10.37) reads

$$\pi_{p,q}(a) = 1 \quad \text{if } p \leq q, \qquad \text{for } a = 0, 1, 2, \ldots. \tag{10.38}$$

Consider a simple random walk $T_0, T_1, \ldots$ starting from $T_0 = 0$ in which each step is to the right with probability p or to the left with probability q, and let C be the event that the walk revisits its starting point: $C = \{T_n = 0 \text{ for some } n \geq 1\}$. From Theorem 10.12,

$$\mathbb{P}(C) = 1 - |p - q|. \tag{10.39}$$

On the other hand, the usual conditioning argument yields

$$\mathbb{P}(C) = \mathbb{P}(C \mid H)\mathbb{P}(H) + \mathbb{P}(C \mid H^c)\mathbb{P}(H^c),$$

where H is the event that the first move is to the right. Now, $\mathbb{P}(C \mid H) = \pi_{p,q}(1)$, since this is the probability that a walk ever reaches 0 having started from 1. Also, $\mathbb{P}(C \mid H^c) = \pi_{q,p}(1)$, since this is the probability that the 'mirror image' walk (in which each step of T is replaced by an opposite step) reaches 0 starting from 1. Thus,

$$\mathbb{P}(C) = p\pi_{p,q}(1) + q\pi_{q,p}(1),$$

which combines with (10.39) to give

$$1 - |p - q| = p\pi_{p,q}(1) + q\pi_{p,q}(1). \tag{10.40}$$

If $p \geq q$, then $\pi_{q,p}(1) = 1$ by (10.38), and (10.40) becomes

$$1 - (p - q) = p\pi_{p,q}(1) + q,$$

implying that

$$\pi_{p,q}(1) = q/p \qquad \text{if } p > q.$$

Substitute this into (10.36) to find that $\beta = 1$ if $p > q$, and the theorem is proved. □

Exercise 10.41 Show that, in the Gambler's Ruin Problem, the game terminates with probability 1. You may find it useful to partition the sequence of coin tosses into disjoint runs of length N, and to consider the event that one of these runs contains only steps to the right.

Exercise 10.42 Use Theorem 10.23 to show that, in the Gambler's Ruin Problem, the probability that B wins the game is

$$\mu(a) = \begin{cases} \dfrac{(p/q)^{N-a} - 1}{(p/q)^N - 1} & \text{if } p \neq q, \\ (N-a)/N & \text{if } p = q = \frac{1}{2}, \end{cases}$$

where $£(N-a)$ is B's initial capital. Deduce that

$$\mathbb{P}(\text{A wins}) + \mathbb{P}(\text{B wins}) = 1,$$

and hence that there is zero probability that the game will fail to terminate.

Exercise 10.43 Consider the Gambler's Ruin Problem with the difference that A's initial capital is uniformly distributed on the set $\{0, 1, 2, \ldots, N\}$. What is the probability A wins the game?

Exercise 10.44 Consider the Gambler's Ruin Problem with $0 < p < 1$ and $p \neq \frac{1}{2}$. Suppose A's initial capital is $£k$, where $1 \leq k < N$, and we are given that A wins the game. What is the probability that the first step was from k to $k-1$?

10.5 Problems

1. Two particles perform independent and simultaneous symmetric random walks starting from the origin. Show that the probability that they are at the same position after n steps is

$$\left(\frac{1}{2}\right)^{2n} \sum_{k=0}^{n} \binom{n}{k}^2.$$

   Hence or otherwise show that

$$\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}.$$

2. Consider a random walk on the integers with absorbing barriers at 0 and N in which, at each stage, the particle may jump one unit to the left (with probability α), remain where it is (with probability β), or jump one unit to the right (with probability γ), where $\alpha, \beta, \gamma > 0$ and $\alpha + \beta + \gamma = 1$. If the particle starts from the point a, where $0 \leq a \leq N$, show that the probability that it is absorbed at N is given by equation (10.24) with $p = 1 - q = \gamma/(\alpha + \gamma)$. Find the mean number of stages before the particle is absorbed at one or other of the barriers.

3. A particle performs a random walk on the set $\{-N, -N+1, \ldots, N-1, N\}$ and is absorbed if it reaches $-N$ or N, where $N > 1$. The probability of a step of size -1 is $q = 1 - p$, with $0 < p < 1$. Suppose that the particle starts at 0. By conditioning on the first step and using Theorem 10.23, or otherwise, show that when $p \neq q$, the probability of the particle being absorbed at N or $-N$ before returning to 0 is

$$\frac{(p-q)(p^N + q^N)}{p^N - q^N}.$$

   What is this probability when $p = q$? (Oxford 1983M)

4. Consider a random walk on the integers in which the particle moves either two units to the right (with probability p) or one unit to the left (with probability $q = 1 - p$) at each stage, where $0 < p < 1$. There is an absorbing barrier at 0 and the particle starts at the point a (> 0). Show that the probability $\pi(a)$ that the particle is ultimately absorbed at 0 satisfies the difference equation

$$\pi(a) = p\pi(a+2) + q\pi(a-1) \qquad \text{for } a = 1, 2, \ldots,$$

and deduce that $\pi(a) = 1$ if $p \leq \frac{1}{3}$.

Suppose that the particle is absorbed whenever it hits either N or $N+1$. Find the probability $\pi_N(a)$ that it is absorbed at 0 rather than at N or $N+1$, having started at a, where $0 \leq a \leq N+1$. Deduce that, as $N \to \infty$,

$$\pi_N(a) \to \begin{cases} 1 & \text{if } p \leq \frac{1}{3}, \\ \theta^a & \text{if } p > \frac{1}{3}, \end{cases}$$

where $\theta = \frac{1}{2}\{\sqrt{1 + (4q/p)} - 1\}$.

Deduce that if a fair coin is tossed repeatedly, the probability that the number of heads ever exceeds twice the number of tails is $\frac{1}{2}(\sqrt{5} - 1)$.

5. Consider a simple random walk with an absorbing barrier at 0 and a 'retaining' barrier at N. That is to say, the walk is not allowed to pass to the right of N, so that its position S_n at time n satisfies

$$\mathbb{P}(S_{n+1} = N \mid S_n = N) = p, \qquad \mathbb{P}(S_{n+1} = N - 1 \mid S_n = N) = q,$$

where $p + q = 1$. Set up a difference equation for the mean number $e(a)$ of jumps of the walk until absorption at 0, starting from a, where $0 \leq a \leq N$. Deduce that

$$e(a) = a(2N - a + 1) \qquad \text{if } p = q = \tfrac{1}{2},$$

and find $e(a)$ if $p \neq q$.

6. Let N be the number of times that an asymmetric simple random walk revisits its starting point. Show that N has mass function

$$\mathbb{P}(N = k) = \alpha(1 - \alpha)^k \qquad \text{for } k = 0, 1, 2, \ldots,$$

where $\alpha = |p - q|$ and p is the probability that each step of the walk is to the right.

7. A slot machine functions as follows. At the first pull, the player wins with probability $\frac{1}{2}$. At later pulls, the player wins with probability $\frac{1}{2}$ if the previous pull was lost, and with probability p ($< \frac{1}{2}$) if won. Show that the probability u_n that the player wins at the nth pull satisfies

$$u_n + (\tfrac{1}{2} - p)u_{n-1} = \tfrac{1}{2} \qquad \text{for } n > 1.$$

Deduce that

$$u_n = \frac{1 + (-1)^{n-1}(\frac{1}{2} - p)^n}{3 - 2p} \qquad \text{for } n \geq 1.$$

8. Consider the two-dimensional random walk of Exercise 10.11, in which a particle inhabits the integer points $\{(i, j) : i, j = \ldots, -1, 0, 1, \ldots\}$ of the plane, moving rightwards, upwards, leftwards or downwards with respective probabilities p, q, r, and s at each step, where

$p, q, r, s > 0$ and $p+q+r+s = 1$. Let $\mathbf{S}_n$ be the particle's position after n steps, and suppose that $\mathbf{S}_0 = (0, 0)$. Let v_n be the probability that $\mathbf{S}_n = (0, 0)$, and prove that

$$v_{2m} = \binom{2m}{m}^2 \left(\frac{1}{4}\right)^{2m} \qquad \text{if } p = q = r = s = \tfrac{1}{4}.$$

Use Stirling's formula to show that

$$\sum_{n=0}^{\infty} v_n = \infty \qquad \text{if } p = q = r = s = \tfrac{1}{4}.$$

Deduce directly that the symmetric random walk in two dimensions is recurrent.

9. Here is another way of approaching the symmetric random walk in two dimensions of Problem 10.5.8. Make the following change of variables. If $\mathbf{S}_n = (i, j)$, set $X_n = i + j$ and $Y_n = i - j$; this is equivalent to rotating the axes through an angle of $\frac{1}{4}\pi$. Show that $X_0, X_1, \ldots$ and $Y_0, Y_1, \ldots$ are independent symmetric random walks in one dimension. Deduce by Theorem 10.6 that

$$\mathbb{P}\big(\mathbf{S}_{2m} = \mathbf{0} \,\big|\, \mathbf{S}_0 = \mathbf{0}\big) = \left[\binom{2m}{m}\left(\frac{1}{2}\right)^{2m}\right]^2,$$

where $\mathbf{0} = (0, 0)$.

10. In the two-dimensional random walk of Problem 10.5.8, let D_n be the Euclidean distance between the origin and $\mathbf{S}_n$. Prove that, if the walk is symmetric,

$$\mathbb{E}(D_n^2) = \mathbb{E}(D_{n-1}^2) + 1 \qquad \text{for } n = 1, 2, \ldots,$$

and deduce that $\mathbb{E}(D_n^2) = n$.

* 11. A particle performs a random walk on the integers starting at the origin. At discrete intervals of time, it takes a step of unit size. The steps are independent and equally likely to be in the positive or negative direction. Determine the probability generating function of the time at which the particle first reaches the integer n (≥ 1).

In a two-dimensional random walk, a particle can be at any of the points (x, y) which have integer coordinates. The particle starts at $(0, 0)$ and at discrete intervals of time, takes a step of unit size. The steps are independent and equally likely to be any of the four nearest points. Show that the probability generating function of the time taken to reach the line $x + y = m$ is

$$\left\{\frac{1 - \sqrt{1 - s^2}}{s}\right\}^m \qquad \text{for } |s| \leq 1.$$

Let (X, Y) be the random point on the line $x + y = m$ which is reached first. What is the probability generating function of $X - Y$? (Oxford 1979F)

12. Consider a symmetric random walk on the integer points of the cubic lattice $\{(i, j, k) : i, j, k = \ldots, -1, 0, 1, \ldots\}$ in three dimensions, in which the particle moves to one of its six neighbouring positions, chosen with equal probability $\frac{1}{6}$, at each stage. Show that the probability w_n that the particle revisits its starting point at the nth stage is given by

$$w_{2m+1} = 0,$$
$$w_{2m} = \left(\frac{1}{2}\right)^{2m} \binom{2m}{m} \sum_{\substack{(i,j,k):\\ i+j+k=m}} \left(\frac{m!}{3^m i!\, j!\, k!}\right)^2.$$

Use Stirling's formula to show that

$$\sum_{n=0}^{\infty} w_n < \infty.$$

Deduce that the symmetric random walk in three dimensions is transient (the general argument in Problem 10.5.6 may be useful here).

13. Show that the symmetric random walk on the integer points

$$\mathbb{Z}^d = \{(i_1, i_2, \ldots, i_d) : i_j = \ldots, -1, 0, 1, \ldots, j = 1, 2, \ldots, d\}$$

is recurrent if $d = 1, 2$ and transient if $d \geq 3$. You should use the results of Problems 10.5.8 and 10.5.12, and you need do no more calculations.

14. The generating-function argument used to prove Theorem 10.12 has a powerful application to the general theory of 'recurrent events'. Let η be an event which may or may not happen at each of the time points $0, 1, 2, \ldots$ (η may be the visit of a random walk to its starting point, or a visit to the dentist, or a car accident outside the department). We suppose that η occurs at time 0. Suppose further that the intervals between successive occurrences of η are independent, identically distributed random variables $X_1, X_2, \ldots$, each having mass function

$$\mathbb{P}(X = k) = f_k \qquad \text{for } k = 1, 2, \ldots,$$

so that η occurs at the times $0, X_1, X_1 + X_2, X_1 + X_2 + X_3, \ldots$. There may exist a time after which η never occurs. That is to say, there may be an X_i which takes the value ∞, and we allow for this by requiring only that $f = f_1 + f_2 + \cdots$ satisfies $f \leq 1$, and we set

$$\mathbb{P}(X = \infty) = 1 - f.$$

We call η *recurrent* if $f = 1$ and *transient* if $f < 1$. Let u_n be the probability that η occurs at time n. Show that the generating functions

$$F(s) = \sum_{n=1}^{\infty} f_n s^n, \qquad U(s) = \sum_{n=0}^{\infty} u_n s^n$$

are related by

$$U(s) = U(s)F(s) + 1,$$

and deduce that η is recurrent if and only if $\sum_n u_n = \infty$.

15. The university buys light bulbs which have random lifetimes. If the bulb in my office fails on day n, then it is replaced by a new bulb which lasts for a random number of days, after which it is changed, and so on. We assume that the lifetimes of the bulbs are independent random variables $X_1, X_2, \ldots$ each having mass function

$$\mathbb{P}(X = k) = (1 - \alpha)\alpha^{k-1} \qquad \text{for } k = 1, 2, \ldots,$$

where α satisfies $0 < \alpha < 1$. A new light bulb is inserted in its socket on day 0. Show that the probability that the bulb has to be changed on day n is $1 - \alpha$, independently of n.

11
Random processes in continuous time

Summary. This account of random processes in continuous-time is centred on the Poisson process. The relationship between the Poisson process and the exponential distribution is exposed, via the lack-of-memory property. The simple birth and birth–death processes are described, followed by an introduction to queueing theory.

11.1 Life at a telephone switchboard

Branching processes and random walks are two examples of random processes. Each is a random sequence, and we call them *discrete-time* processes since they involve observations at the discrete times $n = 0, 1, 2 \ldots$. Many other processes involve observations which are made continuously as time passes, and such processes are called *continuous-time* processes. Rather than being a family $(Z_n : n = 0, 1, 2 \ldots)$ of random variables indexed by the non-negative integers, a continuous-time process is a family $Z = (Z_t : t \geq 0)$ of random variables indexed by the continuum $[0, \infty)$, where we think of Z_t as being the value of the process at time t. The general theory of continuous-time processes is rather deep and quite difficult, but most of the main difficulties are avoided if we restrict our attention to processes which take integer values only, that is, processes for which Z_t is a (random) integer for each t, and all our examples are of this form. The principal difference between studying such continuous-time processes and studying discrete-time processes is merely that which arises in moving from the integers to the reals: instead of establishing *recurrence* equations and *difference* equations (for example, (9.5) and (10.26)), we shall establish *differential* equations.

Here is our basic example. Bill is the head porter at the Grand Hotel, and part of his job is to answer incoming telephone calls. He cannot predict with certainty when the telephone will ring; from his point of view, calls seem to arrive at random. We make two simplifying assumptions about these calls. First, we assume that Bill deals with every call instantaneously, so that no call is lost unless it arrives at exactly the same moment as another (in practice, Bill has to get to the telephone and speak, and this takes time—more complicated models take account of this). Secondly, we assume that calls arrive 'homogeneously' in time, in the sense that the chance that the telephone rings during any given period of time depends only upon the length of this period (this is an absurd assumption of course, but it may be valid for certain portions of the day). We describe *time* by a parameter t taking values in $[0, \infty)$, and propose the following model for the arrivals of telephone calls at the switchboard. We let N_t represent the number of calls which have arrived in the time interval $[0, t]$: that is, N_t is the number of incoming calls which Bill has handled up to and including time t. We suppose that the random process $N = (N_t : t \geq 0)$ evolves in such a way that the following conditions are valid:

A. N_t is a random variable taking values in $\{0, 1, 2, \dots\}$,
B. $N_0 = 0$,
C. $N_s \le N_t$ if $s \le t$,
D. *independence*: if $0 \le s < t$ then the number of calls which arrive during the time interval $(s, t]$ is independent of the arrivals of calls prior to time s,
E. *arrival rate*: there exists a number λ (> 0), called the *arrival rate*, such that,[1] for small positive h,

$$\begin{aligned} \mathbb{P}\big(N_{t+h} = n+1 \,\big|\, N_t = n\big) &= \lambda h + \mathrm{o}(h), \\ \mathbb{P}\big(N_{t+h} = n \,\big|\, N_t = n\big) &= 1 - \lambda h + \mathrm{o}(h). \end{aligned} \tag{11.1}$$

Condition E merits a discussion. It postulates that the probability that a call arrives in some short interval $(t, t+h]$ is approximately a linear function of h, and that this approximation becomes better and better as h becomes smaller and smaller. It follows from (11.1) that the chance of two or more calls in the interval $(t, t+h]$ satisfies

$$\begin{aligned} \mathbb{P}\big(N_{t+h} \ge n+2 \,\big|\, N_t = n\big) &= 1 - \mathbb{P}\big(N_{t+h} \text{ equals } n \text{ or } n+1 \,\big|\, N_t = n\big) \\ &= 1 - [\lambda h + \mathrm{o}(h)] - [1 - \lambda h + \mathrm{o}(h)] \\ &= \mathrm{o}(h), \end{aligned}$$

so that the only two possible events having significant probabilities (that is, with probability greater than $\mathrm{o}(h)$) involve either *no* call arriving in $(t, t+h]$ or exactly *one* call arriving in this time interval.

This is our model for the arrival of telephone calls. It is a primitive model based on the idea of random arrivals, and obtained with the aid of various simplifying assumptions. For a reason which will soon be clear, this random process $N = (N_t : t \ge 0)$ is called a *Poisson process with rate* λ. Poisson processes may be used to model many phenomena, such as

(a) the arrival of customers in a shop,
(b) the clicks emitted by a Geiger counter as it records the detection of radioactive particles,
(c) the incidence of deaths in a small town with a reasonably stable population (neglecting seasonal variations).

The Poisson process provides an exceptionally good model for the emission of radioactive particles when the source has a long half-life and is decaying slowly.

We may represent the outcomes of a Poisson process N by a graph of N_t against t (see Figure 11.1). Let T_i be the time at which the ith call arrives, so that

$$T_i = \inf\{t : N_t = i\}. \tag{11.2}$$

Then $T_0 = 0$, $T_0 \le T_1 \le T_2 \le \cdots$, and $N_t = i$ if t lies in the interval $[T_i, T_{i+1})$. We note that $T_0, T_1, T_2, \dots$ is a sequence of random variables whose values determine the process N completely: if we know the T_i, then N_t is given by

$$N_t = \max\{n : T_n \le t\}.$$

The sequence of T_i may be thought of as the 'inverse process' of N.

[1] Recall Landau's notation from p. 127: $\mathrm{o}(h)$ denotes some function of h which is of smaller order of magnitude than h as $h \to 0$. More precisely, we write $f(h) = \mathrm{o}(h)$ if $f(h)/h \to 0$ as $h \to 0$. The term $\mathrm{o}(h)$ generally represents different functions of h at each appearance. Thus, for example, $\mathrm{o}(h) + \mathrm{o}(h) = \mathrm{o}(h)$.

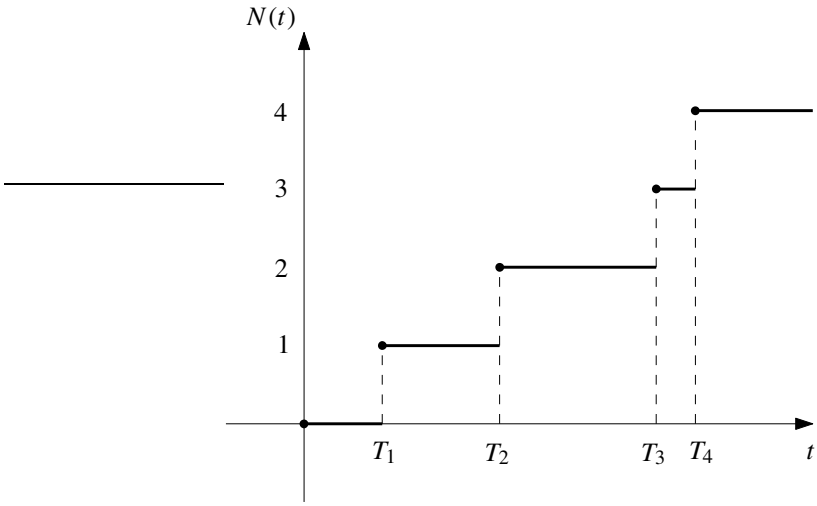


Fig. 11.1 A sketch of the graph of a Poisson process.

Conditions A–E are our postulates for a Poisson process N. In the next two sections, we present some consequences of these postulates, answering such questions as

(a) what is the mass function of N_t for a given value of t?
(b) what can be said about the distribution of the sequence $T_0, T_1, \ldots$ of times at which calls arrive?

Exercise 11.3 If N is a Poisson process with rate λ, show that

$$\mathbb{P}(N_{t+h} = 0) = \big[1 - \lambda h + \mathrm{o}(h)\big]\mathbb{P}(N_t = 0)$$

for small positive values of h. Hence, show that $p(t) = \mathbb{P}(N_t = 0)$ satisfies the differential equation

$$p'(t) = -\lambda p(t).$$

Solve this equation to find $p(t)$.

Exercise 11.4 (Thinning) Suppose that telephone calls arrive at the exchange in the manner of a Poisson process $N = (N_t : t \geq 0)$ with rate λ, and suppose that the equipment is faulty so that each incoming call fails to be recorded with probability q (independently of all other calls). If N'_t is the number of calls recorded by time t, show that $N' = (N'_t : t \geq 0)$ is a Poisson process with rate $\lambda(1-q)$.

Exercise 11.5 (Superposition) Two independent streams of telephone calls arrive at the exchange, the first being a Poisson process with rate λ and the second being a Poisson process with rate μ. Show that the combined stream of calls is a Poisson process with rate $\lambda + \mu$.

11.2 Poisson processes

A *Poisson process* with rate λ is a random process which satisfies postulates A–E of Section 11.1. Our first result establishes the connection to the Poisson distribution.

Theorem 11.6 *For each $t > 0$, the random variable N_t has the Poisson distribution with parameter λt. That is, for $t > 0$,*

$$\mathbb{P}(N_t = k) = \frac{1}{k!}(\lambda t)^k e^{-\lambda t} \qquad \text{for } k = 0, 1, 2, \ldots. \tag{11.7}$$

It follows from (11.7) that the mean and variance of N_t grow linearly in t as t increases:

$$\mathbb{E}(N_t) = \lambda t, \quad \operatorname{var}(N_t) = \lambda t \qquad \text{for } t > 0. \tag{11.8}$$

Proof Just as we set up difference equations for discrete-time processes, here we set up 'differential–difference' equations. Let

$$p_k(t) = \mathbb{P}(N_t = k).$$

Fix $t \geq 0$ and let h be small and positive. The basic step is to express N_{t+h} in terms of N_t as follows. We use the partition theorem, Theorem 1.48, to see that, if $k \geq 1$,

$$\begin{aligned}
\mathbb{P}(N_{t+h} = k) &= \sum_{i=0}^{k} \mathbb{P}(N_{t+h} = k \mid N_t = i)\mathbb{P}(N_t = i) \\
&= \mathbb{P}(N_{t+h} = k \mid N_t = k-1)\mathbb{P}(N_t = k-1) \\
&\qquad + \mathbb{P}(N_{t+h} = k \mid N_t = k)\mathbb{P}(N_t = k) + \mathrm{o}(h) \qquad \text{by (11.1)} \\
&= [\lambda h + \mathrm{o}(h)]\mathbb{P}(N_t = k-1) + [1 - \lambda h + \mathrm{o}(h)]\mathbb{P}(N_t = k) + \mathrm{o}(h) \qquad \text{by (11.1)} \\
&= \lambda h \mathbb{P}(N_t = k-1) + (1 - \lambda h)\mathbb{P}(N_t = k) + \mathrm{o}(h), \qquad (11.9)
\end{aligned}$$

giving that

$$p_k(t+h) - p_k(t) = \lambda h\big[p_{k-1}(t) - p_k(t)\big] + \mathrm{o}(h), \tag{11.10}$$

valid for $k = 1, 2, \ldots$. We divide both sides of (11.10) by h and take the limit as $h \downarrow 0$ to obtain

$$p_k'(t) = \lambda p_{k-1}(t) - \lambda p_k(t) \qquad \text{for } k = 1, 2, \ldots, \tag{11.11}$$

where $p_k'(t)$ is the derivative of $p_k(t)$ with respect to t. When $k = 0$, (11.9) becomes

$$\begin{aligned}
\mathbb{P}(N_{t+h} = 0) &= \mathbb{P}(N_{t+h} = 0 \mid N_t = 0)\mathbb{P}(N_t = 0) \\
&= (1 - \lambda h)\mathbb{P}(N_t = 0) + \mathrm{o}(h),
\end{aligned}$$

giving in the same way

$$p_0'(t) = -\lambda p_0(t). \tag{11.12}$$

Equations (11.11) and (11.12) are a system of 'differential–difference' equations for the functions $p_0(t)$, $p_1(t)$, ..., and we wish to solve them subject to the boundary condition $N_0 = 0$, which is equivalent to the condition

$$p_k(0) = \begin{cases} 1 & \text{if } k = 0, \\ 0 & \text{if } k \neq 0. \end{cases} \tag{11.13}$$

We present two ways of solving this family of equations.

Solution A (by induction) Equation (11.12) involves $p_0(t)$ alone. Its general solution is $p_0(t) = Ae^{-\lambda t}$, and the arbitrary constant A is found from (11.13) to equal 1. Hence,

$$p_0(t) = e^{-\lambda t} \qquad \text{for } t \geq 0. \tag{11.14}$$

Substitute this into (11.11) with $n = 1$ to obtain

$$p_1'(t) + \lambda p_1(t) = \lambda e^{-\lambda t}$$

which, with the aid of an integrating factor and the boundary condition, yields

$$p_1(t) = \lambda t e^{-\lambda t} \qquad \text{for } t \geq 0. \tag{11.15}$$

Continue in this way to find that

$$p_2(t) = \tfrac{1}{2}(\lambda t)^2 e^{-\lambda t}.$$

Now guess the general solution (11.7) and prove it from (11.11) by induction.

Solution B (by generating functions) This method is nicer and has further applications. We use the probability generating function of N_t, namely

$$G(s,t) = \mathbb{E}(s^{N_t}) = \sum_{k=0}^{\infty} p_k(t) s^k.$$

We multiply both sides of (11.11) by s^k and sum over the values $k = 1, 2, \ldots$ to find that

$$\sum_{k=1}^{\infty} p_k'(t) s^k = \lambda \sum_{k=1}^{\infty} p_{k-1}(t) s^k - \lambda \sum_{k=1}^{\infty} p_k(t) s^k.$$

Add (11.12) to this in the obvious way and note that

$$\sum_{k=1}^{\infty} p_{k-1}(t) s^k = sG(s,t)$$

and (plus or minus a dash of mathematical rigour)

$$\sum_{k=0}^{\infty} p_k'(t) s^k = \frac{\partial G}{\partial t},$$

to obtain

$$\frac{\partial G}{\partial t} = \lambda s G - \lambda G, \tag{11.16}$$

a differential equation subject to the boundary condition

$$G(s,0) = \sum_{k=0}^{\infty} p_k(0) s^k = 1 \qquad \text{by (11.13).} \tag{11.17}$$

Equation (11.16) may be written in the form

$$\frac{1}{G}\frac{\partial G}{\partial t} = \lambda(s-1).$$

This resembles a partial differential equation, but for each given value of s it may be integrated in the usual manner with respect to t, giving that

$$\log G = \lambda t(s-1) + A(s),$$

where $A(s)$ is an arbitrary function of s. Use (11.17) to find that $A(s) = 0$ for all s, and hence

$$G(s,t) = e^{-\lambda t(s-1)} = \sum_{k=0}^{\infty} \left(\frac{1}{k!}(\lambda t)^k e^{-\lambda t} \right) s^k.$$

Reading off the coefficient of s^k, we have that

$$p_k(t) = \frac{1}{k!}(\lambda t)^k e^{-\lambda t}$$

as required. □

Exercise 11.18 If N is a Poisson process with rate λ, show that $\operatorname{var}(N_t/t) \to 0$ as $t \to \infty$.

Exercise 11.19 If N is a Poisson process with rate λ, show that, for $t > 0$,

$$\mathbb{P}(N_t \text{ is even}) = e^{-\lambda t}\cosh \lambda t,$$
$$\mathbb{P}(N_t \text{ is odd}) = e^{-\lambda t}\sinh \lambda t.$$

Exercise 11.20 If N is a Poisson process with rate λ, show that the moment generating function of

$$U_t = \frac{N_t - \mathbb{E}(N_t)}{\sqrt{\operatorname{var}(N_t)}}$$

is

$$M_t(x) = \mathbb{E}(e^{xU_t}) = \exp\bigl[-x\sqrt{\lambda t} + \lambda t(e^{x/\sqrt{\lambda t}} - 1)\bigr].$$

Deduce that, as $t \to \infty$,

$$\mathbb{P}(U_t \le u) \to \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}\, dv \qquad \text{for } u \in \mathbb{R}.$$

This is the central limit theorem for a Poisson process.

11.3 Inter-arrival times and the exponential distribution

Let N be a Poisson process with rate λ. The *arrival times* $T_0, T_1, \ldots$ of N are defined as before by $T_0 = 0$ and

$$T_i = \inf\{t : N_t = i\} \qquad \text{for } i = 1, 2, \ldots. \tag{11.21}$$

In other words, T_i is the time of arrival of the ith telephone call. The *inter-arrival times* $X_1, X_2, \ldots$ are the times between successive arrivals,

$$X_i = T_i - T_{i-1} \qquad \text{for } i = 1, 2, \ldots. \tag{11.22}$$

The distributions of the X_i are very simple to describe.

Theorem 11.23 *In a Poisson process with rate λ, the inter-arrival times $X_1, X_2, \ldots$ are independent random variables, each having the exponential distribution with parameter λ.*

This result demonstrates an intimate link between the postulates for a Poisson process and the exponential distribution. Theorem 11.23 is only the tip of the iceberg: a deeper investigation into continuous-time random processes reveals that the exponential distribution is a cornerstone for processes which satisfy an independence condition such as assumption D. The reason for this is that the exponential distribution is the only continous distribution with the so-called *lack-of-memory property*.

Definition 11.24 *A positive random variable X is said to have the* ***lack-of-memory property*** *if*

$$\mathbb{P}\big(X > u + v \,\big|\, X > u\big) = \mathbb{P}(X > v) \qquad \textit{for } u, v \geq 0. \tag{11.25}$$

Thinking about X as the time which elapses before some event A, say, then condition (11.25) requires that if A has not occurred by time u, then the time which elapses subsequently (between u and the occurrence of A) does not depend on the value of u: 'the random variable X does not remember how old it is when it plans its future'.

Theorem 11.26 *The continuous random variable X has the lack-of-memory property if and only if X is exponentially distributed.*

Proof If X is exponentially distributed with parameter λ then, for $u, v \geq 0$,

$$\begin{aligned}
\mathbb{P}\big(X > u + v \,\big|\, X > u\big) &= \frac{\mathbb{P}(X > u + v \text{ and } X > u)}{\mathbb{P}(X > u)} \\
&= \frac{\mathbb{P}(X > u + v)}{\mathbb{P}(X > u)} && \text{since } u \leq u + v \\
&= \frac{e^{-\lambda(u+v)}}{e^{-\lambda u}} && \text{from Example 5.22} \\
&= e^{-\lambda v} = \mathbb{P}(X > v),
\end{aligned}$$

so that X has the lack-of-memory property.

Conversely, suppose that X is positive and continuous, and has the lack-of-memory property. Let $G(u) = \mathbb{P}(X > u)$ for $u \geq 0$. The left-hand side of (11.25) is

$$\mathbb{P}\big(X > u + v \,\big|\, X > u\big) = \frac{\mathbb{P}(X > u + v)}{\mathbb{P}(X > u)} = \frac{G(u+v)}{G(u)},$$

and so G satisfies the 'functional equation'

$$G(u + v) = G(u)G(v) \qquad \text{for } u, v \geq 0. \tag{11.27}$$

The function $G(u)$ is non-increasing in the real variable u, and all non-zero non-increasing solutions of (11.27) are of the form

$$G(u) = e^{-\lambda u} \qquad \text{for } u \geq 0, \tag{11.28}$$

where λ is some constant. It is an interesting exercise in analysis to derive (11.28) from (11.27), and we suggest that the reader check this. First, use (11.27) to show that $G(n) = G(1)^n$ for $n = 0, 1, 2, \ldots$, then deduce that $G(u) = G(1)^u$ for all non-negative rationals u, and finally use monotonicity to extend this from the rationals to the reals. □

Sketch proof of Theorem 11.23 Consider X_1 first. Clearly,

$$\mathbb{P}(X_1 > u) = \mathbb{P}(N_u = 0) \qquad \text{for } u \geq 0,$$

and Theorem 11.6 gives

$$\mathbb{P}(X_1 > u) = e^{-\lambda u} \qquad \text{for } u \geq 0,$$

so that X_1 has the exponential distribution with parameter λ. From the independence assumption D, arrivals in the interval $(0, X_1]$ are independent of arrivals subsequent to X_1, and it follows that the 'waiting time' X_2 for the next arrival after X_1 is independent of X_1. Furthermore, arrivals occur 'homogeneously' in time (since the probability of an arrival in $(t, t + h]$ does not depend on t but only on h—remember (11.1)), giving that X_2 has the same distribution as X_1. Similarly, all the X_i are independent with the same distribution as X_1. □

The argument of the proof above is incomplete, since the step involving independence deals with an interval $(0, X_1]$ of *random* length. It is not entirely a trivial task to make this step rigorous, and it is for this reason that the proof is only sketched here. The required property of a Poisson process is sometimes called the 'strong Markov property', and we return to this for processes in discrete rather than continuous time in Section 12.7.

We have shown above that, if N is a Poisson process with parameter λ, the times $X_1, X_2, \ldots$ between arrivals in this process are independent and identically distributed with the exponential distribution, parameter λ. This conclusion characterizes the Poisson process, in the sense that Poisson processes are the only 'arrival processes' with this property. More properly, we have the following. Let $X_1^*, X_2^*, \ldots$ be independent random variables, each having the exponential distribution with parameter λ (> 0), and suppose that the telephone at the Grand Hotel is replaced by a very special new model which is programmed to ring at the times

$$T_1^* = X_1^*, \quad T_2^* = X_1^* + X_2^*, \quad T_3^* = X_1^* + X_2^* + X_3^*, \quad \ldots,$$

so that the time which elapses between the $(i-1)$th and the ith call equals X_i^*. Let

$$N_t^* = \max\{k : T_k^* \le t\}$$

be the number of calls which have arrived by time t. Then the process $N^* = (N_t^* : t \ge 0)$ is a Poisson process with rate λ, so that from Bill's point of view the new telephone behaves in exactly the same way (statistically speaking) as the old model.

Example 11.29 Suppose that buses for downtown arrive at the bus stop on the corner in the manner of a Poisson process. Knowing this, David expects to wait an exponentially distributed period of time before a bus will pick him up. If he arrives at the bus stop and Doris tells him that she has been waiting 50 minutes already, then this is neither good nor bad news for him, since the exponential distribution has the lack-of-memory property. Similarly, if he arrives just in time to see a bus departing, then he need not worry that his wait will be longer than usual. These properties are characteristics of the Poisson process. △

Exercise 11.30 Let M and N be independent Poisson processes, M having rate λ and N having rate μ. Use the result of Problem 6.9.4 to show that the process $M + N = (M_t + N_t : t \ge 0)$ is a Poisson process with rate $\lambda + \mu$. Compare this method with that of Exercise 11.5.

Exercise 11.31 If T_i is the time of the ith arrival in the Poisson process N, show that $N_t < k$ if and only if $T_k > t$. Use Theorem 11.23 and the central limit theorem, Theorem 8.25, to deduce that, as $t \to \infty$,

$$\mathbb{P}\left(\frac{N_t - \lambda t}{\sqrt{\lambda t}} \le u\right) \to \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}\, dv \qquad \text{for } u \in \mathbb{R}.$$

Compare this with Exercise 11.20.

Exercise 11.32 Calls arrive at a telephone exchange in the manner of a Poisson process with rate λ, but the operator is frequently distracted and answers every other call only. What is the common distribution of the time intervals which elapse between successive calls which elicit responses?

11.4 Population growth, and the simple birth process

The ideas of the last sections have many applications, one of which is a continuous-time model for population growth. We are thinking here of a simple model for phenomena such as the progressive cell division of an amoeba, and we shall formulate the process in these terms. A hypothetical type of amoeba multiplies in the following way. At time $t = 0$, there are a number I, say, of initial amoebas in a large pond. As time passes, these amoebas multiply in number by the process of progressive cell division. When an amoeba divides, the single parent amoeba is replaced by exactly two identical copies of itself. The number of amoebas in the pond grows as time passes, but we cannot say with certainty how many there will be in the future, since cell divisions occur at *random* times (rather as telephone calls arrived at random in the earlier model). We make the following two assumptions about this process.[2]

[2]In practice, amoebas and bacteria multiply at rates which depend upon their environments. Although there is considerable variation between the life cycles of cells in the same environment, these generally lack the degrees of homogeneity and independence which we postulate here.

Division rate: each amoeba present in the pond at time t has a chance of dividing during the short time interval $(t, t+h]$. There exists a constant λ (> 0), called the *birth rate*, such that the probability that any such amoeba

(a) divides *once* in the time interval $(t, t+h]$ equals $\lambda h + \mathrm{o}(h)$,
(b) does *not* divide in the time interval $(t, t+h]$ equals $1 - \lambda h + \mathrm{o}(h)$.

Independence: for each amoeba at time t, all the future divisions of this amoeba occur independently both of its past history and of the activities (past and future) of all other amoebas present at time t.

Each amoeba present at time t has probability $\lambda h + \mathrm{o}(h)$ of dividing into two amoebas by time $t+h$, probability $1 - \lambda h + \mathrm{o}(h)$ of giving rise to one amoeba (itself) by time $t+h$, and consequently probability $\mathrm{o}(h)$ of giving rise to more than two amoebas by $t+h$. Let M_t be the number of amoebas present at time t. From the previous observations, it is not difficult to write down the way in which the distribution of M_{t+h} depends on M_t. Suppose that $M_t = k$. Then $M_{t+h} \geq k$, and

$$\begin{aligned}\mathbb{P}\big(M_{t+h} = k \,\big|\, M_t = k\big) &= \mathbb{P}(\text{no division}) \\ &= [1 - \lambda h + \mathrm{o}(h)]^k \\ &= 1 - \lambda k h + \mathrm{o}(h). \end{aligned} \tag{11.33}$$

Also,

$$\begin{aligned}\mathbb{P}\big(M_{t+h} = k+1 \,\big|\, M_t = k\big) &= \mathbb{P}(\text{exactly one division}) \\ &= \binom{k}{1}[\lambda h + \mathrm{o}(h)][1 - \lambda h + \mathrm{o}(h)]^{k-1} \\ &= \lambda k h + \mathrm{o}(h), \end{aligned} \tag{11.34}$$

since there are k possible choices for the cell division. Finally,

$$\begin{aligned}\mathbb{P}\big(M_{t+h} \geq k+2 \,\big|\, M_t = k\big) &= 1 - \mathbb{P}\big(M_{t+h} \text{ is } k \text{ or } k+1 \,\big|\, M_t = k\big) \\ &= 1 - [\lambda k h + \mathrm{o}(h)] - [1 - \lambda k h + \mathrm{o}(h)] \\ &= \mathrm{o}(h). \end{aligned} \tag{11.35}$$

Consequently, the process M evolves in very much the same general way as the Poisson process N, in that if $M_t = k$, then either $M_{t+h} = k$ or $M_{t+h} = k+1$ with probability $1 - \mathrm{o}(h)$. The big difference between M and N lies in a comparison of (11.34) and (11.1): the rate at which M increases is proportional to M itself, whereas a Poisson process increases at a constant rate. The process M is called a *simple* (*linear*) *birth* process or a *pure birth* process. We treat it with the same techniques which we used for the Poisson process.

Theorem 11.36 *If $M_0 = I$ and $t > 0$, then*

$$\mathbb{P}(M_t = k) = \binom{k-1}{I-1} e^{-I\lambda t}(1 - e^{-\lambda t})^{k-I} \qquad \text{for } k = I, I+1, \ldots. \tag{11.37}$$

Proof Let

$$p_k(t) = \mathbb{P}(M_t = k),$$

as before. We establish differential–difference equations for $p_I(t)$, $p_{I+1}(t), \ldots$ in just the same way as we found (11.11) and (11.12). Thus, we have from the partition theorem that, for $h > 0$,

$$\begin{aligned}\mathbb{P}(M_{t+h} = k) &= \sum_i \mathbb{P}\big(M_{t+h} = k \,\big|\, M_t = i\big)\mathbb{P}(M_t = i) \\ &= \big[1 - \lambda k h + \mathrm{o}(h)\big]\mathbb{P}(M_t = k) + \big[\lambda(k-1)h + \mathrm{o}(h)\big]\mathbb{P}(M_t = k-1) + \mathrm{o}(h)\end{aligned}$$

by (11.33)–(11.35), giving that

$$p_k(t+h) - p_k(t) = \lambda(k-1)hp_{k-1}(t) - \lambda k h p_k(t) + \mathrm{o}(h).$$

Divide this equation by h and take the limit as $h \downarrow 0$ to obtain

$$p_k'(t) = \lambda(k-1)p_{k-1}(t) - \lambda k p_k(t) \qquad \text{for } k = I, I+1, \ldots. \tag{11.38}$$

The equation for $p_I'(t)$ involves $p_{I-1}(t)$, and we note that $p_{I-1}(t) = 0$ for all t. We can solve (11.38) recursively subject to the boundary condition

$$p_k(0) = \begin{cases} 1 & \text{if } k = I, \\ 0 & \text{if } k \neq I. \end{cases} \tag{11.39}$$

That is to say, first find $p_I(t)$, then $p_{I+1}(t), \ldots$, and formula (11.37) follows by induction.

The method of generating functions works also. If we multiply through (11.38) by s^k and sum over k, we obtain the partial differential equation

$$\frac{\partial G}{\partial t} = \lambda s(s-1)\frac{\partial G}{\partial s}, \tag{11.40}$$

where $G = G(s,t)$ is the generating function

$$G(s,t) = \sum_{k=I}^{\infty} p_k(t)s^k.$$

It is not difficult to solve this differential equation subject to the boundary condition $G(s,0) = s^I$, but we do not require such skills from the reader. □

The mean and variance of M_t may be calculated directly from (11.37) in the usual way. These calculations are a little complicated since M_t has a negative binomial distribution, and it is simpler to use the following trick. Writing

$$\mu(t) = \mathbb{E}(M_t) = \sum_{k=I}^{\infty} k p_k(t),$$

we have, by differentiating blithely through the summation again, that

$$\mu'(t) = \sum_{k=I}^{\infty} k p_k'(t) = \sum_{k=I}^{\infty} k\big[\lambda(k-1)p_{k-1}(t) - \lambda k p_k(t)\big] \tag{11.41}$$

from (11.38). We collect the coefficients of $p_k(t)$ together here, to obtain

$$\mu'(t) = \lambda \sum_{k=I}^{\infty} \big[(k+1)k p_k(t) - k^2 p_k(t)\big] \tag{11.42}$$

$$= \lambda \sum_{k=I}^{\infty} k p_k(t) = \lambda \mu(t), \tag{11.43}$$

which is a differential equation in μ with boundary condition

$$\mu(0) = \mathbb{E}(M_0) = I.$$

This differential equation has solution

$$\mu(t) = I e^{\lambda t}, \tag{11.44}$$

showing that (on average) amoebas multiply at an exponential rate (whereas a Poisson process grows linearly on average, remember (11.8)). The same type of argument may be used to calculate $\mathbb{E}(M_t^2)$. This is more complicated and leads to an expression for the variance of M_t,

$$\operatorname{var}(M_t) = I e^{2\lambda t}(1 - e^{-\lambda t}). \tag{11.45}$$

An alternative method of calculating the mean and variance of M_t proceeds by way of the differential equation (11.40) for the probability generating function $G(s, t)$ of M_t. Remember that

$$G(1, t) = 1, \quad \left.\frac{\partial G}{\partial s}\right|_{s=1} = \mu(t).$$

We differentiate throughout (11.40) with respect to s and substitute $s = 1$ to obtain

$$\left.\frac{\partial^2 G}{\partial s \partial t}\right|_{s=1} = \lambda \left.\frac{\partial G}{\partial s}\right|_{s=1}.$$

Assuming that we may interchange the order of differentiation in the first term here, this equation becomes

$$\mu'(t) = \lambda \mu(t),$$

in agreement with (11.43). The variance may be found similarly, by differentiating twice.

Exercise 11.46 Show that, in the simple birth process above, the period of time during which there are exactly k ($\geq I$) individuals is a random variable having the exponential distribution with parameter λk.

Exercise 11.47 Deduce from the result of Exercise 11.46 that the time $T_{I,J}$ required by the birth process to grow from size I to size J ($> I$) has mean and variance given by

$$\mathbb{E}(T_{I,J}) = \sum_{k=I}^{J-1} \frac{1}{\lambda k}, \quad \operatorname{var}(T_{I,J}) = \sum_{k=I}^{J-1} \frac{1}{(\lambda k)^2}.$$

Exercise 11.48 Show that the variance of the simple birth process M_t is given by

$$\mathrm{var}(M_t) = Ie^{2\lambda t}(1 - e^{-\lambda t}).$$

11.5 Birth and death processes

It is usually the case that in any system involving births, there are deaths also. Most telephone calls last for only a finite time, and most bacteria die out after their phases of reproduction. We introduce death into the simple birth process of the last section by replacing the division postulate of the last section by two postulates concerning divisions and deaths, respectively. We shall suppose that our hypothetical type of amoeba satisfies the following.

Division rate: each amoeba present in the pond at time t has a chance of dividing in the short time interval $(t, t + h]$. There exists a constant λ (> 0), called the *birth rate*, such that the probability that any such amoeba

(a) divides *once* during the time interval $(t, t + h]$ equals $\lambda h + \mathrm{o}(h)$,
(b) divides more than once during the time interval $(t, t + h]$ equals $\mathrm{o}(h)$.

Death rate: each amoeba present at time t has a chance of dying, and hence being removed from the population, during the short time interval $(t, t + h]$. There exists a constant μ (> 0), called the *death rate*, such that the probability that any such amoeba dies during the time interval $(t, t + h]$ equals $\mu h + \mathrm{o}(h)$.

We assume also that deaths occur independently of other deaths and of all births. For the time interval $(t, t + h]$, there are now several possibilities for each amoeba present at time t:

(i) death, with probability $\mu h + \mathrm{o}(h)$,
(ii) a single division, with probability $\lambda h + \mathrm{o}(h)$,
(iii) no change of state, with probability $[1-\lambda h+\mathrm{o}(h)][1-\mu h+\mathrm{o}(h)] = 1-(\lambda+\mu)h+\mathrm{o}(h)$,
(iv) some other combination of birth and death (such as division and death, or two divisions), with probability $\mathrm{o}(h)$.

Only possibilities (i)–(iii) have probabilities sufficiently large to be taken into account. Similarly, the probability of two or more amoebas changing their states (by division or death) during the time interval $(t, t + h]$ equals $\mathrm{o}(h)$.

We write L_t for the number of amoebas which are alive at time t, and we find the distribution of L_t in the same way as before. The first step mimics (11.33)–(11.35), and similar calculations to those equations give that, for $k = 0, 1, 2, \ldots$,

$$\mathbb{P}(L_{t+h} = k \mid L_t = k) = 1 - (\lambda + \mu)kh + \mathrm{o}(h), \tag{11.49}$$

$$\mathbb{P}(L_{t+h} = k + 1 \mid L_t = k) = \lambda kh + \mathrm{o}(h), \tag{11.50}$$

$$\mathbb{P}(L_{t+h} = k - 1 \mid L_t = k) = \mu kh + \mathrm{o}(h), \tag{11.51}$$

$$\mathbb{P}(L_{t+h} > k + 1 \text{ or } L_{t+h} < k - 1 \mid L_t = k) = \mathrm{o}(h). \tag{11.52}$$

Note that, if $L_t = k$, the rate of birth in the population is λk and the rate of death is μk. This linearity in k arises because there are k independent possibilities for a birth or a death—remember (11.34).

This *birth–death process* differs from the simple birth process in a very important respect—it has an *absorbing state*, in the sense that if at some time there are no living cells, then there will never be any living cells subsequently.

Unlike the Poisson process and the simple birth process, it is not very easy to write down the mass function of L_t explicitly, since the corresponding differential–difference equations are not easily solved by recursion. The method of generating functions is still useful.

Theorem 11.53 *If $L_0 = I$, then L_t has probability generating function*

$$\mathbb{E}(s^{L_t}) = \begin{cases} \left(\dfrac{\lambda t(1-s)+s}{\lambda t(1-s)+1}\right)^I & \text{if } \mu = \lambda, \\ \left(\dfrac{\mu(1-s)-(\mu-\lambda s)e^{t(\mu-\lambda)}}{\lambda(1-s)-(\mu-\lambda s)e^{t(\mu-\lambda)}}\right)^I & \text{if } \mu \neq \lambda. \end{cases} \tag{11.54}$$

Proof The differential–difference equations for $p_k(t) = \mathbb{P}(L_t = k)$ are

$$p_k'(t) = \lambda(k-1)p_{k-1}(t) - (\lambda+\mu)kp_k(t) + \mu(k+1)p_{k+1}(t), \tag{11.55}$$

valid for $k = 0, 1, 2, \ldots$ subject to the convention that $p_{-1}(t) = 0$ for all t. The boundary condition is

$$p_k(0) = \begin{cases} 1 & \text{if } k = I, \\ 0 & \text{if } k \neq I. \end{cases} \tag{11.56}$$

Recursive solution of (11.55) fails since the equation in $p_0'(t)$ involves both $p_0(t)$ and $p_1(t)$ on the right-hand side. We solve these equations by the method of generating functions, first introducing the probability generating function

$$G(s,t) = \mathbb{E}(s^{L_t}) = \sum_{k=0}^{\infty} p_k(t)s^k.$$

Multiply throughout (11.55) by s^k and sum over k to obtain the partial differential equation

$$\frac{\partial G}{\partial t} = (\lambda s - \mu)(s-1)\frac{\partial G}{\partial s},$$

with boundary condition $G(s,0) = s^I$. The diligent reader may check that the solution is given by (11.54). □

It is possible that this birth–death process L will become extinct ultimately, in that $L_t = 0$ for some t. The probability that this happens is easily calculated from the result of Theorem 11.53.

Theorem 11.57 *Let $L_0 = I$, and write $e(t) = \mathbb{P}(L_t = 0)$ for the probability that the process is extinct by time t. As $t \to \infty$,*

$$e(t) \to \begin{cases} 1 & \text{if } \lambda \leq \mu, \\ (\mu/\lambda)^I & \text{if } \lambda > \mu. \end{cases} \tag{11.58}$$

Hence, extinction is certain if and only if the death rate is at least as big as the birth rate. We shall prove Theorem 11.57 directly, while noting two alternative proofs using random walks and branching processes, respectively.

First, the theorem is reminiscent of the Gambler's Ruin Problem and Theorem 10.32. Actually, (11.58) may be derived directly from the conclusion of Theorem 10.32 by studying what is called the 'embedded random walk' of the birth–death process, see Grimmett and Stirzaker (2001, Sect. 6.11) for the details.

Secondly, there is an 'embedded branching process' of amoebas. On dying, an amoeba is replaced by either no amoeba or by two amoebas, with respective probabilities $\mu/(\lambda+\mu)$ and $\lambda/(\lambda+\mu)$. One may now use the extinction probability theorem for branching processes, Theorem 9.19.

Proof Clearly,

$$e(t) = \mathbb{P}(L_t = 0) = G(0,t).$$

By (11.54),

$$G(0,t) = \begin{cases} \left(\dfrac{\lambda t}{\lambda t+1}\right)^I & \text{if } \lambda = \mu, \\ \left(\dfrac{\mu - \mu e^{t(\mu-\lambda)}}{\lambda - \mu e^{t(\mu-\lambda)}}\right)^I & \text{if } \lambda \neq \mu, \end{cases}$$

and the result follows immediately. □

Exercise 11.59 Let $m(t)$ be the expected size at time t of the population in a birth–death process with birth rate λ and death rate μ. Use (11.55) to show that m satisfies the differential equation

$$m'(t) = (\lambda - \mu)m(t).$$

Hence find $m(t)$ in terms of the initial size of the population.

Exercise 11.60 A birth–death process L has birth rate λ and death rate μ. If the population has size k at time t, show that the subsequent length of time which elapses before there is either a birth or a death is a random variable having the exponential distribution with parameter $(\lambda+\mu)k$.

Exercise 11.61 Let L be a birth–death process with birth rate 1 and death rate 1. Suppose that L_0 is a random variable having the Poisson distribution with parameter α. Show that the probability that the process is extinct by time t is $\exp[-\alpha/(t+1)]$.

11.6 A simple queueing model

We all know how it feels to be waiting in a queue, whether it be buying postage stamps at the post office at lunchtime, calling an insurance company, waiting for a response from a website, or waiting to be called in for a minor operation at the local hospital.

There are many different types of queue, and there are three principal ways in which we may categorize them, according to the ways in which

(a) people arrive in the system,

(b) these people are stored in the system prior to their service,
(c) these people are served.

In many queues, only the method (b) of storage of waiting customers can be predicted with certainty—*first come, first served* is a common 'queue discipline' in shops, although there are many other possibilities. On the other hand, it is generally impossible to predict exactly when people will join the queue and how long they will require for service, and this is the reason why probability theory is important in describing queueing systems. The theory of queues is an old favourite amongst probabilists, as it is a rich source of interesting and diverse problems.

We shall consider a simple model of a queue. There are many others, most of which are more complicated than this one, and the reader may find amusement in devising some of these. Our example goes as follows. In the German delicatessen in the market, Angela is the only shop assistant. Customers arrive in the shop, wait for their turn to be served by Angela, and then leave after their service has been completed. There is randomness in the way in which they arrive and in the lengths of their service times (people rarely visit good delicatessens and buy only one type of cheese). We suppose the following.

Arrivals: Customers arrive in the manner of a Poisson process with rate λ (> 0). That is to say, if N_t is the number who have arrived by time t, then $N = (N_t : t \geq 0)$ is a Poisson process with rate λ.

Service: The service time of each customer is a random variable having the exponential distribution with parameter μ (> 0), and the service times of different customers are independent random variables.

Independence: Service times are independent of arrival times, and Angela works no faster when the shop is crowded than she does when the shop is nearly empty.

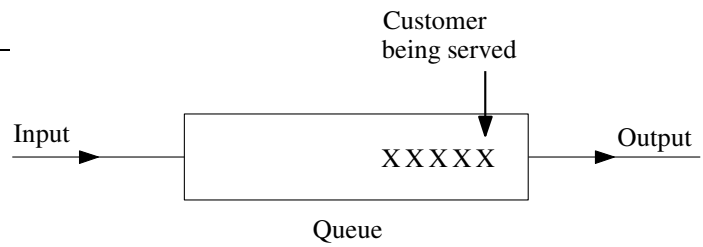


Fig. 11.2 A simple queue.

It is not very important to us how a customer is stored between his or her arrival and departure, but for the sake of definiteness we shall suppose that Angela tolerates no disorder, insisting that her customers form a single line and always serving the customer at the head of this line in the usual way. This queue discipline is called *first come, first served* or *first in, first out*. Thus, Angela's shop looks something like Figure 11.2. Other delicatessens are less disciplined, with customers milling around in the shop and the shop assistant serving people chosen at random.

The arrivals assumption is equivalent to demanding that the times between successive arrivals are independent, exponentially distributed random variables with parameter λ. Our

assumption that both inter-arrival times and service times have the exponential distribution is crucial for this example, since only the exponential distribution has the lack-of-memory property, Theorem 11.26. This assumption has the following consequence. If we glance through the shop window at a certain time t, seeing ten people within, say, then the times of the next arrival and the next departure do not depend on the times of the last arrival and the last departure. Thus, for example, for $h > 0$, the probability of a single arrival during the time interval $(t, t+h]$ equals $\lambda h + \mathrm{o}(h)$, and the probability of no arrival equals $1 - \lambda h + \mathrm{o}(h)$.

Also, if Angela is serving someone at time t, then the probability that she is still serving this person at time $t+h$ equals $\mathbb{P}(S > t+h-\tau \mid S > t-\tau)$, where S is the service time of the customer in question and τ is the time at which the service period began. Now,

$$\begin{aligned}\mathbb{P}\big(S > t+h-\tau \,\big|\, S > t-\tau\big) &= \mathbb{P}(S > h)\\ &= e^{-\mu h} = 1 - \mu h + \mathrm{o}(h),\end{aligned}$$

by Theorem 11.26 and the lack-of-memory property (11.24). Therefore, for $h > 0$, if Angela is serving someone at time t, then the probability that this service is completed during the time interval $(t, t+h]$ equals $\mu h + \mathrm{o}(h)$.

Let Q_t be the number of people in the queue, including the person being served, at time t, and suppose that $Q_0 = 0$. The process $Q = (Q_t : t \geq 0)$ is a type of birth–death process since, if $Q_t = k$, say, then Q_{t+h} equals one of $k-1, k, k+1$ with probability $1 - \mathrm{o}(h)$. The only events which may happen with significant probability (that is, larger than $\mathrm{o}(h)$) during the time interval $(t, t+h]$ are a single departure, a single arrival, or no change of state. More precisely, if $k \geq 1$,

$$\begin{aligned}\mathbb{P}\big(Q_{t+h} = k \,\big|\, Q_t = k\big) &= \mathbb{P}(\text{no arrival, no departure}) + \mathrm{o}(h)\\ &= [1 - \lambda h + \mathrm{o}(h)][1 - \mu h + \mathrm{o}(h)] + \mathrm{o}(h)\\ &= 1 - (\lambda + \mu)h + \mathrm{o}(h), \qquad (11.62)\\ \mathbb{P}\big(Q_{t+h} = k-1 \,\big|\, Q_t = k\big) &= \mathbb{P}(\text{no arrival, one departure}) + \mathrm{o}(h)\\ &= [1 - \lambda h + \mathrm{o}(h)][\mu h + \mathrm{o}(h)] + \mathrm{o}(h)\\ &= \mu h + \mathrm{o}(h), \qquad (11.63)\end{aligned}$$

and, for $k \geq 0$,

$$\begin{aligned}\mathbb{P}\big(Q_{t+h} = k+1 \,\big|\, Q_t = k\big) &= \mathbb{P}(\text{one arrival, no departure}) + \mathrm{o}(h)\\ &= \begin{cases}[\lambda h + \mathrm{o}(h)][1 - \mu h + \mathrm{o}(h)] + \mathrm{o}(h) & \text{if } k \geq 1,\\ \lambda h + \mathrm{o}(h) & \text{if } k = 0,\end{cases}\\ &= \lambda h + \mathrm{o}(h). \qquad (11.64)\end{aligned}$$

Finally,

$$\begin{aligned}\mathbb{P}\big(Q_{t+h} = 0 \,\big|\, Q_t = 0\big) &= \mathbb{P}(\text{no arrival}) + \mathrm{o}(h)\\ &= 1 - \lambda h + \mathrm{o}(h). \qquad (11.65)\end{aligned}$$

Equations (11.62)–(11.65) are very similar to the corresponding equations (11.49)–(11.51) for the simple birth–death process, the only significant difference being that arrivals and departures occur here at rates which do not depend upon the current queue size (unless it is

empty, so that departures are impossible), whereas in the simple birth–death process, the corresponding rates are linear functions of the current population size. It may seem at first sight that this queueing process is simpler than the birth–death process, but the truth turns out to be the opposite: there is no primitive way of calculating the mass function of Q_t for $t > 0$. The difficulty is as follows. As usual, we may use (11.62)–(11.65) to establish a system of differential–difference equations for the probabilities

$$p_k(t) = \mathbb{P}(Q_t = k),$$

and these turn out to be

$$p_k'(t) = \lambda p_{k-1}(t) - (\lambda + \mu)p_k(t) + \mu p_{k+1}(t) \qquad \text{for } k = 1, 2, \ldots, \tag{11.66}$$

and

$$p_0'(t) = \lambda p_0(t) + \mu p_1(t), \tag{11.67}$$

subject to the boundary condition

$$p_k(0) = \begin{cases} 1 & \text{if } k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

We cannot solve this system of equations recursively since the equation (11.67) for $p_0(t)$ involves $p_1(t)$ also. Furthermore, the method of generating functions leads to the differential equation

$$\frac{\partial G}{\partial t} = \frac{s-1}{s}\big[\lambda s G - \mu G + \mu p_0(t)\big]$$

for $G(s, t) = \mathbb{E}(s^{Q_t})$, and this equation involves the unknown function $p_0(t)$. Laplace transforms turn out to be the key to solving (11.66) and (11.67), and the answer is not particularly pretty:

$$p_k(t) = J_k(t) - J_{k+1}(t) \qquad \text{for } k = 0, 1, 2, \ldots, \tag{11.68}$$

where

$$J_k(t) = \int_0^t \left(\frac{\lambda}{\mu}\right)^{\frac{1}{2}k} \frac{k}{s} e^{-s(\lambda+\mu)} I_k(2s\sqrt{\lambda\mu})\, ds$$

and $I_k(t)$ is a modified Bessel function. We shall not prove this here, of course, but refer those interested to Feller (1971, p. 482).

The long-term behaviour of the queue is of major interest. If service times are long in comparison with inter-arrival times, the queue will tend to grow, so that after a long period of time it will be very large. On the other hand, if service times are relatively short, it is reasonable to expect that the queue length will settle down into some 'steady state'. The asymptotic behaviour of the queue length Q_t as $t \to \infty$ is described by the sequence

$$\pi_k = \lim_{t\to\infty} p_k(t) \qquad \text{for } k = 1, 2, \ldots, \tag{11.69}$$

of limiting probabilities, and it is to this sequence $\pi_0, \pi_1, \ldots$ that we turn our attention. It is in fact the case that the limits exist in (11.69), but we shall not prove this. Neither do we prove that

$$0 = \lim_{t\to\infty} p_k'(t) \qquad \text{for } k = 0, 1, 2, \ldots, \tag{11.70}$$

which follows intuitively from (11.69) by differentiating both sides of (11.69) and interchanging the limit and the differential operator. The values of the π_k are found by letting $t \to \infty$ in (11.66)–(11.67) and using (11.69)–(11.70) to obtain the following difference equations for the sequence $\pi_0, \pi_1, \ldots$:

$$0 = \lambda\pi_{k-1} - (\lambda + \mu)\pi_k + \mu\lambda_{k+1} \qquad \text{for } k = 1, 2, \ldots, \tag{11.71}$$

$$0 = -\lambda\pi_0 + \mu\pi_1. \tag{11.72}$$

We call a non-negative sequence $\pi_0, \pi_1, \ldots$ a *steady-state distribution* of the queue if it satisfies (11.71)–(11.72) together with

$$\sum_{k=0}^{\infty} \pi_k = 1. \tag{11.73}$$

Theorem 11.74 *If $\lambda < \mu$, the queue has a unique steady-state distribution given by*

$$\pi_k = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^k \qquad \textit{for } k = 0, 1, 2, \ldots. \tag{11.75}$$

If $\lambda \geq \mu$, there is no steady-state distribution.

We call the ratio $\rho = \lambda/\mu$ the *traffic intensity* of the queue; ρ is the ratio of the arrival rate to the service rate. We may interpret Theorem 11.74 as follows:

(a) if $\rho < 1$, the queue length Q_t settles down as $t \to \infty$ into a steady-state or 'equilibrium' distribution, for which the probability that k customers are present equals $(1 - \rho)\rho^k$,
(b) if $\rho \geq 1$, there is no steady-state distribution, indicating that the rate of arrival of new customers is too large for the single server to cope, and the queue length either grows beyond all bounds or it has fluctuations of large order.

Theorem 11.74 may remind the reader of the final theorem, Theorem 10.32, about a random walk with an absorbing barrier. Just as in the case of the simple birth–death process, there is a random walk embedded in this queueing process, and this random walk has a 'retaining' barrier at 0 which prevents the walk from visiting the negative integers but allows the walk to revisit the positive integers.

Proof of Theorem 11.74 We wish to solve the difference equations (11.71)–(11.72), and we do this recursively rather than using the general method of Appendix B. We find π_1 in terms of π_0 from (11.72):

$$\pi_1 = \rho\pi_0.$$

We substitute this into (11.71) with $k = 1$ to find that

$$\pi_2 = \rho^2\pi_0,$$

and we deduce the general solution

$$\pi_k = \rho^k \pi_0 \qquad \text{for } k = 0, 1, 2, \ldots \tag{11.76}$$

by induction on k. Now, $\pi_0, \pi_1, \ldots$ is a steady-state solution if and only if (11.73) holds. By (11.76),

$$\sum_{k=0}^{\infty} \pi_k = \pi_0 \sum_{k=0}^{\infty} \rho^k.$$

If $\rho < 1$, then (11.73) holds if and only if $\pi_0 = 1 - \rho$. On the other hand, if $\rho \geq 1$, (11.73) holds for no value of π_0, and the proof is complete. □

Exercise 11.77 If Q is the above queueing process with arrival rate λ and service rate μ, and a customer arrives to find exactly k customers waiting ahead (including the person being served), show that this customer leaves the queueing system after a length of time which has the gamma distribution with parameters $k + 1$ and μ.

Exercise 11.78 Show that $p_k(t) = (1 - \rho)\rho^k$ is a solution to equations (11.66)–(11.67) so long as $\rho = \lambda/\mu < 1$. This proves that, if the process begins in its steady-state distribution, then it has this distribution for all time.

Exercise 11.79 A queue has three servers A_1, A_2, A_3, and their service times are independent random variables, A_i's service times having the exponential distribution with parameter μ_i. An arriving customer finds all three servers unoccupied and chooses one at random, each being equally likely. If the customer is still being served at time t, what is the probability that A_1 was chosen?

11.7 Problems

1. If N is a Poisson process with rate λ, what is the distribution of $N_t - N_s$ for $0 \leq s \leq t$?
2. If N is a Poisson process with rate λ, show that $\operatorname{cov}(N_s, N_t) = \lambda s$ if $0 \leq s < t$.
3. Three apparently identical robots, called James, Simon, and John, are set to work at time $t = 0$. Subsequently, each stops working after a random length of time, independently of the other two, and the probability that any given robot stops in the short time interval $(t, t + h)$ is $\lambda h + o(h)$. Show that each robot works for a period of time with the exponential distribution, parameter λ, and that the probability that at least one of the three has stopped by time t is $1 - e^{-3\lambda t}$.

   What is the probability that they stop work in the order James, Simon, John?

* 4. Let $X_1, X_2, X_3, \ldots$ be a sequence of independent, identically distributed random variables having the distribution function

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{for } x < 0, \end{cases}$$

where λ is a positive constant. If $S_n = \sum_{i=1}^{n} X_i$, prove that S_n has density function

$$f_n(x) = \frac{1}{(n-1)!} \lambda^n x^{n-1} e^{-\lambda x} \qquad \text{for } x \geq 0.$$

Deduce that $N_t = \max\{n : S_n \leq t\}$ has a Poisson distribution.

The *excess life* e_t is defined by

$$e_t = S_{N_t+1} - t.$$

If $g(t,x) = \mathbb{P}(e_t > x)$ then, by considering the distribution of e_t conditional on the value of X_1, show that

$$g(t,x) = e^{-\lambda(t+x)} + \int_0^t g(t-u,x)\lambda e^{-\lambda u}\,du.$$

Find a solution of this equation. (Oxford 1976F)

5. Tourist coaches arrive at Buckingham Palace in the manner of a Poisson process with rate λ, and the numbers of tourists in the coaches are independent random variables, each having probability generating function $G(s)$. Show that the total number of tourists who have arrived at the palace by time t has probability generating function

$$\exp\bigl(\lambda t[G(s)-1]\bigr).$$

This is an example of a so-called 'compound' Poisson process.

6. The probability of one failure in a system occurring in the time interval $(t, t+\tau)$ is $\lambda(t)\tau + o(\tau)$, independently of previous failures, and the probability of more than one failure in this interval is $o(\tau)$, where λ is a positive integrable function called the *rate function*.

Prove that the number of failures in $(0,t)$ has the Poisson distribution with mean $\int_0^t \lambda(x)\,dx$.

Let T be the time of occurrence of the first failure. Find the probability density function of T and show that, if $\lambda(t) = c/(1+t)$ where $c > 0$, the expected value of T is finite if and only if $c > 1$. (Oxford 1981F)

This is an example of a so-called 'inhomogeneous' Poisson process.

7. A 'doubly stochastic' Poisson process is an inhomogenous Poisson process in which the rate function $\lambda(t)$ is itself a random process. Show that the simple birth process with birth rate λ is a doubly stochastic Poisson process N for which $\lambda(t) = \lambda N_t$.

8. In a simple birth process with birth rate λ, find the moment generating function of the time required by the process to grow from size I to size J $(> I)$.

9. Show that the moment generating function of the so-called 'extreme-value' distribution with density function

$$f(x) = \exp(-x - e^{-x}) \qquad \text{for } x \in \mathbb{R},$$

is

$$M(t) = \Gamma(1-t) \qquad \text{if } t < 1.$$

Let T_J be the time required by a simple birth process with birth rate λ to grow from size 1 to size J, and let

$$U_J = \lambda T_J - \log J.$$

Show that U_J has moment generating function

$$M_J(t) = \frac{1}{J^t}\prod_{i=1}^{J-1}\left(\frac{i}{i-t}\right) \qquad \text{if } t < 1,$$

and deduce that, as $J \to \infty$,

$$M_J(t) \to M(t) \qquad \text{if } t < 1.$$

[You may use the fact that $J^t\Gamma(J-t)/\Gamma(J) \to 1$ as $J \to \infty$.] It follows that the distribution of U_J approaches the extreme-value distribution as $J \to \infty$.

10. Consider a birth–death process whose birth and death rates satisfy $\lambda = \mu$. If the initial population size is I, show that the time T until the extinction of the process has distribution function

$$\mathbb{P}(T \le t) = \left(\frac{\lambda t}{\lambda t + 1}\right)^I \qquad \text{for } t > 0,$$

and deduce that, as $I \to \infty$, the random variable $U_I = \lambda T / I$ has limiting distribution function given by

$$\mathbb{P}(U_I \le t) \to e^{-1/t} \qquad \text{for } t \ge 0.$$

11. A population develops according to the following rules:
    (a) during the interval $(t, t + dt)$, an individual existing at time t has (independently of its previous history) probability $\lambda dt + \mathrm{o}(dt)$ of having a single offspring (twins, triplets, etc, being impossible) and a probability $\mu dt + \mathrm{o}(dt)$ of dying, where λ and μ are absolute constants,
    (b) in the interval $(t, t + dt)$, there is a probability $\theta dt + \mathrm{o}(dt)$ that a single immigrant will join the population,
    (c) subpopulations descending from distinct individuals develop independently.

    If $p_n(t)$ denotes the probability that the population consists of n individuals at time t, show that

$$\phi(z, t) = \sum_{n=0}^{\infty} z^n p_n(t)$$

satisfies the partial differential equation

$$\frac{\partial \phi}{\partial t} = (\lambda z - \mu)(z - 1)\frac{\partial \phi}{\partial z} + \theta(z - 1)\phi.$$

    In the particular case when $\lambda = \mu = \theta = 1$ and the population is empty at time $t = 0$, show that the size of the population at time t has mean t, and calculate its variance. (Oxford 1964F)

12. Consider the 'birth–death–immigration' process of Problem 11.7.11 and suppose that $\lambda, \mu, \theta > 0$. Use the ideas and methods of Section 11.6 to show that this process has a steady-state distribution if and only if $\lambda < \mu$, and in this case, the steady-state distribution is given by

$$\pi_n = \pi_0 \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n \frac{\Gamma(n + (\theta/\lambda))}{\Gamma(\theta/\lambda)} \qquad \text{for } n = 0, 1, 2, \ldots,$$

    where π_0 is chosen so that $\sum_n \pi_n = 1$.

13. The 'immigration–death' process is obtained from the birth–death–immigration process of Problem 11.7.11 by setting the birth rate λ equal to 0. Let $D = (D_t : t \ge 0)$ be an immigration–death process with positive immigration rate θ and death rate μ. Suppose that $D_0 = I$, and set up the system of differential equations which are satisfied by the probability functions

$$p_k(t) = \mathbb{P}(D_t = k).$$

    Deduce that the probability generating function $G(s, t) = \mathbb{E}(s^{D_t})$ satisfies the partial differential equation

$$\frac{\partial G}{\partial t} = (s - 1)\left(\theta G - \mu \frac{\partial G}{\partial s}\right)$$

    subject to the boundary condition $G(s, 0) = s^I$. Solve this equation to find that

$$G(s, t) = \left[1 + (s - 1)e^{-\mu t}\right]^I \exp\left[\theta(s - 1)(1 - e^{-\mu t})/\mu\right].$$

14. In the immigration–death process of Problem 11.7.13, show that there is a steady-state distribution (in the jargon of Section 11.6) for all positive values of θ and μ. Show further that this distribution is the Poisson distribution with parameter θ/μ.

15. A robot can be in either of two states: state A (idle) and state B (working). In any short time interval $(t, t+h)$, the probability that it changes its state is $\lambda h + o(h)$, where $\lambda > 0$. If $p(t)$ is the probability that it is idle at time t given that it was idle at time 0, show that

$$p'(t) = -2\lambda p(t) + \lambda.$$

Hence find $p(t)$.

Let $q(t)$ be the probability that the robot is working at time t given that it was working at time 0. Find $q(t)$, and find the distribution of the earliest time T at which there is a change of state.

If there are N robots operating independently according to the above laws and $p_k(t)$ is the probability that exactly k are idle at time t, show that

$$p_k'(t) = \lambda(N-k+1)p_{k-1}(t) - \lambda N p_k(t) + \lambda(k+1)p_{k+1}(t),$$

for $k = 0, 1, \ldots, N$, subject to the rule that $p_{-1}(t) = p_{N+1}(t) = 0$.

If all the robots are idle at time 0, show that the number of idle robots at time t has the binomial distribution with parameters N and $e^{-\lambda t}\cosh \lambda t$.

16. Prove that, in a queue whose input is a Poisson process and whose service times have the exponential distribution, the number of new arrivals during any given service time is a random variable with the geometric distribution.

17. Customers arrive in a queue according to a Poisson process with rate λ, and their service times have the exponential distribution with parameter μ. Show that, if there is only one customer in the queue, then the probability that the next customer arrives within time t and has to wait for service is

$$\frac{\lambda}{\lambda+\mu}(1 - e^{-(\lambda+\mu)t}).$$

18. Customers arrive in a queue according to a Poisson process with rate λ, and their service times have the exponential distribution with parameter μ, where $\lambda < \mu$. We suppose that the number Q_0 of customers in the system at time 0 has distribution

$$\mathbb{P}(Q_0 = k) = (1-\rho)\rho^k \qquad \text{for } k = 0, 1, 2, \ldots,$$

where $\rho = \lambda/\mu$, so that the queue is 'in equilibrium' by the conclusion of Exercise 11.78. If a customer arrives in the queue at time t, find the moment generating function of the total time which it spends in the system, including its service time. Deduce that this time has the exponential distribution with parameter $\mu(1-\rho)$.

19. Customers arrive at the door of a shop according to a Poisson process with rate λ, but they are unwilling to enter a crowded shop. If a prospective customer sees k people inside the shop, he or she enters the shop with probability $(\frac{1}{2})^k$ and otherwise leaves, never to return. The service times of customers who enter the shop are random variables with the exponential distribution, parameter μ. If Q_t is the number of people within the shop (excluding the single server) at time t, show that $p_k(t) = \mathbb{P}(Q_t = k)$ satisfies

$$p_k'(t) = \mu p_{k+1}(t) - \left(\frac{\lambda}{2^k} + \mu\right) p_k(t) + \frac{\lambda}{2^{k-1}} p_{k-1}(t) \qquad \text{for } k = 1, 2, \ldots,$$
$$p_0'(t) = \mu p_1(t) - \lambda p_0(t).$$

Deduce that there is a steady-state distribution for all positive values of λ and μ, and that this distribution is given by

$$\pi_k = \pi_0 2^{-\frac{1}{2}k(k-1)}\rho^k \qquad \text{for } k = 0, 1, 2\ldots,$$

where $\rho = \lambda/\mu$ and π_0 is chosen appropriately.

20. (a) The fire alarm in Mill Lane is set off at random times. The probability of an alarm during the time interval $(u, u+h)$ is $\lambda(u)h + \mathrm{o}(h)$, where the 'intensity function' $\lambda(u)$ may vary with the time u. Let $N(t)$ be the number of alarms by time t, and set $N(0) = 0$. Show, subject to reasonable extra conditions to be stated clearly, that $p_i(t) = \mathbb{P}(N(t) = i)$ satisfies

$$p_i'(t) = -\lambda(t)p_i(t), \quad p_i'(t) = \lambda(t)\big[p_{i-1}(t) - p_i(t)\big] \qquad \text{for } i \geq 1.$$

Deduce that $N(t)$ has the Poisson distribution with parameter $\Lambda(t) = \int_0^1 \lambda(u)\,du$.

(b) The fire alarm in Clarkson Road is different. The number $M(t)$ of alarms by time t is such that

$$\mathbb{P}\big(M(t+h) = m+1 \,\big|\, M(t) = m\big) = \lambda_m h + \mathrm{o}(h),$$

where $\lambda_m = \alpha m + \beta$, $m \geq 1$, and $\alpha, \beta > 0$. Show, subject to suitable extra conditions, that $p_m(t) = \mathbb{P}(M(t) = m)$ satisfies a set of differential–difference equations to be specified. Deduce without solving these equations in their entirety that $M(t)$ has mean $\beta(e^{\alpha t} - 1)/\alpha$, and find the variance of $M(t)$.

(Cambridge 2001)

21. (a) Define an inhomogeneous Poisson process with rate function $\lambda(u)$.

(b) Show that the number of arrivals in an inhomogeneous Poisson process during the interval $(0, t)$ has the Poisson distribution with mean $\int_0^t \lambda(u)\,du$.

(c) Suppose that $\Lambda = (\Lambda(t) : t \geq 0)$ is a non-negative, real-valued random process. Conditional on Λ, let $N = (N(t) : t \geq 0)$ be an inhomogeneous Poisson process with rate function $\Lambda(u)$. Such a process N is called a *doubly stochastic Poisson process*. Show that the variance of $N(t)$ cannot be less than its mean.

(d) Now consider the process $M(t)$ obtained by deleting every odd-numbered point in an ordinary Poisson process with rate λ. Check that

$$\mathbb{E}(M(t)) = \tfrac{1}{4}(2\lambda t + e^{-2\lambda t} - 1), \quad \operatorname{var}(M(t)) = \tfrac{1}{16}(4\lambda t - 8\lambda t e^{-2\lambda t} - e^{-4\lambda t} + 1).$$

Deduce that $M(t)$ is not a doubly stochastic Poisson process.

(Cambridge 2011)

22. (a) Give the definition of a Poisson process $N = (N_t : t \geq 0)$ with rate λ, using its transition rates. Show that, for each $t \geq 0$, the distribution of N_t is Poisson with a parameter to be specified.

Let $J_0 = 0$ and let $J_1, J_2, \ldots$ denote the jump times of N. What is the distribution of $(J_{n+1} - J_n : n \geq 0)$? You do not need to justify your answer.

(b) Let $n \geq 1$. Compute the joint probability density function of $(J_1, J_2, \ldots, J_n)$ given $\{N_t = n\}$. Deduce that, given $\{N_t = n\}$, $(J_1, J_2, \ldots, J_n)$ has the same distribution as the non-decreasing rearrangement of n independent uniform random variables on $[0, t]$.

(c) Starting from time 0, passengers arrive on platform $9\frac{3}{4}$ at King's Cross station, with constant rate $\lambda > 0$, in order to catch a train due to depart at time $t > 0$. Using the above results, or otherwise, find the expected total time waited by all passengers (the sum of the passengers' waiting times).

(Cambridge 2012)

12
Markov chains

Summary. The chapter begins with an introduction to discrete-time Markov chains, and to the use of matrix products and linear algebra in their study. The concepts of recurrence and transience are introduced, and a necessary and sufficient criterion for recurrence is proved. This is used to derive Pólya's theorem: symmetric random walk is recurrent in one and two dimensions, and transient in higher dimensions. It is shown how to calculate hitting probabilities and hitting times. Stopping times are introduced, and the strong Markov property is presented. After a section on the classification of states, there is a discussion of invariant distributions. The convergence theorem is proved for positive recurrent chains. A criterion for time reversibility is presented, and applied in the special case of random walk on a finite graph.

12.1 The Markov property

A stochastic process is said to have the 'Markov property' if, conditional on its present value, its future is independent of its past. This is a very restrictive assumption, but it has two benefits. First, many processes in nature may be thus modelled, and secondly, the mathematical theory of such processes is strikingly beautiful and complete.

Let S be a countable set called the *state space*, and let $\mathbf{X} = (X_n : n \geq 0)$ be a sequence of random variables taking values in S. The X_n are functions on some common probability space, but we shall not be specific about that. The following is an informal way of explaining what it means to be a Markov chain: the sequence $\mathbf{X}$ is a Markov chain if, conditional on the present value X_n, the future $(X_r : r > n)$ is independent of the past $(X_m : m < n)$.

Definition 12.1 *The sequence* $\mathbf{X}$ *is called a* ***Markov chain*** *if it satisfies the* ***Markov property***

$$\mathbb{P}\big(X_{n+1} = i_{n+1} \,\big|\, X_0 = i_0,\ X_1 = i_1, \ldots,\ X_n = i_n\big) = \mathbb{P}\big(X_{n+1} = i_{n+1} \,\big|\, X_n = i_n\big) \tag{12.2}$$

for all $n \geq 0$ *and all* $i_0, i_1, \ldots, i_{n+1} \in S$. *The Markov chain is called* ***homogeneous*** *if, for all* $i, j \in S$, *the conditional probability* $\mathbb{P}(X_{n+1} = j \mid X_n = i)$ *does not depend on the value of* n.

Here are some examples of Markov chains. Each has a coherent theory relying on an assumption of independence tantamount to the Markov property.

(a) **Branching processes.** The branching process of Chapter 9 is a simple model of the growth of a population. Each member of the nth generation has a number of offspring that is independent of the past.
(b) **Random walk.** A particle performs a random walk on the line, as in Chapter 10. At each epoch of time, it jumps a random distance that is independent of previous jumps.
(c) **Poisson process.** The Poisson process of Section 11.2 satisfies a Markov property in which time is a *continuous* variable rather than a discrete variable, and thus the Poisson process is an example of a *continuous-time* Markov chain. The Markov property holds since arrivals after time t are independent of arrivals before t.
(d) **Markov chain Monte Carlo.** Here is an important example of the use of Markov chains in statistics. Whereas classical statistics results in a simple *estimate* of an unknown parameter, Bayesian statistics results in a *distribution*. Such 'posterior' distributions are often complicated, and it can be difficult to extract information. The Markov chain Monte Carlo method works as follows. First, construct a Markov chain with the posterior π as its so-called invariant distribution. Secondly, simulate this chain for a sufficiently long time that the outcome is 'nearly' distributed as π.

The basic theory of Markov chains is presented in this chapter. For simplicity, *all Markov chains here will be assumed to be homogeneous*. In order to calculate probabilities associated with such a chain, we need to know two quantities:

(a) the *transition matrix* $P = (p_{i,j} : i, j \in S)$ given by $p_{i,j} = \mathbb{P}(X_1 = j \mid X_0 = i)$, and
(b) the *initial distribution* $\lambda = (\lambda_i : i \in S)$ given by $\lambda_i = \mathbb{P}(X_0 = i)$.

By the assumption of homogeneity,

$$\mathbb{P}(X_{n+1} = j \mid X_n = i) = p_{i,j} \qquad \text{for } n \geq 0.$$

The pair (λ, P) is characterized as follows.

Proposition 12.3

(a) *The vector* λ *is a* ***distribution*** *in that* $\lambda_i \geq 0$ *for* $i \in S$, *and* $\sum_{i \in S} \lambda_i = 1$.
(b) *The matrix* $P = (p_{i,j})$ *is a* ***stochastic matrix*** *in that*
  (i) $p_{i,j} \geq 0$ *for* $i, j \in S$, *and*
  (ii) $\sum_{j \in S} p_{i,j} = 1$ *for* $i \in S$, *so that* P *has row sums* 1.

Proof (a) Since λ_i is a probability, it is non-negative. Also,

$$\sum_{i \in S} \lambda_i = \sum_{i \in S} \mathbb{P}(X_0 = i) = \mathbb{P}(X_0 \in S) = 1.$$

(b) Since $p_{i,j}$ is a probability, it is non-negative. Finally,

$$\begin{aligned} \sum_{j \in S} p_{i,j} &= \sum_{j \in S} \mathbb{P}(X_1 = j \mid X_0 = i) \\ &= \mathbb{P}(X_1 \in S \mid X_0 = i) = 1. \end{aligned}$$

□

The following will be useful later.

Theorem 12.4 *Let λ be a distribution and P a stochastic matrix. The random sequence $\mathbf{X} = (X_n : n \geq 0)$ is a Markov chain with initial distribution λ and transition matrix P if and only if*

$$\mathbb{P}\big(X_0 = i_0,\ X_1 = i_1,\ \ldots,\ X_n = i_n\big) = \lambda_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n} \tag{12.5}$$

for all $n \geq 0$ and $i_0, i_1, \ldots, i_n \in S$.

Proof Write A_k for the event $\{X_k = i_k\}$, so that (12.5) may be written as

$$\mathbb{P}(A_0 \cap A_1 \cap \cdots \cap A_n) = \lambda_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n}. \tag{12.6}$$

Suppose $\mathbf{X}$ is a Markov chain with initial distribution λ and transition matrix P. We prove (12.6) by induction on n. It holds trivially when $n = 0$. Suppose N (≥ 1) is such that (12.6) holds for $n < N$. Then

$$\begin{aligned}\mathbb{P}(A_0 \cap A_1 \cap \cdots \cap A_N) &= \mathbb{P}(A_0 \cap A_1 \cap \cdots \cap A_{N-1})\mathbb{P}\big(A_N \,\big|\, A_0 \cap A_1 \cap \cdots \cap A_{N-1}\big) \\ &= \mathbb{P}(A_0 \cap A_1 \cap \cdots \cap A_{N-1})\mathbb{P}(A_N \mid A_{N-1})\end{aligned}$$

by the Markov property. Now $\mathbb{P}(A_N \mid A_{N-1}) = p_{i_{N-1},i_N}$, and the induction step is complete.

Suppose conversely that (12.6) holds for all n and sequences (i_m). Setting $n = 0$, we obtain the initial distribution $\mathbb{P}(X_0 = i_0) = \lambda_{i_0}$. Now,

$$\mathbb{P}\big(A_{n+1} \,\big|\, A_0 \cap A_1 \cap \cdots \cap A_n\big) = \frac{\mathbb{P}(A_0 \cap A_1 \cap \cdots \cap A_{n+1})}{\mathbb{P}(A_0 \cap A_1 \cap \cdots \cap A_n)}$$

so that, by (12.6),

$$\mathbb{P}\big(A_{n+1} \,\big|\, A_0 \cap A_1 \cap \cdots \cap A_n\big) = p_{i_n,i_{n+1}}. \tag{12.7}$$

Since this does not depend on the states $i_0, i_1, \ldots, i_{n-1}$, $\mathbf{X}$ is a homogeneous Markov chain with transition matrix P.

The last step may be made more formal by writing

$$\mathbb{P}(A_{n+1} \mid A_n) = \frac{\mathbb{P}(A_n \cap A_{n+1})}{\mathbb{P}(A_n)}$$

and

$$\begin{aligned}\mathbb{P}(A_n \cap A_{n+1}) &= \sum_{i_0,i_1,\ldots,i_{n-1}} \mathbb{P}(A_0 \cap A_1 \cap \cdots \cap A_n \cap A_{n+1}) \\ &= \sum_{i_0,i_1,\ldots,i_{n-1}} \mathbb{P}\big(A_{n+1} \,\big|\, A_0 \cap A_1 \cap \cdots \cap A_n\big)\mathbb{P}(A_0 \cap A_1 \cap \cdots \cap A_n) \\ &= p_{i_n,i_{n+1}}\mathbb{P}(A_n),\end{aligned}$$

by (12.7). □

The Markov property (12.2) asserts in essence that the past affects the future only via the present. This is made formal in the next theorem, in which X_n is the present value, F is a future event, and H is a historical event.

Theorem 12.8 (Extended Markov property) *Let* $\mathbf{X}$ *be a Markov chain. For* $n \geq 0$*, for any event* H *given in terms of the past history* $X_0, X_1, \ldots, X_{n-1}$*, and any event* F *given in terms of the future* $X_{n+1}, X_{n+2}, \ldots$*,*

$$\mathbb{P}(F \mid X_n = i,\ H) = \mathbb{P}(F \mid X_n = i) \qquad \text{for } i \in S. \tag{12.9}$$

Proof A slight complication arises from the fact that F may depend on the *infinite* future. There is a general argument of probability theory that allows us to restrict ourselves to the case when F depends on the values of the process at only *finitely* many times, and we do not explain this here.

By the definition of conditional probability and Theorem 12.4,

$$\begin{aligned}
\mathbb{P}(F \mid X_n = i,\ H) &= \frac{\mathbb{P}(H,\ X_n = i,\ F)}{\mathbb{P}(H,\ X_n = i)} \\
&= \frac{\sum_{<n} \sum_{>n} \lambda_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i} p_{i,i_{n+1}} \cdots}{\sum_{<n} \lambda_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i}} \\
&= \sum_{>n} p_{i,i_{n+1}} p_{i_{n+1},i_{n+2}} \cdots \\
&= \mathbb{P}(F \mid X_n = i),
\end{aligned}$$

where $\sum_{<n}$ sums over all sequences $(i_0, i_1, \ldots, i_{n-1})$ corresponding to the event H, and $\sum_{>n}$ sums over all sequences $(i_{n+1}, i_{n+2}, \ldots)$ corresponding to the event F. □

Exercise 12.10 Let X_n be the greatest number shown in the first n throws of a fair six-sided die. Show that $\mathbf{X} = (X_n : n \geq 1)$ is a homogeneous Markov chain, and write down its transition probabilities.

Exercise 12.11 Let $\mathbf{X}$ and $\mathbf{Y}$ be symmetric random walks on the line $\mathbb{Z}$. Is $\mathbf{X} + \mathbf{Y}$ necessarily a Markov chain? Explain.

Exercise 12.12 A square matrix with non-negative entries is called *doubly stochastic* if all its row sums and column sums equal 1. If P is doubly stochastic, show that P^n is doubly stochastic for $n \geq 1$.

12.2 Transition probabilities

Let $\mathbf{X}$ be a Markov chain with transition matrix $P = (p_{i,j})$. The elements $p_{i,j}$ are called the *one-step transition probabilities*. More generally, the *n-step transition probabilities* are given by

$$p_{i,j}(n) = \mathbb{P}(X_n = j \mid X_0 = i),$$

and they form a matrix called the *n-step transition matrix* $P(n) = (p_{i,j}(n) : i, j \in S)$. The matrices $P(n)$ satisfy a collection of equations named after Chapman and Kolmogorov.

Theorem 12.13 (Chapman–Kolmogorov equations) *We have that*

$$p_{i,j}(m+n) = \sum_{k \in S} p_{i,k}(m)\, p_{k,j}(n)$$

for $i, j \in S$ and $m, n \geq 0$. That is to say, $P(m+n) = P(m)P(n)$.

Proof By the definition of conditional probability,

$$\begin{aligned} p_{i,j}(m+n) &= \mathbb{P}(X_{m+n} = j \mid X_0 = i) \\ &= \sum_{k \in S} \mathbb{P}\big(X_{m+n} = j \,\big|\, X_m = k,\ X_0 = i\big)\mathbb{P}(X_m = k \mid X_0 = i). \end{aligned} \tag{12.14}$$

By the extended Markov property, Theorem 12.8,

$$\mathbb{P}\big(X_{m+n} = j \,\big|\, X_m = k,\ X_0 = i\big) = \mathbb{P}(X_{m+n} = j \mid X_m = k),$$

and the claim follows. □

By the Chapman–Kolmogorov equations, Theorem 12.13, the n-step transition probabilities form a matrix $P(n) = (p_{i,j}(n))$ that satisfies $P(n) = P(1)^n = P^n$. One way of calculating the probabilities $p_{i,j}(n)$ is therefore to find the nth power of the matrix P. When the state space is finite, then so is P, and this calculation is usually done best by diagonalizing P. We illustrate this with an example.

Example 12.15 (Two-state Markov chain) Suppose $S = \{1, 2\}$ and

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix},$$

where $\alpha, \beta \in (0, 1)$. Find the n-step transition probabilities.

Solution A (by diagonalization) In order to calculate the n-step transition matrix P^n, we shall diagonalize P. The eigenvalues κ of P are the roots of the equation $\det(P - \kappa I) = 0$, which is to say that $(1-\alpha-\kappa)(1-\beta-\kappa) - \alpha\beta = 0$, with solutions

$$\kappa_1 = 1, \qquad \kappa_2 = 1 - \alpha - \beta.$$

Therefore,

$$P = U^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 1-\alpha-\beta \end{pmatrix} U$$

for some invertible matrix U. It follows that

$$P^n = U^{-1} \begin{pmatrix} 1 & 0 \\ 0 & (1-\alpha-\beta)^n \end{pmatrix} U,$$

and so

$$p_{1,1}(n) = A + B(1-\alpha-\beta)^n, \tag{12.16}$$

for some constants A, B which are found as follows. Since $p_{1,1}(0) = 1$ and $p_{1,1}(1) = 1-\alpha$, we have that $A + B = 1$ and $A + B(1-\alpha-\beta) = 1-\alpha$. Therefore,

$$A = \frac{\beta}{\alpha+\beta}, \qquad B = \frac{\alpha}{\alpha+\beta}.$$

Now, $p_{1,2}(n) = 1 - p_{1,1}(n)$, and $p_{2,2}(n)$ is found by interchanging α and β. In summary,

$$P^n = \frac{1}{\alpha+\beta}\begin{pmatrix} \beta+\alpha(1-\alpha-\beta)^n & \alpha-\alpha(1-\alpha-\beta)^n \\ \beta-\beta(1-\alpha-\beta)^n & \alpha+\beta(1-\alpha-\beta)^n \end{pmatrix}.$$

We note for future reference that

$$P^n \to \frac{1}{\alpha+\beta}\begin{pmatrix} \beta & \alpha \\ \beta & \alpha \end{pmatrix} \qquad \text{as } n \to \infty,$$

which is to say that

$$p_{i,1}(n) \to \frac{\beta}{\alpha+\beta}, \quad p_{i,2}(n) \to \frac{\alpha}{\alpha+\beta} \qquad \text{for } i = 1, 2.$$

This conclusion may be stated as follows. The distribution of X_n settles down to a limiting distribution $(\beta, \alpha)/(\alpha+\beta)$, which does not depend on the choice of initial state i. This hints at a general property of Markov chains to which we shall return in Sections 12.9–12.10.

Solution B (by difference equations) By conditioning on the value of X_n (or, alternatively, by the Chapman–Kolmogorov equations),

$$\begin{aligned} p_{1,1}(n+1) &= \mathbb{P}(X_{n+1} = 1 \mid X_0 = 1) \\ &= \mathbb{P}(X_{n+1} = 1 \mid X_n = 1)p_{1,1}(n) + \mathbb{P}(X_{n+1} = 1 \mid X_n = 2)p_{1,2}(n) \\ &= (1-\alpha)p_{1,1}(n) + \beta p_{1,2}(n) \\ &= (1-\alpha)p_{1,1}(n) + \beta(1 - p_{1,1}(n)). \end{aligned}$$

This is a difference equation with boundary condition $p_{1,1}(0) = 1$. Solving it in the usual way, we obtain (12.16). This is a neat solution when there are only two states, but the solution is more complicated when there are more than two. △

Finally, we summarize the matrix method illustrated in Example 12.15. Suppose the state space is finite, $|S| = N$, say, so that P is an $N \times N$ matrix. It is a general result for stochastic matrices that $\kappa_1 = 1$ is an eigenvalue of P, and no other eigenvalue has larger absolute value.[1] We write $\kappa_1 (= 1), \kappa_2, \ldots, \kappa_N$ for the (possibly complex) eigenvalues of P, arranged in decreasing order of absolute value. *We assume for simplicity that the κ_i are distinct*, since

[1] This is part of the so-called Perron–Frobenius theorem, for which the reader is referred to Grimmett and Stirzaker (2001, Sect. 6.6).

the diagonalization of P is more complicated otherwise. There exists an invertible matrix U such that $P = U^{-1}KU$, where K is the diagonal matrix with entries $\kappa_1, \kappa_2, \ldots, \kappa_N$. Then

$$P^n = (U^{-1}KU)^n = U^{-1}K^nU = U^{-1}\begin{pmatrix} \kappa_1^n & 0 & \cdots & 0 \\ 0 & \kappa_2^n & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \kappa_N^n \end{pmatrix} U, \tag{12.17}$$

from which the individual probabilities $p_{i,j}(n)$ may in principle be found.

The situation is considerably simpler if the chain has two further properties that will be encountered soon, namely 'irreducibility' (see Section 12.3) and 'aperiodicity' (see Definition 12.74 and Theorem 12.75). Under these conditions, by the Perron–Frobenius theorem, $\kappa_1 = 1$ is the unique eigenvalue with absolute value 1, so that $\kappa_k^n \to 0$ as $n \to \infty$, for $k \geq 2$. By (12.17), the long-run transition probabilities of the chain satisfy

$$P^n \to U^{-1}\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 \end{pmatrix} U \qquad \text{as } n \to \infty. \tag{12.18}$$

One may gain further information from (12.18) as follows. The rows of U are the normalized left eigenvectors of P, and the columns of U^{-1} are the normalized right eigenvectors. Since P is stochastic, $P\mathbf{1}' = \mathbf{1}'$, where $\mathbf{1}'$ is the column vector of ones. Therefore, the first column of U^{-1} is constant. By examining the product in (12.18), we find that $p_{i,j}(n) \to \pi_j$ for some vector $\pi = (\pi_j : j \in S)$ that does not depend on the initial state i.

Remark 12.19 (Markov chains and linear algebra) Much of the theory of Markov chains involves the manipulation of vectors and matrices. The equations are usually linear, and thus much of the subject can be phrased in the language of linear algebra. For example, if X_0 has distribution λ, interpreted as a row vector $(\lambda_i : i \in S)$, then

$$\mathbb{P}(X_1 = j) = \sum_{i \in S} \lambda_i p_{i,j} \qquad \text{for } j \in S,$$

so that the distribution of X_1 is the row vector λP. By iteration, X_2 has distribution λP^2, and so on. We therefore adopt the convention that probability distributions are by default row vectors, and they act on the left side of matrices. Thus, λ' denotes the transpose of the row vector λ, and is itself a column vector.

Exercise 12.20 Let $\mathbf{X}$ be a Markov chain with transition matrix P, and let $d \geq 1$. Show that $Y_n = X_{nd}$ defines a Markov chain with transition matrix P^d.

Exercise 12.21 A fair coin is tossed repeatedly. Show that the number H_n of heads after n tosses forms a Markov chain.

Exercise 12.22 A flea hops randomly between the vertices of a triangle. Find the probability that it is back at its starting point after n hops.

12.3 Class structure

An important element in the theory of Markov chains is the interaction between the state space S and the transition mechanism P.

Let $\mathbf{X}$ be a homogeneous Markov chain with state space S and transition matrix P. For $i, j \in S$, we say that i *leads to* j, written $i \to j$, if $p_{i,j}(n) > 0$ for some $n \geq 0$. By setting $n = 0$, we have that $i \to i$ for all $i \in S$. We write $i \leftrightarrow j$ if $i \to j$ and $j \to i$, and in this case we say that i and j *communicate*.

Proposition 12.23 *The relation $\leftrightarrow$ is an equivalence relation.*

Proof We must show that the relation $\leftrightarrow$ is reflexive, symmetric, and transitive. Since $i \to i$, we have that $i \leftrightarrow i$. The relation is trivially symmetric in that $j \leftrightarrow i$ whenever $i \leftrightarrow j$. Suppose that $i, j, k \in S$ satisfy $i \leftrightarrow j$ and $j \leftrightarrow k$. Since $i \to j$ and $j \to k$, there exist $m, n \geq 0$ such that $p_{i,j}(m) > 0$ and $p_{j,k}(n) > 0$. By the Chapman–Kolmogorov equations, Theorem 12.13,

$$\begin{aligned} p_{i,k}(m+n) &= \sum_{l \in S} p_{i,l}(m)\, p_{l,k}(n) \\ &\geq p_{i,j}(m)\, p_{j,k}(n) > 0, \end{aligned}$$

so that $i \to k$. Similarly, $k \to i$, and hence $i \leftrightarrow k$. Therefore, $\leftrightarrow$ is transitive. □

Since $\leftrightarrow$ is an equivalence relation, it has *equivalence classes*, namely the subsets of S of the form $C_i = \{j \in S : i \leftrightarrow j\}$. These classes are called *communicating classes*. The chain $\mathbf{X}$ (or the state space S) is called *irreducible* if there is a single communicating class, which is to say that $i \leftrightarrow j$ for all $i, j \in S$.

A subset $C \subseteq S$ is called *closed* if

$$i \in C,\ i \to j \quad \Rightarrow \quad j \in C. \tag{12.24}$$

If the chain ever hits a closed set C, then it remains in C forever afterwards. If a singleton set $\{i\}$ is closed, we call i an *absorbing* state.

Proposition 12.25 *A subset C of states is closed if and only if*

$$p_{i,j} = 0 \qquad \text{for } i \in C,\ j \notin C. \tag{12.26}$$

Proof Let $C \subseteq S$. If (12.26) fails, then so does (12.24), and C is not closed.

Suppose conversely that (12.26) holds. Let $k \in C, l \in S$ be such that $k \to l$. Since $k \to l$, there exists $m \geq 0$ such that $\mathbb{P}(X_m = l \mid X_0 = k) > 0$, and so there exists a sequence $k_0 (= k), k_1, \ldots, k_m (= l)$ with $p_{k_r, k_{r+1}} > 0$ for $r = 0, 1, \ldots, m-1$. By (12.26), $k_r \in C$ for all r, so that $l \in C$. Statement (12.24) follows. □

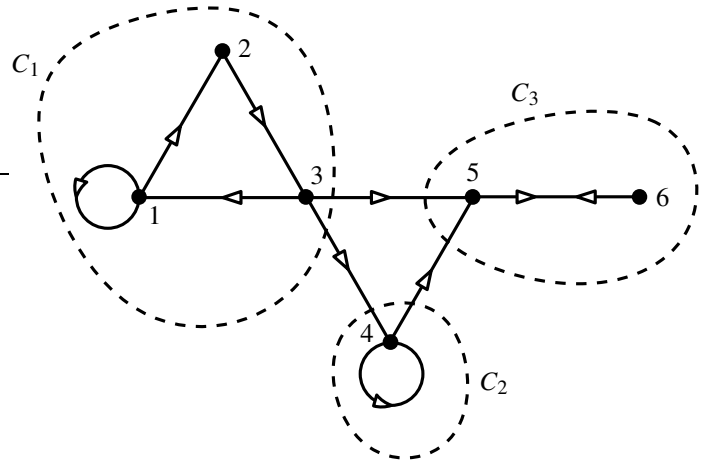


Fig. 12.1 The arrows indicate possible transitions of the chain of Example 12.27. The communicating classes are circled.

Example 12.27 Let $S = \{1, 2, 3, 4, 5, 6\}$ and

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Possible transitions of the chain are illustrated in Figure 12.1. The equivalence classes are $C_1 = \{1, 2, 3\}$, $C_2 = \{4\}$, and $C_3 = \{5, 6\}$. The classes C_1 and C_2 are not closed, but C_3 is closed. △

Exercise 12.28 Find the communicating classes, and the closed communicating classes, when the transition matrix is

$$P = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

It may be useful to draw a diagram.

Exercise 12.29 If the state space is finite, show that there must exist at least one closed communicating class. Give an example of a transition matrix with no such class.

12.4 Recurrence and transience

Let $\mathbf{X}$ be a homogeneous Markov chain with state space S and transition matrix P. For reasons of economy of notation, we write henceforth $\mathbb{P}_i(A)$ for $\mathbb{P}(A \mid X_0 = i)$, and similarly $\mathbb{E}_i(Z)$ for the conditional mean $\mathbb{E}(Z \mid X_0 = i)$.

The *first-passage time* to state j is defined as

$$T_j = \min\{n \geq 1 : X_n = j\},$$

and the *first-passage probabilities* are given by

$$f_{i,j}(n) = \mathbb{P}_i(T_j = n).$$

If a chain starts in state i, is it bound to return to i at some later time?

Definition 12.30 *A state i is called **recurrent** if $\mathbb{P}_i(T_i < \infty) = 1$. A state is called **transient** if it is not recurrent.*[2]

Here is a criterion for recurrence in terms of the transition matrix P and its powers.

Theorem 12.31 *The state i is recurrent if and only if*

$$\sum_{n=0}^{\infty} p_{i,i}(n) = \infty.$$

We saw earlier that simple random walk on the line is recurrent if and only it is unbiased (see Theorem 10.12). The proof used generating functions, and the method may be extended to prove Theorem 12.31. We introduce next the generating functions to be used in the current proof. For $i, j \in S$, let

$$P_{i,j}(s) = \sum_{n=0}^{\infty} p_{i,j}(n)s^n, \qquad F_{i,j}(s) = \sum_{n=0}^{\infty} f_{i,j}(n)s^n,$$

with the conventions that $f_{i,j}(0) = 0$ and $p_{i,j}(0) = \delta_{i,j}$, the Kronecker delta defined by

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{12.32}$$

We let

$$f_{i,j} = F_{i,j}(1) = \mathbb{P}_i(T_j < \infty), \tag{12.33}$$

and note that i is recurrent if and only if $f_{i,i} = 1$.

In proving Theorem 12.31, we shall make use of the following.

[2]The word 'persistent' is sometimes used instead of recurrent.

Theorem 12.34 *For $i, j \in S$, we have that*

$$P_{i,j}(s) = \delta_{i,j} + F_{i,j}(s)P_{j,j}(s), \qquad s \in (-1, 1].$$

Proof By conditioning on the value of T_j,

$$p_{i,j}(n) = \sum_{m=1}^{\infty} \mathbb{P}_i(X_n = j \mid T_j = m)\mathbb{P}_i(T_j = m), \qquad n \geq 1. \tag{12.35}$$

The summand is 0 for $m > n$, since in this case the first passage to j has not taken place by time n. For $m \leq n$,

$$\mathbb{P}_i(X_n = j \mid T_j = m) = \mathbb{P}_i(X_n = j \mid X_m = j,\, H),$$

where $H = \{X_r \neq j \text{ for } 1 \leq r < m\}$ is an event defined prior to time m. By homogeneity and the extended Markov property, Theorem 12.8,

$$\mathbb{P}_i(X_n = j \mid T_j = m) = \mathbb{P}(X_n = j \mid X_m = j) = \mathbb{P}_j(X_{n-m} = j).$$

We substitute this into (12.35) to obtain

$$p_{i,j}(n) = \sum_{m=1}^{n} p_{j,j}(n-m) f_{i,j}(m), \qquad n \geq 1.$$

Multiply through this equation by s^n and sum over $n \geq 1$ to obtain

$$P_{i,j}(s) - p_{i,j}(0) = P_{j,j}(s)F_{i,j}(s).$$

The claim follows since $p_{i,j}(0) = \delta_{i,j}$. □

Proof of Theorem 12.31 By Theorem 12.34 with $i = j$,

$$P_{i,i}(s) = \frac{1}{1 - F_{i,i}(s)} \qquad \text{for } |s| < 1. \tag{12.36}$$

In the limit as $s \uparrow 1$, we have by Abel's lemma[3] that

$$F_{i,i}(s) \uparrow F_{i,i}(1) = f_{i,i}, \qquad P_{i,i}(s) \uparrow \sum_{n=0}^{\infty} p_{i,i}(n).$$

By (12.36),

$$\sum_{n=0}^{\infty} p_{i,i}(n) = \infty \quad \text{if and only if} \quad f_{i,i} = 1,$$

as claimed. □

The property of recurrence is called a *class property*, in that any pair of communicating states are either both recurrent or both transient.

[3] See the footnote on p. 55.

Theorem 12.37 *Let C be a communicating class.*

(a) *Either every state in C is recurrent or every state is transient.*
(b) *Suppose C contains some recurrent state. Then C is closed.*

Proof (a) Let $i \leftrightarrow j$ and $i \neq j$. By Theorem 12.31, it suffices to show that

$$\sum_{n=0}^{\infty} p_{i,i}(n) = \infty \quad \text{if and only if} \quad \sum_{n=0}^{\infty} p_{j,j}(n) = \infty. \tag{12.38}$$

Since $i \leftrightarrow j$, there exist $m, n \geq 1$ such that

$$\alpha := p_{i,j}(m) p_{j,i}(n) > 0.$$

By the Chapman–Kolmogorov equations, Theorem 12.13,

$$p_{i,i}(m + r + n) \geq p_{i,j}(m) p_{j,j}(r) p_{j,i}(n) = \alpha p_{j,j}(r) \qquad \text{for } r \geq 0.$$

We sum over r to obtain

$$\sum_{r=0}^{\infty} p_{i,i}(m + r + n) \geq \alpha \sum_{r=0}^{\infty} p_{j,j}(r).$$

Therefore, $\sum_r p_{i,i}(r) = \infty$ whenever $\sum_r p_{j,j}(r) = \infty$. The converse holds similarly, and (12.38) is proved.

(b) Assume $i \in C$ is recurrent and C is not closed. By Proposition 12.25, there exist $j \in C$, $k \notin C$ such that $p_{j,k} > 0$. Since C is a communicating class and $k \notin C$, we have that $k \not\rightarrow j$. By part (a), j is recurrent. However,

$$\mathbb{P}_j(X_n \neq j \text{ for all } n \geq 1) \geq \mathbb{P}_j(X_1 = k) = p_{j,k} > 0,$$

a contradiction. Therefore, C is closed. □

Theorem 12.39 *Suppose that the state space S is finite.*

(a) *There exists at least one recurrent state.*
(b) *If the chain is irreducible, all states are recurrent.*

Here is a preliminary result which will be useful later.

Proposition 12.40 *Let $i, j \in S$. If j is transient, then $p_{i,j}(n) \to 0$ as $n \to \infty$.*

Proof of Proposition 12.40 Let j be transient. By Theorem 12.31 and Abel's lemma, we have that $P_{j,j}(1) < \infty$. By Theorem 12.34, $P_{i,j}(1) < \infty$, and hence the nth term in this sum, $p_{i,j}(n)$, tends to zero as $n \to \infty$. □

Proof of Theorem 12.39 Suppose $|S| < \infty$.

(a) We have that

$$1 = \sum_{j \in S} \mathbb{P}_i(X_n = j) = \sum_{j \in S} p_{i,j}(n). \tag{12.41}$$

Assume every state is transient. By Proposition 12.40, for all $j \in S$, $p_{i,j}(n) \to 0$ as $n \to \infty$. This contradicts (12.41).

(b) Suppose the chain is irreducible. By Theorem 12.37, either every state is recurrent or every state is transient, and the claim follows by part (a). □

Exercise 12.42 A Markov chain **X** has an absorbing state s to which all other states lead. Show that all states except s are transient.

Exercise 12.43

(a) Let j be a recurrent state of a Markov chain. Show that $\sum_n p_{i,j}(n) = \infty$ for all states i such that $i \to j$.

(b) Let j be a transient state of a Markov chain. Show that $\sum_n p_{i,j}(n) < \infty$ for all states i.

12.5 Random walks in one, two, and three dimensions

One-dimensional random walks were explored in some detail in Chapter 10. The purpose of the current section is to extend the theory to higher dimensions within the context of Markov chains.

The graphs in this section are d-dimensional lattices. Let $\mathbb{Z} = \{\ldots, -1, 0, 1 \ldots\}$ denote the integers, and let $\mathbb{Z}^d$ be the set of all d-vectors of integers, written $x = (x_1, x_2, \ldots, x_d)$ with each $x_i \in \mathbb{Z}$. The set $\mathbb{Z}^d$ may be interpreted as a graph with vertex set $\mathbb{Z}^d$, and with edges joining any two vectors x, y which are separated by Euclidean distance 1. Two such vertices are declared *adjacent* and are said to be *neighbours*. We denote the ensuing graph by $\mathbb{Z}^d$ also, and note that each vertex has exactly $2d$ neighbours. The graphs $\mathbb{Z}$ and $\mathbb{Z}^2$ are drawn in Figure 12.2. Note that $\mathbb{Z}^d$ is *connected* in that any given pair of vertices is joined by a path of edges.

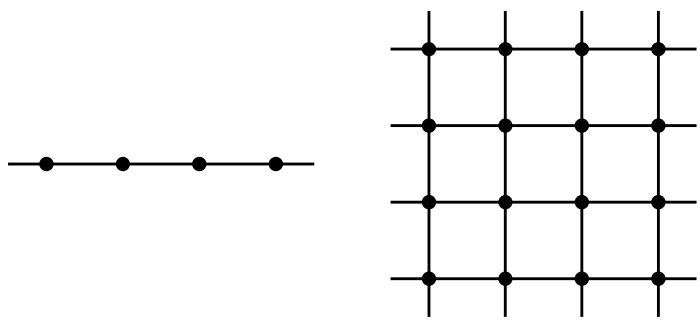


Fig. 12.2 The line $\mathbb{Z}$ and the square lattice $\mathbb{Z}^2$.

Let $d \geq 1$. The symmetric random walk on $\mathbb{Z}^d$ is the Markov chain on the state space $\mathbb{Z}^d$ which, at each step, jumps to a uniformly chosen neighbour. The transition matrix is given by

$$p_{x,y} = \begin{cases} \dfrac{1}{2d} & \text{if } y \text{ is a neighbour of } x, \\ 0 & \text{otherwise.} \end{cases}$$

Since the graph is connected, the chain is irreducible. By Theorem 12.37, either every state is recurrent or every state is transient.

Theorem 12.44 (Pólya's theorem) *The symmetric random walk on $\mathbb{Z}^d$ is recurrent if $d = 1, 2$ and transient if $d \geq 3$.*

The case $d = 1$ was proved at Theorem 10.12, and the cases $d = 2, 3$ featured in Exercise 10.11 and Problems 10.5.8 and 10.5.12.

Proof Let $d = 1$ and $X_0 = 0$. The walker can return to 0 only after an even number of steps. The probability of return after $2n$ steps is the probability that, of the first $2n$ steps, exactly n are to the right. Therefore,

$$p_{0,0}(2n) = \left(\frac{1}{2}\right)^{2n} \binom{2n}{n}. \tag{12.45}$$

By Stirling's formula, Theorem A.4,

$$p_{0,0}(2n) = \left(\frac{1}{2}\right)^{2n} \frac{(2n)!}{(n!)^2} \sim \frac{1}{\sqrt{\pi n}}. \tag{12.46}$$

In particular, $\sum_n p_{0,0}(2n) = \infty$. By Theorem 12.31, the state 0 is recurrent.

Suppose that $d = 2$. There is a clever but special way to handle this case, which we defer until after this proof. We develop instead a method that works also when $d \geq 3$. The walk is at the origin $\mathbf{0} := (0, 0)$ at time $2n$ if and only if it has taken equal numbers of leftward and rightward steps, and also equal numbers of upward and downward steps. Therefore,

$$p_{\mathbf{0},\mathbf{0}}(2n) = \left(\frac{1}{4}\right)^{2n} \sum_{m=0}^{n} \frac{(2n)!}{[m!\,(n-m)!]^2}.$$

Now,

$$\sum_{m=0}^{n} \frac{(2n)!}{[m!\,(n-m)!]^2} = \binom{2n}{n} \sum_{m=0}^{n} \binom{n}{m}\binom{n}{n-m} = \binom{2n}{n}^2,$$

by (A.2). Therefore,

$$p_{\mathbf{0},\mathbf{0}}(2n) = \left(\frac{1}{2}\right)^{4n} \binom{2n}{n}^2. \tag{12.47}$$

This is simply the square of the one-dimensional answer (12.45) (this is no coincidence), so that

$$p_{\mathbf{0},\mathbf{0}}(2n) \sim \frac{1}{\pi n}. \tag{12.48}$$

Therefore, $\sum_n p_{\mathbf{0},\mathbf{0}}(2n) = \infty$, and hence $\mathbf{0}$ is recurrent.

Suppose finally that $d = 3$, the general case $d \geq 3$ is handled similarly. By the argument that led to (12.47), and a little reorganization,

$$\begin{aligned} p_{\mathbf{0},\mathbf{0}}(2n) &= \left(\frac{1}{6}\right)^{2n} \sum_{i+j+k=n} \frac{(2n)!}{(i!\,j!\,k!)^2} \\ &= \left(\frac{1}{2}\right)^{2n} \binom{2n}{n} \sum_{i+j+k=n} \left(\frac{n!}{3^n i!\,j!\,k!}\right)^2 \\ &\leq \left(\frac{1}{2}\right)^{2n} \binom{2n}{n} M \sum_{i+j+k=n} \frac{n!}{3^n i!\,j!\,k!}, \end{aligned} \tag{12.49}$$

where

$$M = \max\left\{\frac{n!}{3^n i!\,j!\,k!} : i, j, k \geq 0,\ i + j + k = n\right\}.$$

It is not difficult to see that the maximum M is attained when i, j, and k are all closest to $\frac{1}{3}n$, so that

$$M \leq \frac{n!}{3^n\left(\lfloor \frac{1}{3}n \rfloor!\right)^3}.$$

Furthermore, the final summation in (12.49) equals 1, since the summand is the probability that, in allocating n balls randomly to three urns, the urns contain i, j, and k balls, respectively. It follows that

$$p_{\mathbf{0},\mathbf{0}}(2n) \leq \frac{(2n)!}{12^n n!\left(\lfloor \frac{1}{3}n \rfloor!\right)^3}$$

which, by Stirling's formula, is no bigger than $Cn^{-\frac{3}{2}}$ for some constant C. Therefore,

$$\sum_{n=0}^{\infty} p_{\mathbf{0},\mathbf{0}}(2n) < \infty,$$

implying that the origin $\mathbf{0}$ is transient. □

This section closes with an account of the 'neat' way of studying the two-dimensional random walk (see also Problem 10.5.9). It is the precisely 'squared' form of (12.47) that suggests an explanation using independence. Write $X_n = (A_n, B_n)$ for the position of the walker at time n, and let $Y_n = (U_n, V_n)$, where

$$U_n = A_n - B_n, \quad V_n = A_n + B_n.$$

Thus, Y_n is derived from X_n by referring to a rotated and rescaled coordinate system, as illustrated in Figure 12.3.

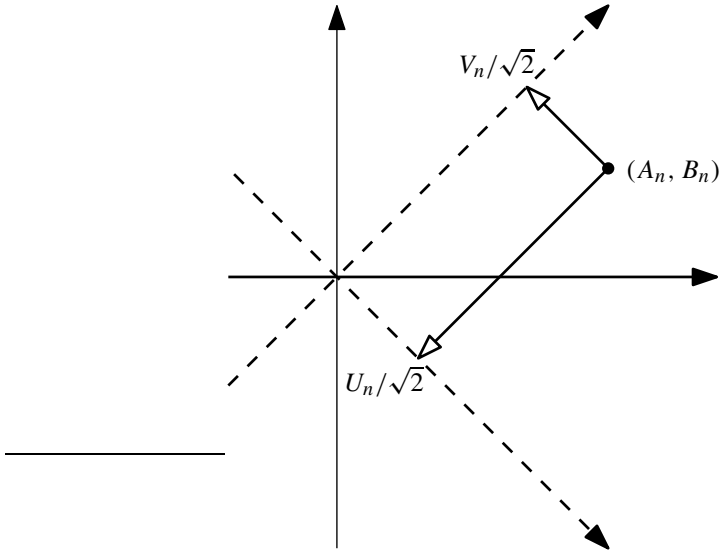


Fig. 12.3 The new coordinate system for the process $\mathbf{Y} = (Y_n)$.

The key fact is that $\mathbf{U} = (U_n)$ and $\mathbf{V} = (V_n)$ are independent, symmetric random walks on the line $\mathbb{Z}$. This is checked by a set of four calculations of the following type. First,

$$\begin{aligned}\mathbb{P}\big(Y_{n+1} - Y_n = (1,1)\big) &= \mathbb{P}(A_{n+1} - A_n = 1) \\ &= \mathbb{P}\big(X_{n+1} - X_n = (1,0)\big) = \tfrac{1}{4},\end{aligned}$$

and similarly for the other three possibilities for $Y_{n+1} - Y_n$, namely $(-1, 1)$, $(1, -1)$, and $(-1, -1)$. It follows that $\mathbf{U}$ and $\mathbf{V}$ are symmetric random walks. Furthermore, they are independent since

$$\mathbb{P}\big(Y_{n+1} - Y_n = (u, v)\big) = \mathbb{P}(U_{n+1} - U_n = u)\mathbb{P}(V_{n+1} - V_n = v) \qquad \text{for } u, v = \pm 1.$$

Finally, $X_n = \mathbf{0}$ if and only if $Y_n = \mathbf{0}$, and this occurs if and only if both $U_n = 0$ and $V_n = 0$. Therefore, in agreement with (12.47),

$$p_{\mathbf{0},\mathbf{0}}(2n) = \mathbb{P}_0(U_n = 0)\mathbb{P}_0(V_n = 0) = \left[\left(\frac{1}{2}\right)^{2n}\binom{2n}{n}\right]^2,$$

by (12.45). The corresponding argument is invalid in three or more dimensions.

Exercise 12.50 The infinite binary tree T is the tree-graph in which every vertex has exactly three neighbours. Show that a random walk on T is transient.

Exercise 12.51 Consider the asymmetric random walk on the line $\mathbb{Z}$ that moves one step rightwards with probability p, or one step leftwards with probability q $(= 1 - p)$. Show that the walk is recurrent if and only if $p = \frac{1}{2}$.

Exercise 12.52 In a variant of Exercise 12.51, the walker moves two steps rightwards with probability p, and otherwise one step leftwards. Show that the walk is recurrent if and only if $p = \frac{1}{3}$.

12.6 Hitting times and hitting probabilities

Let $A \subseteq S$. The *hitting time* of the subset A is the earliest epoch n of time at which $X_n \in A$:

$$H^A = \inf\{n \geq 0 : X_n \in A\}. \tag{12.53}$$

The infimum of an empty set is taken by convention to be ∞, so that H^A takes values in the extended integers $\{0, 1, 2, \dots\} \cup \{\infty\}$. Note that $H^A = 0$ if $X_0 \in A$.

In this section, we study the *hitting probability*

$$h_i^A = \mathbb{P}_i(H^A < \infty)$$

of ever hitting A starting from i, and also the mean value of H^A. If A is closed, then h_i^A is called an *absorption probability*.

Theorem 12.54 *The vector $h^A = (h_i^A : i \in S)$ is the minimal non-negative solution to the equations*

$$h_i^A = \begin{cases} 1 & \text{for } i \in A, \\ \displaystyle\sum_{j \in S} p_{i,j} h_j^A & \text{for } i \notin A. \end{cases} \tag{12.55}$$

In saying that h^A is the *minimal* non-negative solution, we mean the following: for any non-negative solution $(x_i : i \in S)$ of (12.55), we have that $h_i^A \leq x_i$ for all $i \in S$. The vector $h^A = (h_i^A)$ multiplies P on its right side in (12.55), and is therefore best considered as a column vector.

Proof We show first that the hitting probabilities satisfy (12.55). Certainly $h_i^A = 1$ for $i \in A$, since $H^A = 0$ in this case. For $i \notin A$, we condition on the first step of the chain to obtain

$$h_i^A = \sum_{j \in S} p_{i,j} \mathbb{P}_i(H^A < \infty \mid X_1 = j) = \sum_{j \in S} p_{i,j} h_j^A$$

as required for (12.55).

We show next that the h_i^A are minimal. Let $x = (x_i : i \in S)$ be a non-negative solution to (12.55). In particular, $h_i^A = x_i = 1$ for $i \in A$. Let $i \notin A$. Since x satisfies (12.55),

$$x_i = \sum_{j \in S} p_{i,j} x_j = \sum_{j \in A} p_{i,j} x_j + \sum_{j \notin A} p_{i,j} x_j. \tag{12.56}$$

Since $x_j = 1$ for $j \in A$, and x is non-negative, we have that

$$x_i \geq \sum_{j \in A} p_{i,j}$$
$$= \mathbb{P}_i(X_1 \in A) = \mathbb{P}_i(H^A = 1).$$

We iterate this as follows. By expanding the final summation in (12.56),

$$x_i = \mathbb{P}_i(X_1 \in A) + \sum_{j \notin A} p_{i,j} \left(\sum_{k \in A} p_{j,k} x_k + \sum_{k \notin A} p_{j,k} x_k \right)$$
$$\geq \mathbb{P}_i(X_1 \in A) + \mathbb{P}_i(X_1 \notin A,\ X_2 \in A)$$
$$= \mathbb{P}_i(H^A \leq 2).$$

By repeated substitution, we obtain $x_i \geq \mathbb{P}_i(H^A \leq n)$ for all $n \geq 0$. Take the limit as $n \to \infty$ to deduce as required that $x_i \geq \mathbb{P}_i(H^A < \infty) = h_i^A$. □

We turn now to the mean hitting times, and we write

$$k_i^A = \mathbb{E}_i(H^A),$$

noting that $k_i^A = \infty$ if $\mathbb{P}_i(H^A = \infty) > 0$.

Theorem 12.57 *The vector $k^A = (k_i^A : i \in S)$ is the minimal non-negative solution to the equations*

$$k_i^A = \begin{cases} 0 & \text{for } i \in A, \\ 1 + \sum_{j \in S} p_{i,j} k_j^A & \text{for } i \notin A. \end{cases} \tag{12.58}$$

Proof This is very similar to the last proof. We show first that the k_i^A satisfy (12.58). Certainly, $k_i^A = 0$ for $i \in A$, since $H^A = 0$ in this case. For $i \notin A$, we condition on the first step of the chain to obtain

$$k_i^A = \sum_{j \in S} p_{i,j} [1 + \mathbb{E}_j(H^A)] = 1 + \sum_{j \in S} p_{i,j} k_j^A$$

as required for (12.58).

We show next that the k_i^A are minimal. Let $y = (y_i : i \in S)$ be a non-negative solution to (12.58). In particular, $k_i^A = y_i = 0$ for $i \in A$. Let $i \notin A$. Since y satisfies (12.58),

$$y_i = 1 + \sum_{j \in S} p_{i,j} y_j = 1 + \sum_{j \notin A} p_{i,j} y_j$$
$$= 1 + \sum_{j \notin A} p_{i,j} \left(1 + \sum_{k \notin A} p_{j,k} y_k \right)$$
$$\geq \mathbb{P}_i(H^A \geq 1) + \mathbb{P}_i(H^A \geq 2).$$

By iteration,

$$y_i \geq \sum_{m=1}^{n} \mathbb{P}_i(H^A \geq m) \qquad \text{for } n \geq 1,$$

and we send $n \to \infty$ to obtain

$$y_i \geq \sum_{m=1}^{\infty} \mathbb{P}_i(H^A \geq m) = k_i^A,$$

as required. We have used the elementary fact that $\mathbb{E}M = \sum_{m=1}^{\infty} \mathbb{P}(M \geq m)$ for a random variable M taking non-negative integer values (see Problem 2.6.6). □

Example 12.59 (Gambler's ruin) Let S be the non-negative integers $\{0, 1, 2, \dots\}$, and $p \in (0, 1)$. A random walk on S moves one unit rightwards with probability p, and one unit leftwards with probability q $(= 1 - p)$, and has an absorbing barrier at 0. Find the probability of ultimate absorption from a given starting point.

Solution Let h_i be the probability of absorption starting at i. By Theorem 12.54, (h_i) is the minimal non-negative solution to the equations

$$h_0 = 1, \qquad h_i = ph_{i+1} + qh_{i-1} \quad \text{for } i \geq 1.$$

Suppose $p \neq q$. The difference equation has general solution

$$h_i = A + B(q/p)^i \qquad \text{for } i \geq 0.$$

If $p < q$, the boundedness of the h_i forces $B = 0$, and the fact $h_0 = 1$ implies $A = 1$. Therefore, $h_i = 1$ for all $i \geq 0$.

Suppose $p > q$. Since $h_0 = 1$, we have $A + B = 1$, so that

$$h_i = (q/p)^i + A\big(1 - (q/p)^i\big).$$

Since $h_i \geq 0$, we have $A \geq 0$. By the minimality of the h_i, we have $A = 0$, and hence $h_i = (q/p)^i$, in agreement with Theorem 10.32.

Suppose finally that $p = q = \frac{1}{2}$. The difference equation has solution

$$h_i = A + Bi,$$

and the above arguments yield $B = 0$, $A = 1$, so that $h_i = 1$. △

Example 12.60 (Birth–death chain) Let $(p_i : i \geq 1)$ be a sequence of numbers satisfying $p_i = 1 - q_i \in (0, 1)$. The above gambler's ruin example may be extended as follows. Let $\mathbf{X}$ be a Markov chain on $\{0, 1, 2, \dots\}$ with transition probabilities

$$p_{i,i+1} = p_i, \quad p_{i,i-1} = q_i \qquad \text{for } i \geq 1,$$

and $p_{0,0} = 1$. What is the probability of ultimate absorption at 0, having started at i?

Solution As in Example 12.59, the required probabilities h_i are the minimal non-negative solutions of

$$h_0 = 1, \qquad h_i = p_i h_{i+1} + q_i h_{i-1} \qquad \text{for } i \geq 1.$$

Set $u_i = h_{i-1} - h_i$ and reorganize this equation to obtain that

$$u_{i+1} = (q_i/p_i)u_i \qquad \text{for } i \geq 1,$$

so that $u_{i+1} = \gamma_i u_1$, where

$$\gamma_i = \frac{q_1 q_2 \cdots q_i}{p_1 p_2 \cdots p_i}.$$

Now, $u_1 + u_2 + \cdots + u_i = h_0 - h_i$, so that

$$h_i = 1 - u_1(\gamma_0 + \gamma_1 + \cdots + \gamma_{i-1}) \qquad \text{for } i \geq 1,$$

where $\gamma_0 = 1$. It remains to determine the constant u_1.

There are two situations. Suppose first that $S = \sum_{k=0}^{\infty} \gamma_k$ satisfies $S = \infty$. Since $h_i \geq 0$ for all i, we have that $u_1 = 0$, and therefore $h_i = 1$ for all i. On the other hand, if $S < \infty$, the h_i are minimized when $1 - u_1 S = 0$, which is to say that $u_1 = 1/S$ and

$$h_i = \sum_{k=i}^{\infty} \gamma_k \Big/ \sum_{k=0}^{\infty} \gamma_k \qquad \text{for } i \geq 0.$$

Thus, $h_i < 1$ for $i \geq 1$ if and only if $S < \infty$. △

Exercise 12.61 Let $\mathbf{X}$ be a Markov chain on the non-negative integers $\{0, 1, 2, \dots\}$ with transition probabilities satisfying

$$p_{0,1} = 1, \quad p_{i,i+1} + p_{i,i-1} = 1, \quad p_{i,i+1} = p_{i,i-1}\left(\frac{i+1}{i}\right)^2 \quad \text{for } i \geq 1.$$

Show that $\mathbb{P}_0(X_n \geq 1 \text{ for all } n \geq 1) = 6/\pi^2$. You may use the fact that $\sum_{k=1}^{\infty} k^{-2} = \frac{1}{6}\pi^2$.

Exercise 12.62 Consider Exercise 12.61 with the difference that

$$p_{i,i+1} = p_{i,i-1}\left(\frac{i+1}{i}\right)^{\alpha} \quad \text{for } i \geq 1,$$

where $\alpha > 0$. Find the probability $\mathbb{P}_0(X_n \geq 1 \text{ for all } n \geq 1)$ in terms of α.

12.7 Stopping times and the strong Markov property

The Markov property of Definition 12.1 requires that, conditional on the value of the chain at a given time n, the future evolution of the chain is independent of its past. We frequently require an extension of this property to a *random* time n. It is not hard to see that the Markov property cannot be true for *all* random times, and it turns out that the appropriate times are those satisfying the following definition.

Definition 12.63 *A random variable $T : \Omega \to \{0, 1, 2, \dots\} \cup \{\infty\}$ is called a **stopping time** for the chain $\mathbf{X}$ if, for all $n \geq 0$, the event $\{T = n\}$ is given in terms of $X_0, X_1, \dots, X_n$ only.*

That is to say, a random time T is a stopping time if you can tell whether it equals any given time by examining only the present and past of the chain. Random times that 'look into the future' are not stopping times.

The principal examples of stopping times are the hitting times of Section 12.6. Let $A \subseteq S$, and consider the hitting time H^A given in (12.53). Note that

$$\{H^A = n\} = \{X_n \in A\} \cap \left(\bigcap_{0 \leq m < n} \{X_m \notin A\} \right),$$

so that H^A is indeed a stopping time: one can tell whether or not $H^A = n$ by examining $X_0, X_1, \dots, X_n$ only.

Two related examples: it is easily checked that $T = H^A + 1$ is a stopping time, and that $T = H^A - 1$ is not. See Exercise 12.70 for a further example.

Theorem 12.64 (Strong Markov property) *Let $\mathbf{X}$ be a Markov chain with transition matrix P, and let T be a stopping time. Given that $T < \infty$ and $X_T = i$, the sequence $\mathbf{Y} = (Y_k : k \geq 0)$, given by $Y_k = X_{T+k}$, is a Markov chain with transition matrix P and initial state $Y_0 = i$. Furthermore, given that $T < \infty$ and $X_T = i$, $\mathbf{Y}$ is independent of $X_0, X_1, \dots, X_{T-1}$.*

Proof Let H be an event given in terms of $X_0, X_1, \dots, X_{T-1}$. It is required to show that

$$\begin{aligned} \mathbb{P}\big(X_{T+1} = i_1, X_{T+2} = i_2, \dots, X_{T+n} = i_n, H \,\big|\, T < \infty, X_T = i\big) \\ = \mathbb{P}_i(X_1 = i_1, X_2 = i_2, \dots, X_n = i_n)\mathbb{P}(H \mid T < \infty, X_T = i). \end{aligned} \tag{12.65}$$

The event $H \cap \{T = m\}$ is given in terms of $X_1, X_2, \dots, X_m$ only. Furthermore, $X_T = X_m$ when $T = m$. We condition on the event $H \cap \{T = m\} \cap \{X_m = i\}$ and use the Markov property (12.9) at time m to deduce that

$$\begin{aligned} \mathbb{P}(X_{T+1} = i_1, X_{T+2} = i_2, \dots, X_{T+n} = i_n, H, T = m, X_T = i) \\ = \mathbb{P}_i(X_1 = i_1, X_2 = i_2, \dots, X_n = i_n)\mathbb{P}(H, T = m, X_T = i). \end{aligned}$$

Now sum over $m = 0, 1, 2, \dots$ and divide by $\mathbb{P}(T < \infty, X_T = i)$ to obtain (12.65). □

Example 12.66 (Gambler's ruin) Let S be the non-negative integers $\{0, 1, 2, \dots\}$, and $p \in (0, 1)$. Consider a random walk $\mathbf{X}$ on S which moves one step rightwards with probability p, one step leftwards with probability q $(= 1 - p)$, and with an absorbing barrier at 0. Let H be the time until absorption at 0. Find the distribution (and mean) of H given $X_1 = 1$.

Solution We shall work with the probability generating function of H. A problem arises since it may be the case that $\mathbb{P}_1(H = \infty) > 0$. One may either work with the conditional generating function $\mathbb{E}_1(s^H \mid H < \infty)$, or, equivalently, we can use the fact that, when $|s| < 1$, $s^n \to 0$ as $n \to \infty$. That is, we write

$$G(s) = \mathbb{E}_1(s^H) = \sum_{n=0}^{\infty} s^n \mathbb{P}_1(H = n) \qquad \text{for } |s| < 1,$$

valid regardless of whether or not $\mathbb{P}_1(H = \infty) = 0$. Henceforth, we assume that $|s| < 1$, and later we shall use Abel's lemma to take the limit as $s \uparrow 1$.[4]

By conditioning on the first step of the walk, we find that

$$G(s) = p\mathbb{E}_1(s^H \mid X_1 = 2) + q\mathbb{E}_1(s^H \mid X_1 = 0).$$

Let us consider the first expectation. The first step is from state 1 to state 2. Having arrived at state 2, we require the probability generating function of the first-passage time to state 0. This is the sum of the first-passage time (denoted H') to state 1 plus the consequent passage time (denoted H'') to state 0. The random variables H' and H'' have the same distribution as H. Furthermore, by conditioning on H' and using the strong Markov property, H' and H'' are independent. It follows that

$$\begin{aligned}\mathbb{E}_1(s^H \mid X_1 = 2) &= \mathbb{E}_1(s^{1+H'+H''}) \\ &= s\mathbb{E}_2(s^{H'})\mathbb{E}_1(s^{H''}) = sG(s)^2.\end{aligned}$$

Therefore,

$$G(s) = psG(s)^2 + qs. \tag{12.67}$$

This is a quadratic in $G(s)$ with solutions

$$G(s) = \frac{1 \pm \sqrt{1 - 4pqs^2}}{2ps}. \tag{12.68}$$

Since G is continuous wherever it is finite, we must choose one of these solutions and stick with it for all $|s| < 1$. Since $G(0) \leq 1$ and the positive root diverges as $s \downarrow 0$, we take the negative root in (12.68) for all $|s| < 1$.

The mass function of H is obtained from the coefficients in the expansion of $G(s)$ as a power series:

$$\mathbb{P}_1(H = 2k - 1) = \binom{1/2}{k}(-1)^{k-1}\frac{(4pq)^k}{2p} = \frac{(2k-2)!}{k!\,(k-1)!} \cdot \frac{(pq)^k}{p},$$

for $k = 1, 2, \ldots$. This uses the extended binomial theorem, Theorem A.3.

[4] See the footnote on p. 55 for a statement of Abel's lemma.

It is not certain that $H < \infty$. Since $\mathbb{P}_1(H < \infty) = \lim_{s\uparrow 1} G(s)$, we have by (12.68) and Abel's lemma that

$$\mathbb{P}_1(H < \infty) = \frac{1 - \sqrt{1 - 4pq}}{2p}.$$

It is convenient to write

$$1 - 4pq = 1 - 4p + 4p^2 = (1 - 2p)^2 = |p - q|^2,$$

so that

$$\mathbb{P}_1(H < \infty) = \frac{1 - |p - q|}{2p} = \begin{cases} 1 & \text{if } p \le q, \\ q/p & \text{if } p > q, \end{cases}$$

as in Theorem 10.32 and Example 12.59.

We turn to the mean value $\mathbb{E}_1(H)$. When $p > q$, $\mathbb{P}_1(H = \infty) > 0$, and so $\mathbb{E}_1(H) = \infty$. Suppose $p \le q$. By differentiating (12.67),

$$pG^2 + 2psGG' - G' + q = 0 \qquad \text{for } |s| < 1, \tag{12.69}$$

which we solve for G' to find that

$$G'(s) = \frac{pG(s)^2 + q}{1 - 2psG(s)} \qquad \text{for } |s| < 1.$$

By Abel's lemma, $\mathbb{E}_1(H) = \lim_{s\uparrow 1} G'(s)$, so that

$$\mathbb{E}_1(H) = \lim_{s\uparrow 1} \left(\frac{pG^2 + q}{1 - 2psG} \right) = \begin{cases} \infty & \text{if } p = q, \\ \dfrac{1}{q - p} & \text{if } p < q. \end{cases} \qquad \triangle$$

Exercise 12.70

(a) Let H^A be the hitting time of the set A. Show that $T = H^A - 1$ is not generally a stopping time.

(b) Let L^A be the time of the last visit of a Markov chain to the set A, with the convention that $L^A = \infty$ if infinitely many visits are made. Show that L^A is not generally a stopping time.

12.8 Classification of states

We saw in Definition 12.30 that a state i is *recurrent* if, starting from i, the chain returns to i with probability 1. The state is *transient* if it is not recurrent. If the starting state i is recurrent, the chain is bound to return to it. Indeed, it is bound to return infinitely often.

Theorem 12.71 *Suppose $X_0 = i$, and let $V_i = |\{n \geq 1 : X_n = i\}|$ be the number of subsequent visits by the Markov chain to i. Then V_i has the geometric distribution*

$$\mathbb{P}_i(V_i = r) = (1 - f) f^r \qquad \text{for } r = 0, 1, 2, \ldots, \tag{12.72}$$

where $f = f_{i,i}$ is the return probability, $f_{i,i} = \mathbb{P}_i(X_n = i \text{ for some } n \geq 1)$. In particular,

(a) $\mathbb{P}_i(V_i = \infty) = 1$ *if i is recurrent,*
(b) $\mathbb{P}_i(V_i < \infty) = 1$ *if i is transient.*

We return in Theorem 12.105 to the more detailed question of the rate of divergence of the number of visits to i in the recurrent case. The proof makes use of the recurrence time of a state i. Let

$$T_i = \inf\{n \geq 1 : X_n = i\} \tag{12.73}$$

be the first-passage time to i. If $X_0 = i$, then T_i is the *recurrence time* of i, with mean $\mu_i = \mathbb{E}_i(T_i)$.

Proof Let $f_{i,i} = \mathbb{P}_i(T_i < \infty)$, so that i is recurrent if $f_{i,i} = 1$ and transient if $f_{i,i} < 1$. Let T_i^r be the epoch of the rth visit to i, with $T_i^r = \infty$ if $V_i < r$. Since the T_i^r are increasing,

$$\begin{aligned}
\mathbb{P}_i(V_i \geq r) &= \mathbb{P}_i(T_i^r < \infty) \\
&= \mathbb{P}_i\big(T_i^r < \infty \,\big|\, T_i^{r-1} < \infty\big)\mathbb{P}_i(T_i^{r-1} < \infty) \\
&= f_{i,i}\,\mathbb{P}_i(T_i^{r-1} < \infty) \qquad \text{for } r \geq 1,
\end{aligned}$$

by the strong Markov property, Theorem 12.64. By iteration, $\mathbb{P}_i(V_i \geq r) = f_{i,i}^r$, as required for (12.72). We send $r \to \infty$ to find that

$$\mathbb{P}_i(V_i = \infty) = \begin{cases} 1 & \text{if } f_{i,i} = 1, \\ 0 & \text{if } f_{i,i} < 1, \end{cases}$$

and the theorem is proved. □

Definition 12.74

(a) *The **mean recurrence time** μ_i of the state i is defined by*

$$\mu_i = \mathbb{E}_i(T_i) = \begin{cases} \sum_{n=1}^{\infty} n f_{i,i}(n) & \text{if } i \text{ is recurrent,} \\ \infty & \text{if } i \text{ is transient.} \end{cases}$$

(b) *If i is recurrent, we call it **null** if $\mu_i = \infty$, and **positive** (or **non-null**) if $\mu_i < \infty$.*
(c) *The **period** d_i of the state i is given by*[5]

$$d_i = \gcd\{n : p_{i,i}(n) > 0\}.$$

*The state i is called **aperiodic** if $d_i = 1$, and **periodic** if $d_i > 1$.*
(d) *State i is called **ergodic** if it is aperiodic and positive recurrent.*

It was proved in Theorem 12.37 that recurrence is a class property. This conclusion may be extended as follows.

Theorem 12.75 *If $i \leftrightarrow j$, then*

(a) *i and j have the same period,*
(b) *i is recurrent if and only if j is recurrent,*
(c) *i is positive recurrent if and only if j is positive recurrent,*
(d) *i is ergodic if and only if j is ergodic.*

We may therefore speak of a communicating class C as being recurrent, transient, ergodic, and so on. An irreducible chain has a single communicating class, and thus we may attribute these adjectives (when appropriate) to the chain itself.

Proof We may assume $i \neq j$.
(a) Since $i \leftrightarrow j$, there exist $m, n \geq 1$ such that

$$\alpha := p_{i,j}(m) p_{j,i}(n) > 0.$$

By the Chapman–Kolmogorov equations, Theorem 12.13,

$$p_{i,i}(m+r+n) \geq p_{i,j}(m) p_{j,j}(r) p_{j,i}(n) = \alpha p_{j,j}(r) \qquad \text{for } r \geq 0. \tag{12.76}$$

In particular, $p_{i,i}(m+n) \geq \alpha > 0$, so that $d_i \mid m+n$. Therefore, if $d_i \nmid r$, then $d_i \nmid m+r+n$, so that $p_{i,i}(m+n+r) = 0$. In this case, by (12.76), $p_{j,j}(r) = 0$, and hence $d_j \nmid r$. Therefore, $d_i \mid d_j$. By the reverse argument, $d_j \mid d_i$, and hence $d_i = d_j$.
(b) This was proved at Theorem 12.37.
(c) For this proof, we look ahead slightly to Theorem 12.83. Suppose that i is positive recurrent, and let C be the communicating class of states containing i. Since i is recurrent, by

[5] The greatest common divisor of the set N is denoted $\gcd\{N\}$.

Theorem 12.37(b), C is closed. If $X_0 \in C$, then $X_n \in C$ for all n, and the chain is irreducible on the state space C. By part (a) of Theorem 12.83, it possesses an invariant distribution, and by part (b) every state (of C) is positive recurrent. If $i \leftrightarrow j$ then $j \in C$, so j is positive recurrent.

(d) This follows from (a), (b), and (c). □

Finally in this section, we note that recurrent states are visited regardless of the initial distribution. This will be useful later.

Proposition 12.77 *If the chain is irreducible and $j \in S$ is recurrent, then*

$$\mathbb{P}(X_n = j \text{ for some } n \geq 1) = 1,$$

regardless of the distribution of X_0.

Proof Let i, j be distinct states and recall the passage probability

$$f_{i,j} = \mathbb{P}_i(X_n = j \text{ for some } n \geq 1).$$

Since the chain is irreducible, there exists a least integer m (≥ 1) such that $p_{j,i}(m) > 0$. Since m is least, it is the case that

$$p_{j,i}(m) = \mathbb{P}_j(X_m = i,\ X_r \neq j \text{ for } 1 \leq r < m). \tag{12.78}$$

Suppose $X_0 = j$, $X_m = i$, and no return to j takes place after time m. By (12.78), with conditional probability 1 no return to j ever takes place. It follows by the Markov property at time m that

$$p_{j,i}(m)(1 - f_{i,j}) \leq 1 - f_{j,j}.$$

If j is recurrent, then $f_{j,j} = 1$, so that $f_{i,j} = 1$ for all $i \in S$.

Let $\lambda_i = \mathbb{P}(X_0 = i)$ for $i \in S$. With $T_j = \inf\{n \geq 1 : X_n = j\}$ as usual,

$$\mathbb{P}(T_j < \infty) = \sum_{i \in S} \lambda_i f_{i,j} = 1,$$

by conditioning on X_0. □

Exercise 12.79 Let $\mathbf{X}$ be an irreducible Markov chain with period d. Show that $Y_n = X_{nd}$ defines an aperiodic Markov chain.

Exercise 12.80 Let $0 < p < 1$. Classify the states of the Markov chains with transition matrices

$$\begin{pmatrix} 0 & p & 0 & 1-p \\ 1-p & 0 & p & 0 \\ 0 & 1-p & 0 & p \\ p & 0 & 1-p & 0 \end{pmatrix}, \qquad \begin{pmatrix} 1-2p & 2p & 0 \\ p & 1-2p & p \\ 0 & 2p & 1-2p \end{pmatrix}.$$

Exercise 12.81 Let i be an aperiodic state of a Markov chain. Show that there exists $N \geq 1$ such that $p_{i,i}(n) > 0$ for all $n \geq N$.

12.9 Invariant distributions

We turn now towards the study of the long-term behaviour of a Markov chain: what can be said about X_n in the limit as $n \to \infty$? Since the sequence $(X_n : n \geq 0)$ is subject to random fluctuations, it does not (typically) converge to any given state. On the other hand, we will see in the next section that its distribution settles into an equilibrium. In advance of stating this limit theorem, we first explore the possible limits. Any distributional limit is necessarily invariant under the evolution of the chain, and we are led to the following definition.

Definition 12.82 *Let* $\mathbf{X}$ *be a Markov chain with transition matrix* P*. The vector* $\pi = (\pi_i : i \in S)$ *is called an* ***invariant distribution***[6] *of the chain if:*

(a) $\pi_i \geq 0$ *for all* $i \in S$*, and* $\sum_{i \in S} \pi_i = 1$,
(b) $\pi = \pi P$.

An invariant distribution is invariant under the passage of time: if X_0 has distribution π, then X_n has distribution πP^n, and $\pi P^n = \pi P \cdot P^{n-1} = \pi P^{n-1} = \cdots = \pi$, so that every X_n has distribution π.

Theorem 12.83 *Consider an irreducible Markov chain.*

(a) *There exists an invariant distribution* π *if and only if some state is positive recurrent.*
(b) *If there exists an invariant distribution* π*, then every state is positive recurrent, and*

$$\pi_i = \frac{1}{\mu_i} \qquad \text{for } i \in S,$$

where μ_i *is the mean recurrence time of state* i*. In particular,* π *is the unique invariant distribution.*

We shall prove Theorem 12.83 by exhibiting an explicit solution of the vector equation $\rho = \rho P$. In looking for a solution, it is natural to consider a vector ρ with entries indicative of the proportions of time spent in the various states. Towards this end, we fix a state $k \in S$ and start the chain from this state. Let W_i be the number of subsequent visits to state i before the first return to the initial state k. Thus, W_i may be expressed in either of the forms

$$W_i = \sum_{m=1}^{\infty} 1(X_m = i,\ T_k \geq m) = \sum_{m=1}^{T_k} 1(X_m = i), \qquad i \in S, \tag{12.84}$$

where $T_k = \inf\{n \geq 1 : X_n = k\}$ is the first return time to the starting state k, and $1(A) = 1_A$ is the indicator function of A. Note that $W_k = 1$ if $T_k < \infty$. Our candidate for the vector ρ is given by

$$\rho_i = \mathbb{E}_k(W_i), \qquad i \in S. \tag{12.85}$$

Recall that $\mathbb{E}_k(Z)$ denotes the mean of Z given that $X_0 = k$.

[6] Also known as a *stationary*, or *equilibrium*, or *steady-state* distribution. An invariant distribution is sometimes referred to as an *invariant measure*, but it is more normal to reserve this expression for a non-negative solution π of the equation $\pi = \pi P$ with no assumption of having sum 1, or indeed of even having finite sum.

Proposition 12.86 *For an irreducible, recurrent chain, and any given $k \in S$, the vector $\rho = (\rho_i : i \in S)$ satisfies:*

(a) $\rho_k = 1$,
(b) $\sum_{i \in S} \rho_i = \mu_k$, *regardless of whether or not* $\mu_k < \infty$,
(c) $\rho = \rho P$,
(d) $0 < \rho_i < \infty$ *for* $i \in S$.

Here is a useful consequence of the last theorem and proposition. Consider an irreducible, positive recurrent Markov chain, and fix a state k. Since the chain is positive, we have $\mu_k < \infty$. By Theorem 12.83, there exists a unique invariant distribution π, and $\pi_k = 1/\mu_k$. By Proposition 12.86(b, c), $\nu := \pi_k \rho$ satisfies $\nu = \nu P$ and $\sum_{i \in S} \nu_i = 1$. By the uniqueness of the invariant distribution, $\pi = \nu$. Therefore, $\rho_i = \nu_i / \pi_k = \pi_i / \pi_k$. We state this conclusion as a corollary.

Corollary 12.87 *Let $i, k \in S$ be distinct states of an irreducible, positive recurrent Markov chain with invariant distribution π. The mean number of visits to state i between two consecutive visits to k equals π_i / π_k.*

Proof of Proposition 12.86 (a) Since the chain is assumed recurrent, $\mathbb{P}_k(T_k < \infty) = 1$. By (12.84), $W_k = 1$, so that $\rho_k = \mathbb{E}_k(1) = 1$.

(b) Since the time between two visits to state k must be spent somewhere, we have that

$$T_k = \sum_{i \in S} W_i,$$

where W_i is given by (12.84). By an interchange of expectation and summation,[7]

$$\mu_k = \mathbb{E}_k(T_k) = \sum_{i \in S} \mathbb{E}_k(W_i) = \sum_{i \in S} \rho_i.$$

(c) By (12.84) and a further interchange, for $j \in S$,

$$\rho_j = \sum_{m=1}^{\infty} \mathbb{P}_k(X_m = j,\ T_k \geq m). \tag{12.88}$$

The event $\{T_k \geq m\}$ depends only on $X_0, X_1, \ldots, X_{m-1}$. By the extended Markov property, Theorem 12.8, for $m \geq 1$,

$$\begin{aligned} \mathbb{P}_k(X_m = j,\ T_k \geq m) &= \sum_{i \in S} \mathbb{P}_k(X_{m-1} = i,\ X_m = j,\ T_k \geq m) \\ &= \sum_{i \in S} \mathbb{P}_k\big(X_m = j \,\big|\, X_{m-1} = i,\ T_k \geq m\big) \mathbb{P}_k(X_{m-1} = i,\ T_k \geq m) \\ &= \sum_{i \in S} p_{i,j} \mathbb{P}_k(X_{m-1} = i,\ T_k \geq m). \end{aligned} \tag{12.89}$$

[7] Care is necessary when interchanging limits. This interchange is justified by the footnote on p. 40. The forthcoming interchange at (12.90) holds since the order of summation is irrelevant to the value of a double sum of non-negative reals.

By (12.88)–(12.89) and another interchange of limits,

$$\rho_j = \sum_{i \in S} \sum_{m=1}^{\infty} p_{i,j} \mathbb{P}_k(X_{m-1} = i,\ T_k \geq m). \tag{12.90}$$

We rewrite this with $r = m - 1$ to find that

$$\rho_j = \sum_{i \in S} p_{i,j} \sum_{r=0}^{\infty} \mathbb{P}_k(X_r = i,\ T_k \geq r + 1) = \sum_{i \in S} p_{i,j} \rho_i,$$

where the last equality holds by separate consideration of the two cases $i = k$ and $i \neq k$. In summary, $\rho = \rho P$.

(d) We shall use the fact that $\rho_k = 1$. Since the chain is irreducible, there exist $m, n \geq 0$ such that $p_{i,k}(m),\ p_{k,i}(n) > 0$. Since $\rho = \rho P$ and hence $\rho = \rho P^r$ for $r \geq 1$, we have that

$$\rho_k \geq \rho_i\, p_{i,k}(m), \qquad \rho_i \geq \rho_k\, p_{k,i}(n).$$

Since $\rho_k = 1$,

$$p_{k,i}(m) \leq \rho_i \leq \frac{1}{p_{i,k}(n)},$$

and the proof is complete. □

Proof of Theorem 12.83 (a) Suppose $k \in S$ is positive recurrent, so that $\mu_k < \infty$. Let ρ be given by (12.85). By Proposition 12.86, $\pi := (1/\mu_k)\rho$ is an invariant distribution.

(b) Suppose that π is an invariant distribution of the chain. We show first that

$$\pi_i > 0 \qquad \text{for } i \in S. \tag{12.91}$$

Since $\sum_{i \in S} \pi_i = 1$, there exists $k \in S$ with $\pi_k > 0$. Let $i \in S$. By irreducibility, there exists $m \geq 1$ such that $p_{k,i}(m) > 0$. We have that $\pi = \pi P$, so that $\pi = \pi P^m$. Therefore,

$$\pi_i = \sum_{j \in S} \pi_j\, p_{j,i}(m) \geq \pi_k\, p_{k,i}(m) > 0,$$

and (12.91) follows.

By irreducibility and Theorem 12.37, either all states are transient or all are recurrent. If all states are transient, then $p_{i,j}(n) \to 0$ as $n \to \infty$ by Proposition 12.40. Since $\pi = \pi P^n$,

$$\pi_j = \sum_i \pi_i\, p_{i,j}(n) \to 0 \quad \text{as} \quad n \to \infty, \qquad \text{for } i, j \in S, \tag{12.92}$$

which contradicts (12.91). Therefore, all states are recurrent. A small argument is needed to justify the limit in (12.92) when S is infinite, and this is deferred to Lemma 12.95.

We show next that the existence of π implies that all states are positive, and that $\pi_i = \mu_i^{-1}$ for $i \in S$. Suppose that X_0 has distribution π. Since π is invariant,

$$(X_1, X_2, \ldots, X_{n-1}) \text{ and } (X_0, X_1, \ldots, X_{n-2}) \text{ have the same joint distributions.} \quad (12.93)$$

Now,[8]

$$\pi_i \mu_i = \mathbb{P}(X_0 = i) \sum_{n=1}^{\infty} \mathbb{P}_i(T_i \geq n) = \sum_{n=1}^{\infty} \mathbb{P}(T_i \geq n,\ X_0 = i).$$

However, $\mathbb{P}(T_i \geq 1,\ X_0 = i) = \mathbb{P}(X_0 = i)$, and for $n \geq 2$,

$$\begin{aligned}
\mathbb{P}(T_i \geq n,\ X_0 = i) &= \mathbb{P}(X_0 = i,\ X_m \neq i \text{ for } 1 \leq m \leq n-1) \\
&= \mathbb{P}(X_m \neq i \text{ for } 1 \leq m \leq n-1) - \mathbb{P}(X_m \neq i \text{ for } 0 \leq m \leq n-1) \\
&= \mathbb{P}(X_m \neq i \text{ for } 0 \leq m \leq n-2) - \mathbb{P}(X_m \neq i \text{ for } 0 \leq m \leq n-1) \\
&= a_{n-2} - a_{n-1}
\end{aligned}$$

by (12.93), where

$$a_r = \mathbb{P}(X_m \neq i \text{ for } 0 \leq m \leq r).$$

We sum over n to obtain

$$\pi_i \mu_i = \mathbb{P}(X_0 = i) + a_0 - \lim_{n \to \infty} a_n = 1 - \lim_{n \to \infty} a_n.$$

However, $a_n \to \mathbb{P}(X_m \neq i \text{ for all } m) = 0$ as $n \to \infty$, by the recurrence of i and Proposition 12.77.

We have shown that

$$\pi_i \mu_i = 1, \quad (12.94)$$

so that $\mu_i = \pi_i^{-1} < \infty$ by (12.91). Hence $\mu_i < \infty$, and all states of the chain are positive. Furthermore, (12.94) specifies π_i uniquely as μ_i^{-1}. □

Here is the little lemma used to establish the limit in (12.92). It is a form of the so-called bounded convergence theorem.

Lemma 12.95 *Let $\lambda = (\lambda_i : i \in S)$ be a distribution on the countable set S. Let $\alpha_i(n)$ satisfy $|\alpha_i(n)| \leq M < \infty$ for all $i \in S$ and all n, and in addition*

$$\lim_{n \to \infty} \alpha_i(n) = 0 \qquad \text{for } i \in S.$$

Then

$$\sum_{i \in S} \lambda_i \alpha_i(n) \to 0 \qquad \text{as } n \to \infty.$$

[8] If T takes values in the non-negative integers, $\mathbb{E}(T) = \sum_{n=1}^{\infty} \mathbb{P}(T \geq n)$.

Proof Let F be a finite subset of S, and write

$$\begin{aligned}\sum_{i\in S}|\lambda_i\alpha_i(n)| &\le \sum_{i\in F}\lambda_i|\alpha_i(n)| + M\sum_{i\notin F}\lambda_i \\ &\to M\sum_{i\notin F}\lambda_i \qquad \text{as } n\to\infty, \text{ since } F \text{ is finite} \\ &\to 0 \qquad \text{as } F\uparrow S, \text{ since } \sum_{i\in S}\lambda_i < \infty.\end{aligned}$$

□

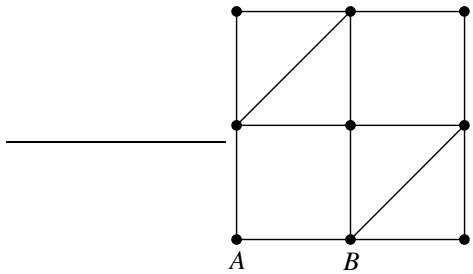


Fig. 12.4 Find the mean number of visits to B before returning to the starting state A.

Exercise 12.96 A particle starts at A and executes a symmetric random walk on the graph of Figure 12.4. At each step it moves to a neighbour of the current vertex chosen uniformly at random. Find the invariant distribution of the chain. Using the remark after Proposition 12.86 or otherwise, find the expected number of visits to B before the particle returns to A.

Exercise 12.97 Consider the symmetric random walk on the line $\mathbb{Z}$. Show that any invariant distribution π satisfies $\pi_n = \frac{1}{2}(\pi_{n-1} + \pi_{n+1})$, and deduce that the walk is null recurrent.

12.10 Convergence to equilibrium

The principal result for discrete-time Markov chains is that, subject to reasonable conditions, its distribution converges to the unique invariant distribution.

Theorem 12.98 (Convergence theorem for Markov chains) *Consider a Markov chain that is aperiodic, irreducible, and positive recurrent. For $i, j \in S$,*

$$p_{i,j}(n) \to \pi_j \qquad \text{as } n\to\infty,$$

where π is the unique invariant distribution of the chain.

Proof The proof uses an important technique known as 'coupling'. Construct an ordered pair $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ of *independent* Markov chains $\mathbf{X} = (X_n : n \ge 0)$, $\mathbf{Y} = (Y_n : n \ge 0)$,

each of which has state space S and transition matrix P. Then $\mathbf{Z} = (Z_n : n \geq 0)$ is given by $Z_n = (X_n, Y_n)$, and it is easy to check that $\mathbf{Z}$ is a Markov chain with state space $S \times S$ and transition probabilities

$$\begin{aligned} p_{ij,kl} &= \mathbb{P}\big(Z_{n+1} = (k,l) \,\big|\, Z_n = (i,j)\big) \\ &= \mathbb{P}(X_{n+1} = k \mid X_n = i)\mathbb{P}(Y_{n+1} = l \mid Y_n = j) \quad \text{by independence} \\ &= p_{i,k}\, p_{j,l}. \end{aligned}$$

Since $\mathbf{X}$ is irreducible and aperiodic, for $i, j, k, l \in S$ there exists $N = N(i, j, k, l)$ such that $p_{i,k}(n)\, p_{j,l}(n) > 0$ for all $n \geq N$ (see Exercise 12.81). Therefore, $\mathbf{Z}$ is irreducible. Only here is the aperiodicity used.

Suppose that $\mathbf{X}$ is positive recurrent. By Theorem 12.83, $\mathbf{X}$ has a unique stationary distribution π, and it is follows that $\mathbf{Z}$ has the stationary distribution $\nu = (\nu_{i,j} : i, j \in S)$ given by $\nu_{i,j} = \pi_i \pi_j$. Therefore, $\mathbf{Z}$ is also positive recurrent, by Theorem 12.83. Let $X_0 = i$ and $Y_0 = j$, so that $Z_0 = (i, j)$. Fix $s \in S$ and let

$$T = \min\{n \geq 1 : Z_n = (s, s)\}$$

be the first-passage time of $\mathbf{Z}$ to (s, s). By the recurrence of Z and Proposition 12.77,

$$\mathbb{P}_{ij}(T < \infty) = 1, \tag{12.99}$$

where $\mathbb{P}_{ij}$ denotes the probability measure conditional on $Z_0 = (i, j)$.

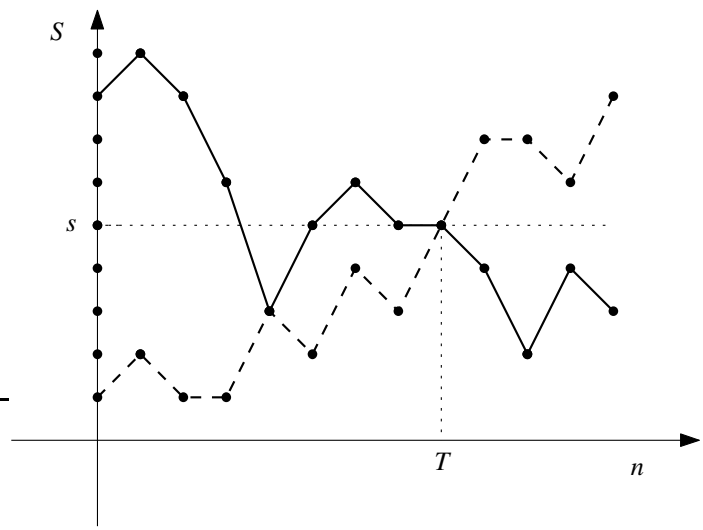


Fig. 12.5 The two chains $\mathbf{X}$ and $\mathbf{Y}$ evolve until the first time T at which both chains are simultaneously in state s. Conditional on the event $\{T \leq n\}$, X_n and Y_n have the same distribution.

The central idea of the proof is the following observation, illustrated in Figure 12.5. Since T is a stopping time, by the strong Markov property, X_n and Y_n have the same conditional

distributions given the event $\{T \le n\}$. We shall use this fact, together with the finiteness of T, to show that the limiting distributions of X and Y are independent of their starting points.

More precisely,

$$\begin{aligned} p_{i,k}(n) &= \mathbb{P}_{ij}(X_n = k) \\ &= \mathbb{P}_{ij}(X_n = k,\ T \le n) + \mathbb{P}_{ij}(X_n = k,\ T > n) \\ &= \mathbb{P}_{ij}(Y_n = k,\ T \le n) + \mathbb{P}_{ij}(X_n = k,\ T > n) \\ &\qquad \text{since, given that } T \le n,\ X_n \text{ and } Y_n \text{ are identically distributed} \\ &\le \mathbb{P}_{ij}(Y_n = k) + \mathbb{P}_{ij}(T > n) \\ &= p_{j,k}(n) + \mathbb{P}_{ij}(T > n), \end{aligned}$$

where we have used the strong Markov property. This, and the related inequality with i and j interchanged, yields

$$\big|p_{i,k}(n) - p_{j,k}(n)\big| \le \mathbb{P}_{ij}(T > n) \to 0 \qquad \text{as } n \to \infty$$

by (12.99). Therefore,

$$p_{i,k}(n) - p_{j,k}(n) \to 0 \quad \text{as } n \to \infty, \qquad \text{for } i, j, k \in S. \tag{12.100}$$

Thus, if the limit $\lim_{n\to\infty} p_{ik}(n)$ exists, then it does not depend on the choice of i. To show that the limit exists, write

$$\pi_k - p_{j,k}(n) = \sum_{i\in S} \pi_i \big[p_{i,k}(n) - p_{j,k}(n)\big] \to 0 \qquad \text{as } n \to \infty, \tag{12.101}$$

by Lemma 12.95. The proof is complete. □

Example 12.102 Here is an elementary example which highlights the necessity of aperiodicity in the convergence theorem, Theorem 12.98. Let $\mathbf{X}$ be a Markov chain with state space $S = \{1, 2\}$ and transition matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Thus, $\mathbf{X}$ alternates deterministically between the two states. It is immediate that $P^{2m} = I$ and $P^{2m+1} = P$ for $m \ge 0$, and, in particular, the limit $\lim_{n\to\infty} p_{i,j}(n)$ exists for no $i, j \in S$.

The proof of Theorem 12.98 fails since the paired chain $\mathbf{Z}$ is not irreducible: for example, if $Z_0 = (0, 1)$, then $Z_n \ne (0, 0)$ for all n. △

Example 12.103 (Coupling game) A pack of playing cards is shuffled, and the cards dealt (face up) one by one. A friend is asked to select some card, secretly, from amongst the first six or seven cards, say. If the face value of this card is m (aces count 1 and court cards count 10), the next $m - 1$ cards are allowed to pass, and your friend is asked to note the face value of the mth card. Continuing according to this rule, there arrives a last card in this sequence, with face value X, say, and with fewer than X cards remaining. We call X your friend's 'score'.

With high probability, you are able to guess accurately your friend's score, as follows. You follow the same rules as the friend, starting for simplicity at the first card. You obtain thereby a score Y, say. There is a high probability that $X = Y$.

Why is this the case? Suppose your friend picks the m_1th card, m_2th card, and so on, and you pick the n_1 $(= 1)$th, n_2th, ... If $m_i = n_j$ for some i, j, the two of you are 'stuck together' forever after. When this occurs first, we say that 'coupling' has occurred. Prior to coupling, each time you read the value of a card, there is a positive probability that you will arrive at the next stage on exactly the same card as the other person. If the pack of cards were infinitely large, then coupling would take place sooner or later. It turns out that there is a reasonable chance that coupling takes place before the last card of a regular pack of 52 cards has been dealt. △

A criterion for transience or recurrence was presented at Theorem 12.31. We now have a criterion for null recurrence.

Theorem 12.104 *Let* $\mathbf{X}$ *be an irreducible, recurrent Markov chain. The following are equivalent.*

(a) *There exists a state* i *such that* $p_{i,i}(n) \to 0$ *as* $n \to \infty$.
(b) *Every state is null recurrent.*

As an application, consider a symmetric random walk on the graphs $\mathbb{Z}$ or $\mathbb{Z}^2$ of Section 12.5. By (12.46) or (12.48) as appropriate, $p_{\mathbf{0},\mathbf{0}}(n) \to 0$ as $n \to \infty$, from which we deduce that the one- and two-dimensional random walks are null recurrent. This may be compared with the method of Exercise 12.97.

Proof We shall prove only that (a) implies (b). See Grimmett and Stirzaker (2001, Thm 6.2.9) for the other part. If the chain $\mathbf{X}$ is positive recurrent and, in addition, aperiodic, then

$$p_{i,i}(n) \to \frac{1}{\mu_i} > 0,$$

by Theorems 12.83 and 12.98. Therefore, (a) does not hold. The same argument may be applied in the periodic case by considering the chain $Y_n = X_{nd}$ where d is the period of the chain. Thus (a) implies (b). □

This section closes with a discussion of the long-run proportion of times at which a Markov chain is in a given state. Let $i \in S$ and let

$$V_i(n) = \sum_{k=1}^{n} 1(X_k = i)$$

denote the number of visits to i up to time n. Recall from Definition 8.45 that $Z_n \Rightarrow Z$ means Z_n converges to Z in distribution.

Theorem 12.105 *Let $i \in S$. If the chain is irreducible and positive recurrent,*

$$\frac{1}{n} V_i(n) \Rightarrow \frac{1}{\mu_i} \qquad \text{as } n \to \infty,$$

irrespective of the initial distribution of the chain.

There are various modes of convergence of random variables, of which we have chosen convergence in distribution for the sake of simplicity. (It is equivalent to convergence in probability in this case, see Theorem 8.47.) A more powerful result is valid, but it relies on the so-called strong law of large numbers, which is beyond the range of this volume.

Proof The law of large numbers tells us about the asymptotic behaviour of the sum of independent, identically distributed random variables, and the key to the current proof is to write $V_i(n)$ in terms of such a sum. Let

$$U_1 = \inf\{n \geq 1 : X_n = i\}$$

be the time until the first visit to i, and for $m \geq 1$, let U_m be the time between the mth and $(m+1)$th visits. Since the chain is assumed positive recurrent, we have that $\mathbb{P}(U_m < \infty) = 1$ and $\mu_i = \mathbb{E}_i(U_1) < \infty$. The *first* passage time U_1 may have a different distribution from the remaining U_m if $X_0 \neq i$.

By the strong Markov property, the random variables $U_1, U_2, \ldots$ are independent, and $U_2, U_3, \ldots$ are identically distributed. Moreover,

$$V_i(n) \geq x \quad \text{if and only if} \quad S_{\lceil x \rceil} \leq n,$$

where $\lceil x \rceil$ is the least integer not less than x, and

$$S_m = \sum_{r=1}^{m} U_r$$

is the time of the mth visit to i. Therefore,

$$\mathbb{P}\left(\frac{1}{n} V_i(n) \geq \frac{1+\epsilon}{\mu_i}\right) = \mathbb{P}\left(S_N \leq n\right), \tag{12.106}$$

where $N = \left\lceil (1+\epsilon) n/\mu_i \right\rceil$. By the weak law of large numbers, Theorem 8.17,

$$\begin{aligned} \frac{1}{N} S_N &= \frac{1}{N} U_1 + \frac{1}{N} \sum_{r=2}^{N} U_r \\ &\Rightarrow \mu_i \qquad \text{as } n \to \infty, \end{aligned} \tag{12.107}$$

where we have used the fact that $U_1/N \to 0$ in probability (see Theorem 8.47 and Problem 8.6.9). Since $n/N \to \mu_i/(1+\epsilon)$, we have by (12.106)–(12.107) that

$$\mathbb{P}\left(\frac{1}{n}V_i(n) \geq \frac{1+\epsilon}{\mu_i}\right) = \mathbb{P}\left(\frac{1}{N}S_N \leq \frac{n}{N}\right) \to \begin{cases} 0 & \text{if } \epsilon > 0, \\ 1 & \text{if } \epsilon < 0. \end{cases}$$

There is a gap in this proof, since Theorem 8.17 assumed that a typical summand U_2, say, has finite variance. If that is not known, then it is necessary to appeal to the more powerful conclusion of Example 8.52 whose proof uses the method of characteristic functions. □

Exercise 12.108 Let π be the unique invariant distribution of an aperiodic, irreducible Markov chain $\mathbf{X}$. Show that $\mathbb{P}(X_n = j) \to \pi_j$ as $n \to \infty$, regardless of the initial distribution of X_0.

12.11 Time reversal

An important observation of physics is that many equations are valid irrespective of whether time flows forwards or backwards. Invariance under time-reversal is an important property of certain Markov chains.

Let $\mathbf{X} = (X_n : 0 \leq n \leq N)$ be an irreducible, positive recurrent Markov chain, with transition matrix P and invariant distribution π. Suppose further that X_0 has distribution π, so that X_n has distribution π for every n. The 'reversed chain' $\mathbf{Y} = (Y_n : 0 \leq n \leq N)$ is given by reversing time: $Y_n = X_{N-n}$ for $0 \leq n \leq N$. Recall from Theorem 12.83(b) that $\pi_i > 0$ for $i \in S$.

Theorem 12.109 *The sequence* $\mathbf{Y}$ *is an irreducible Markov chain with transition matrix* $\widehat{P} = (\widehat{p}_{i,j} : i, j \in S)$ *given by*

$$\widehat{p}_{i,j} = \frac{\pi_j}{\pi_i} p_{j,i} \qquad \text{for } i, j \in S, \tag{12.110}$$

and with invariant distribution π.

Proof We check first that $\widehat{P}$ is a stochastic matrix. Certainly its entries are non-negative, and also

$$\sum_{j \in S} \widehat{p}_{i,j} = \frac{1}{\pi_i} \sum_{j \in S} \pi_j p_{j,i} = \frac{1}{\pi_i} \pi_i = 1,$$

since $\pi = \pi P$.

Next we show that π is invariant for $\widehat{P}$. By (12.110),

$$\sum_{i \in S} \pi_i \widehat{p}_{i,j} = \sum_{i \in S} \pi_j p_{j,i} = \pi_j,$$

since P has row sums 1.

By Theorem 12.4,

$$\begin{aligned} \mathbb{P}\big(Y_0 = i_0,\, Y_1 = i_1,\, \ldots,\, Y_n = i_n\big) &= \mathbb{P}\big(X_{N-n} = i_n,\, X_{N-n+1} = i_{n-1},\, \ldots,\, X_N = i_0\big) \\ &= \pi_{i_n} p_{i_n, i_{n-1}} \cdots p_{i_1, i_0} \\ &= \pi_{i_0} \widehat{p}_{i_0, i_1} \cdots \widehat{p}_{i_{n-1}, i_n} \qquad \text{by (12.110).} \end{aligned}$$

By Theorem 12.4 again, $\mathbf{Y}$ has transition matrix $\widehat{P}$ and initial distribution π. □

We call the chain $\mathbf{Y}$ the *time reversal* of the chain $\mathbf{X}$, and we say that $\mathbf{X}$ is *reversible* if $\mathbf{X}$ and its time reversal have the same transition probabilities.

Definition 12.111 *Let $\mathbf{X} = (X_n : 0 \le n \le N)$ be an irreducible Markov chain such that X_0 has the invariant distribution π. The chain is **reversible** if $\mathbf{X}$ and its time reversal $\mathbf{Y}$ have the same transition matrices, which is to say that*

$$\pi_i p_{i,j} = \pi_j p_{j,i} \qquad \text{for } i, j \in S. \tag{12.112}$$

Equations (12.112) are called the *detailed balance equations*, and they are pivotal to the study of reversible chains. More generally we say that a transition matrix P and a distribution λ are *in detailed balance* if

$$\lambda_i p_{i,j} = \lambda_j p_{j,i} \qquad \text{for } i, j \in S.$$

An irreducible chain $\mathbf{X}$ with invariant distribution π is said to be *reversible in equilibrium* if its transition matrix P is in detailed balance with π.

It turns out that, for an irreducible chain, P is in detailed balance with a distribution λ if and only if λ is the unique invariant distribution. This provides a good way of finding the invariant distribution of a reversible chain.

Theorem 12.113 *Let P be the transition matrix of an irreducible chain $\mathbf{X}$, and suppose that π is a distribution satisfying*

$$\pi_i p_{i,j} = \pi_j p_{j,i} \qquad \text{for } i, j \in S. \tag{12.114}$$

Then π is the unique invariant distribution of the chain. Furthermore, $\mathbf{X}$ is reversible in equilibrium.

Proof Suppose that π is a distribution that satisfies (12.114). Then

$$\sum_{i \in S} \pi_i p_{i,j} = \sum_{i \in S} \pi_j p_{j,i} = \pi_j \sum_{i \in S} p_{j,i} = \pi_j,$$

since P has row sums 1. Therefore, $\pi = \pi P$, whence π is invariant. The reversibility in equilibrium of $\mathbf{X}$ follows by Definition 12.111. □

The above discussion of reversibility is restricted to an irreducible, positive recurrent Markov chain with only *finitely* many time points $0, 1, 2, \ldots, N$. Such a chain on the *singly infinite* time set $0, 1, 2, \ldots$ is called reversible if the finite subsequences $X_0, X_1, \ldots, X_N$ are reversible for all $N \ge 0$. The discussion may also be extended to the *doubly infinite* time set $\ldots, -2, -1, 0, 1, 2, \ldots$, subject to the assumption that X_n has the invariant distribution π for all n.

Time reversibility is a very useful concept in the theory of random networks. There is a valuable analogy using the language of flows. Let $\mathbf{X}$ be a Markov chain with state space S and invariant distribution π. To this chain there corresponds the following directed network

(or graph). The vertices of the network are the states of the chain, and an arrow is placed from vertex i to vertex j if $p_{i,j} > 0$. One unit of a notional material ('probability') is distributed about the vertices and allowed to flow along the arrows. A proportion π_i of the material is placed initially at vertex i. At each epoch of time and for each vertex i, a proportion $p_{i,j}$ of the material at i is transported to each vertex j.

It is immediate that the amount of material at vertex i after one epoch is $\sum_j \pi_j p_{j,i}$, which equals π_i since $\pi = \pi P$. That is to say, the deterministic flow of probability is in equilibrium: there is 'global balance' in the sense that the total quantity leaving each vertex is balanced by an equal quantity arriving there. There may or may not be 'local balance', in the sense that, for every $i, j \in S$, the amount flowing from i to j equals the amount flowing from j to i. Local balance occurs if and only if $\pi_i p_{i,j} = \pi_j p_{j,i}$ for $i, j \in S$, which is to say that P and π are in detailed balance.

Example 12.115 (Birth–death chain with retaining barrier) Consider a random walk $\mathbf{X} = (X_n : n \geq 0)$ on the non-negative integers $\{0, 1, 2, \dots\}$ which, when at $i \geq 1$, moves one step rightwards with probability p_i, or one step leftwards with probability q_i $(= 1 - p_i)$. When at $i = 0$, it stays at 0 with probability q_0 and otherwise moves to 1. We assume for simplicity that $0 < p_i < 1$ for all i. This process differs from the birth–death chain of Example 12.60 in its behaviour at 0.

Under what conditions on the p_i is the Markov chain $\mathbf{X}$ reversible in equilibrium? If this holds, find the invariant distribution.

Solution We look for a solution to the detailed balance equations (12.114), which may be written as

$$\pi_{i-1} p_{i-1} = \pi_i q_i \qquad \text{for } i \geq 1.$$

By iteration, the solution is

$$\pi_i = \rho_i \pi_0 \qquad \text{for } i \geq 0, \tag{12.116}$$

where $\rho_0 = 1$ and

$$\rho_i = \frac{p_{i-1} p_{i-2} \cdots p_0}{q_i q_{i-1} \cdots q_1} \qquad \text{for } i \geq 1.$$

The vector π is a distribution if and only if $\sum_i \pi_i = 1$. By (12.116),

$$\sum_{i \in S} \pi_i = \pi_0 \sum_{i \in S} \rho_i.$$

We may choose π_0 appropriately if and only if $S = \sum_i \rho_i$ satisfies $S < \infty$, in which case we set $\pi_0 = 1/S$.

By Theorem 12.113, $\mathbf{X}$ is reversible in equilibrium if and only if $S < \infty$, in which case the invariant distribution is given by $\pi_i = \rho_i / S$. △

Example 12.117 (Ehrenfest dog–flea model) Two dogs, Albert and Beatrice, are infested by a total of m fleas that jump from one dog to the other at random. We assume that, at each epoch of time, one flea, picked uniformly at random from the m available, passes from its

current host to the other dog. Let X_n be the number of fleas on Albert after n units of time has passed. Thus, $\mathbf{X} = (X_n : n \geq 0)$ is an irreducible Markov chain with transition matrix

$$p_{i,i+1} = 1 - \frac{i}{m}, \quad p_{i,i-1} = \frac{i}{m} \qquad \text{for } 0 \leq i \leq m.$$

Rather than solve the equation $\pi = \pi P$ to find the invariant distribution, we look for solutions of the detailed balance equations $\pi_i p_{i,j} = \pi_j p_{j,i}$. These equations amount to

$$\pi_{i-1}\left(\frac{m-i+1}{m}\right) = \pi_i \cdot \frac{i}{m} \qquad \text{for } 1 \leq i \leq m.$$

By iteration,

$$\pi_i = \binom{m}{i}\pi_0,$$

and we choose $\pi_0 = 2^{-m}$ so that π is a distribution. By Theorem 12.113, π is the unique invariant distribution. △

Exercise 12.118 Consider a random walk on a triangle, illustrated in Figure 12.6. The state space is $S = \{1, 2, 3\}$, and the transition matrix is

$$P = \begin{pmatrix} 0 & \alpha & 1-\alpha \\ 1-\alpha & 0 & \alpha \\ \alpha & 1-\alpha & 0 \end{pmatrix},$$

where $0 < \alpha < 1$. Show that the detailed balance equations possess a solution if and only if $\alpha = \frac{1}{2}$.

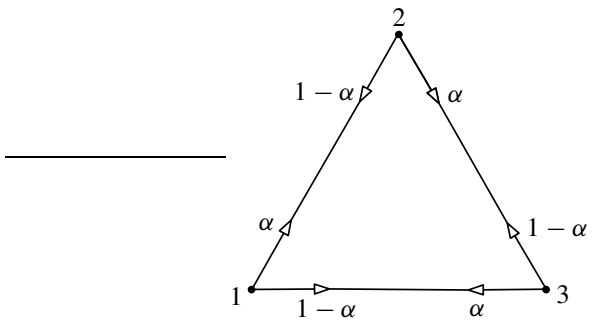


Fig. 12.6 Transition probabilities for a random walk on a triangle.

Exercise 12.119 Can a reversible Markov chain be periodic? Explain.

Exercise 12.120 A random walk moves on the finite set $\{0, 1, 2, \ldots, N\}$. When in the interior of the interval, it moves one step rightwards with probability p, or one step leftwards with probability q $(= 1 - p)$. When it is at either endpoint, 0 or N, and tries to leave the interval, it is retained at its current position. Assume $0 < p < 1$, and use the detailed balance equations to find the invariant distribution.

12.12 Random walk on a graph

A *graph* $G = (V, E)$ is a set V of vertices, pairs of which are joined by edges. That is, the edge set E is a set of distinct unordered pairs $\langle u, v\rangle$ of distinct elements of V. A graph is usually represented in the manner illustrated in Figure 12.7. The lattice graphs $\mathbb{Z}^d$ in Section 12.5 are examples of infinite graphs.

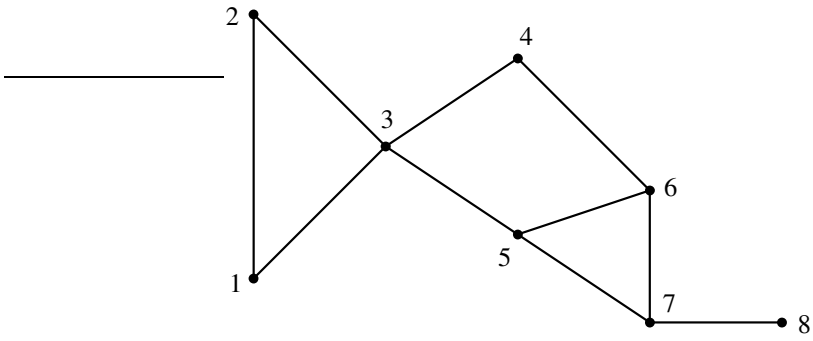


Fig. 12.7 A graph G with 8 vertices. A random walk on G moves around the vertex set. At each step, it moves to a uniformly random neighbour of its current position.

Here is some language and notation concerning graphs. A graph is *connected* if, for every distinct pair $u, v \in V$, there exists a path of edges from u to v. We write $u \sim v$ if $\langle u, v\rangle \in E$, in which case we say that u and v are *neighbours*. The *degree* $d(v)$ of vertex v is the number of edges containing v, that is, $d(v) = |\{u \in V : v \sim u\}|$.

There is a rich theory of random walks on finite and infinite graphs. Let $G = (V, E)$ be a connected graph with $d(v) < \infty$ for all $v \in V$. A particle moves about the vertices of G, taking steps along the edges. Let X_n be the position of the particle at time n. At time $n + 1$, it moves to a uniformly random neighbour of X_n. More precisely, a random walk is the Markov chain $\mathbf{X} = (X_n : n \geq 0)$ with state space V and transition matrix

$$p_{u,v} = \begin{cases} \dfrac{1}{d(u)} & \text{if } v \sim u, \\ 0 & \text{otherwise.} \end{cases} \tag{12.121}$$

When G is infinite, the main question is to understand the long-run behaviour of the walk, such as whether or not it is transient or recurrent. This was the question addressed in Section 12.5 for the lattice graphs $\mathbb{Z}^d$. In this section, we consider a *finite* connected graph G. It will be useful to note that

$$\sum_{v \in V} d(v) = 2|E|, \tag{12.122}$$

since each edge contributes 2 to the summation.

Theorem 12.123 *Random walk on the finite connected graph $G = (V, E)$ is an irreducible Markov chain with unique invariant distribution*

$$\pi_v = \frac{d(v)}{2|E|} \qquad \text{for } v \in V.$$

The chain is reversible in equilibrium.

Proof Since G is connected, the chain is irreducible. The vector π is certainly a distribution since $\pi_v \geq 0$ for $v \in V$, and $\sum_{v \in V} \pi_v = 1$ by (12.122). By Theorem 12.113, it suffices to check the detailed balance equations (12.114), namely

$$\frac{d(u)}{2|E|} p_{u,v} = \frac{d(v)}{2|E|} p_{v,u}, \qquad \text{for } u, v \in V.$$

This holds by the definition (12.121) of the transition probabilities. □

Example 12.124 (Erratic knights) A knight is the sole inhabitant of a chess board, and it performs random moves. Each move is chosen at random from the set of currently permissible moves, as illustrated in Figure 12.8. What is the invariant distribution of the Markov chain describing the knight's motion?

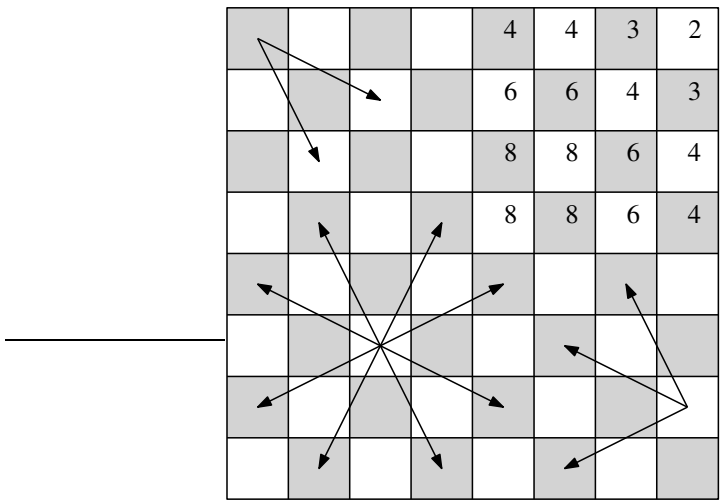


Fig. 12.8 A map for the erratic knight. The arrows indicate permissible moves. If the knight is at a square from which there are m permissible moves, then it selects one of these with equal probability $1/m$. The numbers are the degrees of the corresponding graph vertices.

Solution Let $G = (V, E)$ be the graph given as follows. The vertex set V is the set of squares of the chess board, and the edge set E is given as follows: two vertices u, v are joined by an edge if and only if the move between u and v is a legal knight-move. The knight performs

a random walk on G. In order to find the invariant distribution, we must count the vertex degrees. The four corners have degree 2, and so on, as indicated in the upper right corner of Figure 12.8. The sum of the vertex degrees is

$$\sum_{v \in V} d(v) = 4 \cdot 2 + 8 \cdot 3 + 20 \cdot 4 + 16 \cdot 6 + 16 \cdot 8 = 336,$$

and the invariant distribution is given by Theorem 12.123 as $\pi_v = d(v)/336$. △

Exercise 12.125 An erratic king performs random (but legal) moves on a chess board. Find his invariant distribution.

12.13 Problems

1. A transition matrix is called *doubly stochastic* if its column sums equal 1, that is, if $\sum_{i \in S} p_{i,j} = 1$ for $j \in S$.

   Suppose an irreducible chain with N ($< \infty$) states has a doubly stochastic transition matrix. Find its invariant distribution. Deduce that all states are positive recurrent and that, if the chain is aperiodic, then $p_{i,j}(n) \to 1/N$ as $n \to \infty$.

2. Let $\mathbf{X}$ be a discrete-time Markov chain with state space $S = \{1, 2\}$ and transition matrix

   $$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

   Classify the states of the chain. Suppose that $0 < \alpha\beta < 1$. Find the n-step transition probabilities and show directly that they converge to the unique invariant distribution. For what values of α and β is the chain reversible in equilibrium?

3. We distribute N black balls and N white balls in two urns in such a way that each contains N balls. At each epoch of time, one ball is selected at random from each urn, and these two balls are interchanged. Let X_n be the number of black balls in the first urn after time n. Write down the transition matrix of this Markov chain, and find the unique invariant distribution. Is the chain reversible in equilibrium?

4. Consider a Markov chain on the set $S = \{0, 1, 2, \dots\}$ with transition probabilities

   $$p_{i,i+1} = a_i, \quad p_{i,0} = 1 - a_i,$$

   where $(a_i : i \geq 0)$ is a sequence of constants satisfying $0 < a_i < 1$ for all i. Let $b_0 = 1$ and $b_i = a_0 a_1 \cdots a_{i-1}$ for $i \geq 1$. Show that the chain is

   (a) recurrent if and only if $b_i \to 0$ as $i \to \infty$,
   (b) positive recurrent if and only if $\sum_i b_i < \infty$,

   and write down the invariant distribution when the last condition holds.

5. At each time n, a random number S_n of students enter the lecture room, where $S_0, S_1, S_2, \dots$ are independent and Poisson distributed with parameter λ. Each student remains in the room for a geometrically distributed time with parameter p, different times being independent. Let X_n be the number of students present at time n. Show that $\mathbf{X}$ is a Markov chain, and find its invariant distribution.

6. Each morning, a student takes one of three books (labelled 1, 2, and 3) from her shelf. She chooses book i with probability α_i, and choices on successive days are independent. In the evening, she replaces the book at the left-hand end of the shelf. If p_n denotes the probability that on day n she finds the books in the order 1, 2, 3 from left to right, show that p_n converges as $n \to \infty$, and find the limit.

7. Let $\mathbf{X}$ be an irreducible, positive recurrent, aperiodic Markov chain with state space S. Show that $\mathbf{X}$ is reversible in equilibrium if and only if

$$p_{i_1,i_2} p_{i_2,i_3} \cdots p_{i_{n-1},i_n} p_{i_n,i_1} = p_{i_1,i_n} p_{i_n,i_{n-1}} \cdots p_{i_2,i_1},$$

for all finite sequences $i_1, i_2, \ldots, i_n \in S$.

8. A special die is thrown repeatedly. Its special property is that, on each throw, the outcome is equally likely to be any of the five numbers that are different from the immediately previous number. If the first score is 1, find the probability that the $(n+1)$th score is 1.

9. A particle performs a random walk about the eight vertices of a cube. Find
   (a) the mean number of steps before it returns to its starting vertex S,
   (b) the mean number of visits to the opposite vertex T to S before its first return to S,
   (c) the mean number of steps before its first visit to T.

10. *Markov chain Monte Carlo.* We wish to simulate a discrete random variable Z with mass function satisfying $\mathbb{P}(Z = i) \propto \pi_i$, for $i \in S$ and S countable. Let $\mathbf{X}$ be an irreducible Markov chain with state space S and transition matrix $P = (p_{i,j})$. Let $Q = (q_{i,j})$ be given by

$$q_{i,j} = \begin{cases} \min\{p_{i,j}, (\pi_j/\pi_i) p_{j,i}\} & \text{if } i \neq j, \\ 1 - \displaystyle\sum_{j:\, j \neq i} q_{i,j} & \text{if } i = j. \end{cases}$$

Show that Q is the transition matrix of a Markov chain which is reversible in equilibrium, and has invariant distribution equal to the mass function of Z.

11. Let i be a state of an irreducible, positive recurrent Markov chain $\mathbf{X}$, and let V_n be the number of visits to i between times 1 and n. Let $\mu = \mathbb{E}_i(T_i)$ and $\sigma^2 = \mathbb{E}_i([T_i - \mu]^2)$ be the mean and variance of the first return time to the starting state i, and assume $0 < \sigma^2 < \infty$.

Suppose $X_0 = i$. Show that

$$U_n = \frac{V_n - (n/\mu)}{\sqrt{n\sigma^2/\mu^3}}$$

converges in distribution to the normal distribution N(0, 1) as $n \to \infty$.

12. Consider a pack of cards labelled $1, 2, \ldots, 52$. We repeatedly take the top card and insert it uniformly at random in one of the 52 possible places, that is, on the top or on the bottom or in one of the 50 places inside the pack. How long on average will it take for the bottom card to reach the top?

Let p_n denote the probability that after n iterations, the cards are found to be in increasing order from the top. Show that, irrespective of the initial ordering, p_n converges as $n \to \infty$, and determine the limit p. You should give precise statements of any general results to which you appeal.

Show that, at least until the bottom card reaches the top, the ordering of the cards inserted beneath it is uniformly random. Hence or otherwise show that, for all n,

$$|p_n - p| \leq \frac{52(1 + \log 51)}{n}.$$

(Cambridge 2003)

13. Consider a collection of N books arranged in a line along a bookshelf. At successive units of time, a book is selected randomly from the collection. After the book has been consulted, it is replaced on the shelf one position to the left of its original position, with the book in that position moved to the right by one. That is, the selected book and its neighbour to the left swap positions. If the selected book is already in the leftmost position, it is returned there. All but one of the books have plain covers and are equally likely to be selected. The other book has a red cover. At each time unit, the red book will be selected with probability p, where $0 < p < 1$. Each other book will be selected with probability $(1 - p)/(N - 1)$. Successive choices of book are independent.

    Number the positions on the shelf from 1 (at the left) to N (at the right). Write X_n for the position of the red book after n units of time. Show that $\mathbf{X}$ is a Markov chain, with non-zero transition probabilities given by:

$$\begin{aligned} p_{i,i-1} &= p && \text{for } i = 2, 3, \ldots, N, \\ p_{i,i+1} &= \frac{1-p}{N-1} && \text{for } i = 1, 2, \ldots, N-1, \\ p_{i,i} &= 1 - p - \frac{1-p}{N-1} && \text{for } i = 2, 3, \ldots, N-1, \\ p_{1,1} &= 1 - \frac{1-p}{N-1}, \\ p_{N,N} &= 1 - p. \end{aligned}$$

    If $(\pi_i : i = 1, 2, \ldots, N)$ is the invariant distribution of the Markov chain $\mathbf{X}$, show that

$$\pi_2 = \frac{1-p}{p(N-1)}\pi_1, \qquad \pi_3 = \frac{1-p}{p(N-1)}\pi_2.$$

    Deduce the invariant distribution. (Oxford 2005)

* 14. Consider a Markov chain with state space $S = \{0, 1, 2, \ldots\}$ and transition matrix given by

$$p_{i,j} = \begin{cases} qp^{j-i+1} & \text{for } i \geq 1 \text{ and } j \geq i-1, \\ qp^j & \text{for } i = 0 \text{ and } j \geq 0, \end{cases}$$

    and $p_{i,j} = 0$ otherwise, where $0 < p = 1 - q < 1$.

    For each $p \in (0, 1)$, determine whether the chain is transient, null recurrent, or positive recurrent, and in the last case find the invariant distribution. (Cambridge 2007)

15. Let $(X_n : n \geq 0)$ be a simple random walk on the integers: the random variables $\xi_n := X_n - X_{n-1}$ are independent, with distribution

$$\mathbb{P}(\xi = 1) = p, \quad \mathbb{P}(\xi = -1) = q,$$

    where $0 < p < 1$ and $q = 1 - p$. Consider the hitting time $\tau = \inf\{n : X_n = 0 \text{ or } X_n = N\}$, where $N > 1$ is a given integer. For fixed $s \in (0, 1)$, define

$$H_k = \mathbb{E}\big(s^\tau 1(X_\tau = 0) \,\big|\, X_0 = k\big) \qquad \text{for } k = 0, 1, \ldots, N.$$

    Show that the H_k satisfy a second-order difference equation, and hence find H_k. (Cambridge 2009)

16. An erratic bishop starts at the bottom left of a chess board and performs random moves. At each stage, she picks one of the available legal moves with equal probability, independently of

earlier moves. Let X_n be her position after n moves. Show that $(X_n : n \geq 0)$ is a reversible Markov chain, and find its invariant distribution.

What is the mean number of moves before she returns to her starting square?

17. A frog inhabits a pond with an infinite number of lily pads, numbered $1, 2, 3 \ldots$. She hops from pad to pad in the following manner: if she happen to be on pad i at a given time, she hops to one of the pads $(1, 2, \ldots, i, i+1)$ with equal probability.

    (a) Find the equilibrium distribution of the corresponding Markov chain.
    (b) Suppose the frog starts on pad k and stops when she returns to it. Show that the expected number of times the frog hops is $e(k-1)!$, where $e = 2.718 \ldots$. What is the expected number of times she will visit the lily pad $k+1$?

    (Cambridge 2010)

18. Let $(X_n : n \geq 0)$ be a simple, symmetric random walk on the integers $\{\ldots, -1, 0, 1, \ldots\}$, with $X_0 = 0$ and

$$\mathbb{P}\big(X_{n+1} = i \pm 1 \,\big|\, X_n = i\big) = \tfrac{1}{2}.$$

    For each integer $a \geq 1$, let $T_a = \inf\{n \geq 0 : X_n = a\}$. Show that T_a is a stopping time.

    Define a random variable Y_n by the rule

$$Y_n = \begin{cases} X_n & \text{if } n < T_a, \\ 2a - X_n & \text{if } n \geq T_a. \end{cases}$$

    Show that $(Y_n : n \geq 0)$ is also a simple, symmetric random walk.

    Let $M_n = \max\{X_i : 0 \leq i \leq n\}$. Explain why $\{M_n \geq a\} = \{T_a \leq n\}$ for $a \geq 1$. By using the process $(Y_n : n \geq 0)$ constructed above, show that, for $a \geq 1$,

$$\mathbb{P}\big(M_n \geq a,\ X_n \leq a - 1\big) = \mathbb{P}(X_n \geq a + 1),$$

    and thus,

$$\mathbb{P}(M_n \geq a) = \mathbb{P}(X_n \geq a) + \mathbb{P}(X_n \geq a + 1).$$

    Hence compute $\mathbb{P}(M_n = a)$, where a and n are positive integers with $n \geq a$. [Hint: if n is even, then X_n must be even, and if n is odd, then X_n must be odd.] (Cambridge 2010)

Appendix A
Elements of combinatorics

The number of *permutations* of n distinct objects is

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

and is pronounced 'n factorial'. The number of ordered subsequences of length r from these n objects is $n(n-1)\cdots(n-r+1)$, which may be written as

$${}_nP_r := \frac{n!}{(n-r)!}.$$

If the ordering of these r objects is not important, then any of the $r!$ possible orderings gives rise to the same subset, and the number of such *combinations* is then

$${}_nC_r := \frac{1}{r!}\,{}_nP_r = \frac{n!}{r!\,(n-r)!}.$$

This is usually called the *binomial coefficient*, written as

$$\binom{n}{r} := \frac{n!}{r!\,(n-r)!},$$

and pronounced 'n choose r'. It is useful to note that

$$\binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!}, \tag{A.1}$$

since this formal definition makes sense even when n is a general real number.

There are entire volumes devoted to combinatorial identities. Of the many such identities we highlight one, namely the following:

$$\binom{2n}{n} = \sum_{r=0}^{n}\binom{n}{r}\binom{n}{n-r}, \tag{A.2}$$

the proof of which is left as a small exercise.

The binomial theorem states that

$$(1+x)^n = \sum_{r=0}^{n}\binom{n}{r}x^r,$$

valid for $x \in \mathbb{R}$ and $n = 1, 2, \ldots$. There is a more general version that holds even when n is fractional, or even negative.

Theorem A.3 (Extended binomial theorem) *Let $y \in \mathbb{R}$. We have that*

$$(1+x)^y = \sum_{r=0}^{\infty} \binom{y}{r} x^r \qquad \text{for } |x| < 1,$$

where the binomial coefficients are given by (A.1)*.*

It is often necessary to compare the rate of growth of $n!$ with polynomial and exponential functions of n. The requisite formula is called Stirling's formula.

Theorem A.4 (Stirling's formula) *We have that*

$$n! \sim (n/e)^n \sqrt{2\pi n} \qquad \text{as } n \to \infty,$$

where $a_n \sim b_n$ means $a_n/b_n \to 1$.

A 'short' proof of Stirling's formula has been given by Romik (2000), and a proof of a slightly weaker result without the identification of the constant $\sqrt{2\pi}$ is provided in Norris (1997, Sect. 1.12).

Partial proof We prove only the weaker 'logarithmic asymptotic'

$$\log n! \sim n \log n \qquad \text{as } n \to \infty. \tag{A.5}$$

Since log is increasing on the interval $[1, \infty)$, we have that

$$\int_1^n \log x \, dx \le \sum_{k=1}^{n} \log k \le \int_1^{n+1} \log x \, dx,$$

which is to say that

$$n \log n - n + 1 \le \log n! \le (n+1)\log(n+1) - n.$$

Equation (A.5) follows by dividing by $n \log n$ and letting $n \to \infty$. □

Readers are referred to Graham *et al.* (1994) for further information about these and related topics.

Appendix B

Difference equations

We say that the sequence $x_0, x_1, \ldots$ satisfies a *difference equation* if

$$a_0 x_{n+k} + a_1 x_{n+k-1} + \cdots + a_k x_n = 0 \qquad \text{for } n = 0, 1, 2, \ldots, \tag{B.1}$$

where $a_0, a_1, \ldots, a_k$ is a given sequence of real numbers and $a_0 \neq 0$. We generally suppose that $a_k \neq 0$, and in this case we call (B.1) a difference equation of *order* k. Difference equations occur quite often in the study of random processes, particularly random walks, and it is useful to be able to solve them. We describe here how to do this.

Just as in solving differential equations, we require boundary conditions in order to solve difference equations. To see this, note that (B.1) may be rewritten as

$$x_{n+k} = -\frac{1}{a_0}(a_1 x_{n+k-1} + a_2 x_{n+k-2} + \cdots + a_k x_n) \qquad \text{for } n = 0, 1, 2, \ldots \tag{B.2}$$

since $a_0 \neq 0$. Thus, if we know the values of $x_0, x_1, \ldots, x_{k-1}$, equation (B.2) with $n = 0$ provides the value of x_k. Next, equation (B.2) with $n = 1$ tells us the value of x_{k+1}, and so on. That is to say, there is a unique solution of (B.1) with specified values for $x_0, x_1, \ldots, x_{k-1}$. It follows that, if $a_k \neq 0$, the general solution of (B.1) contains exactly k independent arbitrary constants, and so exactly k independent boundary conditions are required in order to solve (B.1) explicitly.

The principal step involved in solving (B.1) is to find the roots of the *auxiliary equation*

$$a_0 \theta^k + a_1 \theta^{k-1} + \cdots + a_{k-1}\theta + a_k = 0, \tag{B.3}$$

a polynomial in θ of degree k. We denote the (possibly complex) roots of this polynomial by $\theta_1, \theta_2, \ldots, \theta_k$. The general solution of (B.1) is given in the next theorem.

Theorem B.4 *Let $a_0, a_1, \ldots, a_k$ be a sequence of real numbers with $a_0 \neq 0$.*

(a) *If the roots $\theta_1, \theta_2, \ldots, \theta_k$ of the auxiliary equation are distinct, the general solution of* (B.1) *is*

$$x_n = c_1\theta_1^n + c_2\theta_2^n + \cdots + c_k\theta_k^n \qquad \text{for } n = 0, 1, 2, \ldots, \tag{B.5}$$

where $c_1, c_2, \ldots, c_k$ are arbitrary constants.

(b) *More generally, if $\theta_1, \theta_2, \ldots, \theta_r$ are the distinct roots of the auxiliary equation and m_i is the multiplicity of θ_i for $i = 1, 2, \ldots, r$, the general solution of* (B.1) *is*

$$\begin{aligned} x_n = {} & (a_1 + a_2 n + \cdots + a_{m_1} n^{m_1 - 1})\theta_1^n \\ & + (b_1 + b_2 n + \cdots + b_{m_2} n^{m_2 - 1})\theta_2^n + \cdots \\ & + (c_1 + c_2 n + \cdots + c_{m_r} n^{m_r - 1})\theta_r^n \qquad \text{for } n = 0, 1, 2, \ldots, \end{aligned} \tag{B.6}$$

where the k numbers $a_1, \ldots, a_{m_1}, b_1 \ldots, b_{m_2}, \ldots, c_1, \ldots, c_{m_r}$ are arbitrary constants.

The auxiliary equation may not possess k *real* roots, and thus some or all of $\theta_1, \theta_2, \ldots, \theta_k$ may have non-zero imaginary parts. Similarly, the arbitrary constants in Theorem B.4 need not necessarily be real, and the general solution (B.6) is actually the general solution for *complex* solutions of the difference equation (B.1). If we seek *real* solutions only of (B.1), then this fact should be taken into account when finding the values of the constants.

We do not prove this theorem, but here are two ways of going about proving it, should one wish to do so. The first way is constructive, and uses the generating functions of the sequences of a_i and x_i (see Hall (1967, p. 20)). The second way is to check that (B.5) and (B.6) are indeed solutions of (B.1) and then to note that they contain the correct number of arbitrary constants.

Here is an example of the theorem in action.

Example B.7 Find the solution of the difference equation

$$x_{n+3} - 5x_{n+2} + 8x_{n+1} - 4x_n = 0$$

subject to the boundary conditions $x_0 = 0$, $x_1 = 3$, $x_3 = 41$.

Solution The auxiliary equation is

$$\theta^3 - 5\theta^2 + 8\theta - 4 = 0$$

with roots $\theta = 1, 2, 2$. The general solution is therefore

$$x_n = a1^n + (b + cn)2^n,$$

where the constants a, b, c are found from the boundary conditions to be given by $a = 1$, $b = -1$, $c = 2$. △

An important generalization of Theorem B.4 deals with difference equations of the form

$$a_0 x_{n+k} + a_1 x_{n+k-1} + \cdots + a_k x_n = g(n) \qquad \text{for } n = 0, 1, 2, \ldots, \tag{B.8}$$

where g is a given function of n, not always equal to 0. There are two principal steps in solving (B.8). First, we find a solution of (B.8) by any means available, and we call this a *particular solution*. Secondly, we find the general solution to the difference equation obtained by setting $g(n) = 0$ for all n:

$$a_0 x_{n+k} + a_1 x_{n+k-1} + \cdots + a_k x_n = 0 \qquad \text{for } n = 0, 1, 2, \ldots,$$

this solution is called the *complementary solution*.

Theorem B.9 *Suppose that $a_0, a_1, \ldots, a_k$ is a given sequence of real numbers and that $a_0 \neq 0$. The general solution of* (B.8) *is*

$$x_n = \kappa_n + \pi_n \qquad \textit{for } n = 0, 1, 2 \ldots, \tag{B.10}$$

where $\kappa_0, \kappa_1, \ldots$ is the complementary solution and $\pi_0, \pi_1, \ldots$ is a particular solution.

This may be proved in the same general way as Theorem B.4. We finish with an example.

Example B.11 Find the solution of the difference equation

$$x_{n+2} - 5x_{n+1} + 6x_n = 4n + 2 \tag{B.12}$$

subject to the boundary conditions $x_0 = 5$, $x_4 = -37$.

Solution The right-hand side of (B.12) is a polynomial function of n, and this suggests that there may be a particular solution which is a polynomial. Trial and error shows that

$$x_n = 2n + 4 \qquad \text{for } n = 0, 1, 2, \ldots$$

is a *particular solution*. The general solution of the difference equation

$$x_{n+2} - 5x_{n+1} + 6x_n = 0$$

is

$$x_n = a2^n + b3^n \qquad \text{for } n = 0, 1, 2, \ldots,$$

where a and b are arbitrary constants. It follows that the general solution of (B.12) is

$$x_n = a2^n + b3^n + 2n + 4 \qquad \text{for } n = 0, 1, 2 \ldots.$$

The constants a and b are found from the boundary conditions to be given by $a = 2, b = -1$.

△

Answers to exercises

Chapter 1

1.17. Yes.
1.21. $\frac{6}{10}$.
1.30. Compare $1-(\frac{5}{6})^4$ with $1-(\frac{35}{36})^{24}$.
1.36. $\frac{1}{8}$.
1.43. Either A or B must have zero probability.
1.46. (a) $(1-p)^m$, (b) $\frac{1}{2}[1+(q-p)^n]$, where $p+q=1$.
1.52. (a) $\frac{46}{63}$, (b) $\frac{16}{37}$.
1.53. $\frac{4}{3}(\frac{2}{3})^n - \frac{1}{3}(-\frac{1}{3})^n$.

Chapter 2

2.11. Only V is a random variable.
2.12. $c=1$.
2.26. $\mathbb{P}(Y=0)=e^{-\lambda}\cosh\lambda$, $\mathbb{P}(Y=1)=e^{-\lambda}\sinh\lambda$.
2.37. npq.

Chapter 3

3.8.

	$x=0$	$x=1$	$x=2$
$y=0$	$\frac{11\cdot43}{13\cdot51}$	$\frac{88}{13\cdot51}$	$\frac{1}{13\cdot17}$
$y=1$	$\frac{88}{13\cdot51}$	$\frac{8}{13\cdot51}$	0
$y=2$	$\frac{1}{13\cdot17}$	0	0

3.9. $p_X(i)=\theta^i(\theta+\theta^2+\theta^3)$ for $i=0,1,2$.
3.29. Take $Y=X$, so that $X+Y$ takes even values only—it cannot then have the Poisson distribution.

Chapter 4

4.4. (a) $V(s)=2U(s)$, (b) $V(s)=U(s)+(1-s)^{-1}$, (c) $V(s)=sU'(s)$.
4.5. $u_{2n}=\binom{2n}{n}p^nq^n$, $u_{2n+1}=0$.
4.32. $\mathbb{E}(X)=-\left.\frac{d\Delta}{ds}\right|_{s=1}$. $\mathbb{E}(\log X)=\lim_{y\to0}\frac{\Delta(-y)-1}{y}$.

Chapter 5

5.12. Yes.

5.13. $F_Y(y) = \begin{cases} F_X(y) & \text{if } y \geq 0, \\ 0 & \text{if } y < 0. \end{cases}$

5.19. $c = \frac{1}{2}$.

5.30. $F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x^2 & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases}$

5.31. $F(x) = \begin{cases} \frac{1}{2}e^x & \text{if } x \leq 0, \\ 1 - \frac{1}{2}e^{-x} & \text{if } x > 0. \end{cases}$

5.32. $f(x) = F'(x)$ if $x \neq 0$.

5.33. $F(x) = \exp(-e^{-x})$.

5.45. $w = 1$, arbitrary positive λ.

5.48. $2/\pi$.

5.54. (a) $f_A(x) = \frac{1}{2}\lambda \exp\bigl(-\frac{1}{2}\lambda[x-5]\bigr)$ if $x > 5$.
(b) $f_B(x) = \lambda e^{-\lambda - 1}$ if $x > 1$.
(c) $f_C(x) = \lambda x^{-2} \exp\bigl(-\lambda[x^{-1} - 1]\bigr)$ if $x < 1$.
(d) $f_D(x) = \frac{1}{2}\lambda x^{-\frac{3}{2}} \exp\bigl(-\lambda[x^{-\frac{1}{2}} - 1]\bigr)$ if $x < 1$.

5.68. $c = 6$, $\mathbb{E}(X) = \frac{1}{2}$, $\text{var}(X) = \frac{1}{20}$.

5.69. e^2.

5.70. $\frac{2}{3}$.

Chapter 6

6.25. $c = \frac{6}{7}$ and $F(x, y) = \frac{6}{7}(\frac{1}{3}x^3 y + \frac{1}{8}x^2 y^2)$ if $0 \leq x \leq 1$, $0 \leq y \leq 2$.

6.26. $\mathbb{P}(X + Y \leq 1) = 1 - 2e^{-1}$, $\mathbb{P}(X > Y) = \frac{1}{2}$.

6.35. $c = 3$, $f_X(x) = 3x^2$ if $0 < x < 1$, and $f_Y(y) = \frac{3}{2}(1 - y^2)$ if $0 < y < 1$. X and Y are dependent.

6.36. X, Y, and Z are independent, and $\mathbb{P}(X > Y) = \mathbb{P}(Y > Z) = \frac{1}{2}$.

6.45. $f_{X+Y}(u) = \frac{1}{2}u^2 e^{-u}$ if $u > 0$.

6.47. $X + Y$ has the normal distribution with mean 0 and variance 2.

6.54. $f_{U,V}(u, v) = \dfrac{1}{4\pi\sigma^2} \exp\left(-\dfrac{1}{4\sigma^2}[u^2 + (v - 2\mu)^2]\right)$, which factorizes, so that U and V are independent.

6.60. $f_{X|Y}(x \mid y) = y^{-1}$ and $f_{Y|X}(y \mid x) = e^{x-y}$, if $0 < x < y < \infty$.

6.70. $\mathbb{E}\sqrt{X^2 + Y^2} = \frac{2}{3}$ and $\mathbb{E}(X^2 + Y^2) = \frac{1}{2}$.

6.71. Let X and Z be independent, X having the normal distribution with mean 0, and variance 1, and Z taking the values ± 1 each with probability $\frac{1}{2}$. Define $Y = XZ$.

6.72. $\mathbb{E}(X \mid Y = y) = \frac{1}{2}y$ and $\mathbb{E}(Y \mid X = x) = x + 1$.

Chapter 7

7.59. (a) $[\lambda/(\lambda - t)]^w$ if $t < \lambda$, (b) $\exp(-\lambda + \lambda e^t)$.

7.60. $\mu^3 + 3\mu\sigma^2$.
7.72. The discrete distribution that places probability $\frac{1}{2}$ on the each value 0 and 2μ.
7.97. (a) $[\lambda/(\lambda - it)]^w$, (b) $\exp(-\lambda + \lambda e^{it})$.

Chapter 8

8.32. $a = -\sqrt{6}$, $b = 2\sqrt{6}$.
8.42. $\Lambda^*(a) = \frac{1}{2}a^2$.
8.44. $\dfrac{1}{2} - \dfrac{1}{\pi}\tan^{-1} a$.

Chapter 9

9.11. $1 - p^n$.

Chapter 10

10.4. $\mathbb{E}(S_n) = n(p - q)$, $\text{var}(S_n) = 4pqn$.
10.5. p^n.
10.9. $\dbinom{2n+1}{n} p^{n+1} q^n$.
10.43. If $p = \frac{1}{2}$, it is $\frac{1}{2}$. If $p \neq \frac{1}{2}$, it is $\dfrac{1}{\theta^N - 1}\left(\dfrac{\theta^{N+1} - 1}{(N+1)(\theta - 1)} - 1\right)$, where $\theta = q/p$.
10.44. $q\left(\dfrac{\theta^{k-1} - 1}{\theta^k - 1}\right)$, where $\theta = q/p$.

Chapter 11

11.3. $p(t) = e^{-\lambda t}$ if $t \geq 0$.
11.32. The gamma distribution with parameters 2 and λ.
11.59. $m(t) = m(0)e^{(\lambda - \mu)t}$.
11.79. $e^{-\mu_1 t}/(e^{-\mu_1 t} + e^{-\mu_2 t} + e^{-\mu_3 t})$.

Chapter 12

12.10. $S = \{1, 2, 3, 4, 5, 6\}$ and

$$P = \begin{pmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & \frac{2}{3} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & \frac{5}{6} & \frac{1}{6} \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

12.11. Not necessarily. For example, let $\mathbf{X} = (X_n)$ be a symmetric random walk with $X_0 = 0$. Let R be independent of $\mathbf{X}$ and equally likely to take the values ± 1, and set $Y_n = RX_n$ and $Z_n = X_n + Y_n$. Now consider $\mathbb{P}(Z_{n+1} = 0 \mid Z_n = 0,\ Z_1 = z)$ for $z = 0, 2$.

12.22. $\frac{1}{3} + \frac{2}{3}(-\frac{1}{2})^n$.
12.28. $\{1, 5\}$, $\{3\}$, $\{2, 4\}$. The first two are closed.
12.29. The process $\mathbf{X}$ on the integers $\mathbb{Z}$ given by $X_n = n$.
12.62. $1/\zeta(\alpha)$, where $\zeta(s) = \sum_{k=1}^{\infty} k^{-s}$ is the Riemann zeta function.
12.80. All states are positive recurrent. They have period 2 in the first case. In the second case, thay have period 2 if $p = \frac{1}{2}$ and are aperiodic otherwise.
12.96. $\frac{1}{14}$ for the corners and $\frac{1}{7}$ for the other vertices. The mean number is 2.
12.119. Yes. Take $S = \{1, 2\}$, $\mathbb{P}(X_0 = 1) = \frac{1}{2}$, $p_{1,2} = p_{2,1} = 1$.
12.120. $\pi_k = \theta^k(1 - \theta)/(1 - \theta^{N+1})$, where $\theta = p/q$.
12.125. $\frac{1}{140}$ for the corners, $\frac{1}{84}$ for other side squares, $\frac{2}{105}$ for other squares.

Remarks on problems

Chapter 1

1. Expand $(1+x)^n + (1-x)^n$.
2. No.
6. $\frac{79}{140}$ and $\frac{40}{61}$.
7. $\frac{11}{50}$.
8. $\sqrt{3/(4\pi n)}(\frac{27}{32})^n$.
9. If X and Y are the numbers of heads obtained,

$$\begin{aligned}\mathbb{P}(X=Y) &= \sum_k \mathbb{P}(X=k)\mathbb{P}(Y=k) = \sum_k \mathbb{P}(X=k)\mathbb{P}(Y=n-k)\\ &= \mathbb{P}(X+Y=n).\end{aligned}$$

10. $1-(1-p)(1-p^2)^2$ and $1-(1-p)(1-p^2)^2 - p + p[1-(1-p)^2]^2$.
12. To do this rigorously is quite complicated. You need to show that the proportion $\frac{1}{10}$ is correct for any single one of the numbers $0, 1, 2, \ldots, 9$.
13. Use the Partition Theorem 1.48 to obtain the difference equations. Either iterate these directly to solve them, or set up a matrix recurrence relation, and iterate this.
14. (a) Induction. (b) Let A_i be the event that the ith key is hung on its own hook.
15. Use the result of Problem 1.11.14(a).
16. Conditional probabilities again. The answer is $\frac{1}{4}(2e^{-1} + e^{-2} + e^{-4})$.
18. $\bigcup_{i=1}^n A_i \to \bigcup_{i=1}^\infty A_i$ as $n \to \infty$.
19. Show $n = 6$.

Chapter 2

2. Use Theorem 2.42 with X and B_i chosen appropriately. The answer is $m(r) = r/p$.
3. $\mathbb{E}(X^2) = \sum x^2 \mathbb{P}(X = x)$, the sum of non-negative terms.
4. $\alpha < -1$ and $c = 1/\zeta(-\alpha)$, where $\zeta(p) = \sum_k k^{-p}$ is the Riemann zeta function.
5. For the last part, show that $G(n) = \mathbb{P}(X > n)$ satisfies $G(m+n) = G(m)G(n)$, and solve this relation.
6. The summation here is $\sum_{k=0}^\infty \sum_{i=k+1}^\infty \mathbb{P}(X = i)$. Change the order of summation. For the second part, use the result of Exercise 1.20.
7. This generalizes the result of Problem 2.6.6.
8. This is sometimes called Banach's matchbox problem. First, condition on which pocket is first emptied. You may find the hint more comprehensible if you note that $2(n-h)p_h = (2n-h)p_{h+1}$. The mean equals $(2n+1)p_0 - 1$.
9. $(1-p^n)/[p^n(1-p)]$.

Chapter 3

1. Use the result of Exercise 1.35, with Theorem 3.27.
2. $a = b = \frac{1}{2}$. No.
4. $\mathbb{P}(U_n = k) = \mathbb{P}(U_n \geq k) - \mathbb{P}(U_n \geq k+1)$, and $\mathbb{P}(U_n \geq k) = \left(1 - \dfrac{k-1}{N}\right)^n$.
5. $\mathbb{P}(U > k) = \mathbb{P}(X > k)\mathbb{P}(Y > k)$.
6. Let 1_k be the indicator function of the event that, when there are $2k$ ends, a new hoop is created at the next step. Then $\mathbb{E}(1_k) = k\big/\binom{2k}{2} = 1/(2k-1)$. The mean final number of hoops is $\sum_{k=1}^{n} \mathbb{E}(1_k)$.
7. Use Theorem 2.42 with $B_i = \{N = i - 1\}$.
9. (a) $\frac{1}{2}$, (b) $\frac{1}{6}(3\sqrt{5} - 1)$, (c) $\frac{5}{6}$.
10. Let Z_i be the indicator function that the ith box is empty. The total number of empty boxes is $S = Z_1 + Z_2 + \cdots + Z_M$. Also, $\mathbb{E}(Z_i) = (M-1)^N/M^N$ and $\mathbb{E}(S) = M\mathbb{E}(Z_1)$.
11. Adapt the hint for Problem 3.6.10.
12. In calculating the mean, remember that the expectation operator $\mathbb{E}$ is linear. The answer here is $c\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{c}\right)$, a much more elegant solution than that proposed for Problem 2.6.7.
13. $c[1 - (1 - c^{-1})^n]$, by using indicator functions.
14. Condition on the value of N. X has the Poisson distribution with parameter λp.
15. $\operatorname{var}(U_n) = (n-1)pq - (3n-5)(pq)^2$.
16. (a) The means are 7 and 0, and both variances equal $\frac{35}{6}$. To locate the extremal probabilities, find the number of ways in which the various possible outcomes can occur. For example, $\mathbb{P}(X = x)$ is maximized at $x = 7$. To verify (in)dependence, it is convenient to use a simple fact of the type $\mathbb{P}(X = 3,\ Y = 0) = 0$.

Chapter 4

1. Note that $\mathbb{P}(X = k) = u_{k-1} - u_k$.
2. $(\frac{1}{6})^7 \left(\dfrac{13!}{6!\,7!} - 49\right)$.
3. $\frac{9}{19}$, $\frac{6}{19}$, $\frac{4}{19}$. The mean number of throws is 3.
4. $[q/(1-ps)]^N$. The variance is $Np(1-p)^{-2}$.
5. The first part of this problem may be done by way of Theorem 4.36, with $N+1$ having a geometric distribution and the X_i having the Bernoulli distribution. Alternatively, use the methods of Chapter 3. The answer to the first part is $2(1-p)p^r(2-p)^{-r-1}$, and to the second part $\binom{n}{r}(\frac{1}{2})^{n+1}p^{n-r}(2-p)^{r+1}$.
6. For the third part, find the real part of $G_X(\theta)$, where θ is a primitive complex root of unity.
8. $G_X(s) = G_N(\frac{1}{2} + \frac{1}{2}s)$, giving by independence that $G = G_N$ satisfies the functional equation $G(s) = G(\frac{1}{2} + \frac{1}{2}s)^2$. Iterate this to obtain $G(s) = G\big(1 + (s-1)/m\big)^m$, where $m = 2^n$, use Taylor's theorem and take the limit as $n \to \infty$.
9. This is an alternative derivation of the result of Problem 3.6.12.
10. $\mathbb{P}(\text{A wins}) = a/(a + b - ab)$. The mean number of shots is $(2-a)/(a+b-ab)$.
11. This is essentially a reprise of Problem 4.5.8.

Chapter 5

3. $n \geq 4$.
4. $f_Y(y) = \sqrt{2/\pi}\exp(-\frac{1}{2}y^2)$ for $y > 0$. $\sqrt{2/\pi}$ and $1 - (2/\pi)$.
5. Let $F^{-1}(y) = \sup\{x : F(x) = y\}$. Find $\mathbb{P}(F(X) \leq y)$.
6. Note that $x \leq F(y)$ if and only if $F^{-1}(x) \leq y$, whenever $0 < F(y) < 1$.
7. Integrate by parts. You are proving that $\mathbb{E}(X) = \int \mathbb{P}(X > x)\,dx$, the continuous version of Problem 2.6.6.
8. Apply the conclusion of Problem 5.8.7 to $Y = g(X)$, express the result as a double integral and change the order of integration.
10. $f_Y(y) = \dfrac{3}{(1-y)^2}\exp\left(-\dfrac{y+2}{1-y}\right)$ for $-2 < y < 1$.
11. This distance has distribution function $(2/\pi)\tan^{-1}x$ for $0 \leq x < \infty$.
12. Assume that the centre is uniform on the rectangle $[0, a] \times [0, b]$, and that the acute angle θ between the needle and a line of the first grid is uniform on $[0, \frac{1}{2}\pi]$. There is no intersection if and only if the centre lies within an inner rectangle of size $(a - \ell\cos\theta) \times (b - \ell\sin\theta)$. Hence, the probability of an intersection is

$$\frac{2}{\pi ab}\int_0^{\pi/2}\big[ab - (a - \ell\cos\theta)(b - \ell\sin\theta)\big]\,d\theta = \frac{2\ell}{\pi ab}(a + b - \tfrac{1}{2}\ell).$$

13. Let X_k be the position of the kth break (in no special order). The pieces form a polygon if no piece is longer than the sum of the other lengths, which is equivalent to each piece having length less than $\frac{1}{2}$. This fails to occur if and only if the disjoint union $A_0 \cup A_1 \cup \cdots \cup A_n$ occurs, where A_0 is the event there is no break in $(0, \frac{1}{2}]$, and A_k is the event of no break in $(X_k, X_k + \frac{1}{2}]$ for $k \geq 1$ (remember the permanent break at 1). Now, $\mathbb{P}(A_0) = (\frac{1}{2})^n$, and for $k \geq 1$,

$$\mathbb{P}(A_k) = \int_0^1 \mathbb{P}(A_k \mid X_k = x)\,dx = \int_0^{\frac{1}{2}} (\tfrac{1}{2})^{n-1}\,dx = (\tfrac{1}{2})^n.$$

14. Find $\mathbb{P}(Y \leq y)$ for $y \in \mathbb{R}$.

Chapter 6

1. For the first part, find the joint density function of X and XY by the method of change of variables, and then find the marginal density function of XY.
2. No.
3. The region $\{(x, y, z) : \sqrt{4xz} < y \leq 1,\ 0 \leq x, z \leq 1\}$ has volume $\frac{5}{36} + \frac{1}{6}\log 2$.
4. $\min\{X, Y\} > u$ if and only if $X > u$ and $Y > u$.
5. Show that $G(y) = \mathbb{P}(Y > y)$ satisfies $G(x + y) = G(x)G(y)$, and solve this equation. The corresponding question for integer-valued random variables appears at Problem 2.6.5.
6. If you can do Problem 6.9.4 then you should be able to do this one.

$$\mathbb{P}(U \leq x,\ V \leq y) = F(y)^n - [F(y) - F(x)]^n \quad \text{for} \quad x < y.$$

9. $f_Y(y) = \frac{1}{4}(3y+1)e^{-y}$ for $0 < y < \infty$.
11. Use Theorem 6.62 with $g(x, y) = \sqrt{x^2+y^2}$ and change to polar coordinates. The variance equals $\sigma^2(2 - \frac{1}{2}\pi)$.
12. Draw the regions in question in the (x, y)-plane. It is useful to prove that $R^2 = X^2 + Y^2$ and $\Theta = \tan^{-1}(Y/X)$ are independent, having an exponential and uniform distributions, respectively.
    (a) $1 - \exp(-\frac{1}{2}a^2/\sigma^2)$.
    (b) $\alpha/(2\pi)$.
13. (a) (i) $1 - e^{-\lambda X}$ is uniformly distributed on (0, 1).
    (ii) $\min\{X, Y\}$ has the exponential distribution with parameter 2λ.
    (iii) $X - Y$ has the bilateral exponential distribution.
    (b) The answer is 0 if $a < 1$ and $a/(1+a)$ if $a \geq 1$.
14. This is largely an exercise in changes of variables, but there is a much better argument which shows $\frac{1}{2}$ to be the answer to (b).
20. $\mathbb{E}(Y \mid X = x) = \frac{1}{2}x$ for $0 < x < 1$.
21. You could use Problem 5.8.7 for the first part. The covariance is $\frac{1}{36}$.
22. $f_\Phi(\phi) = \frac{1}{2}\sin\phi$ for $0 < \phi < \pi$. The intersection of a plane through C with the surface of the planet is called a *great circle*. There is a two–one correspondence between hemispheres and great circles. The key fact for the second part is that the intersection of the surfaces of two hemispheres with radius 1 has area 2α, where α is the angle between the two corresponding great circles at their points of intersection. The answers to the next two parts are $(\pi + \phi)/(2\pi)$ and $(\pi - \phi)/(2\pi)$.
23. Picking a random point on the surface is equivalent to picking a random hemisphere (with pole at that point), and this in turn is equivalent to picking a random great circle, and then flipping a fair coin to choose a hemisphere. The answer is $1-(a_n/2^n)$, where $a_n = n^2 - n + 2$.
24. The final answer is no, and an explanation would be appreciated.
25. Use the Jacobian method to find the joint pdf of X and Z.
27. For part (iii) of (a), use the Jacobian method with $U = X$.
28. (d) One needs the derivative of the quadratic to be negative at -1 and positive at 1.

Chapter 7

2. The identity $\sum(X_i - \overline{X})^2 = \sum\big[X_i - \mu - (\overline{X} - \mu)\big]^2 = \sum(X_i - \mu)^2 - n(\overline{X} - \mu)^2$ may be useful.
3. $\mathbb{E}(S_n/S_n) = n\mathbb{E}(X_1/S_n)$ and $\mathbb{E}(S_m/S_n) = m\mathbb{E}(X_1/S_n)$. The result is generally false when $m > n$. See also Problem 3.6.8.
4. $\big|\{x : F(x) - \lim_{y\uparrow x} F(y) > 1/n\}\big| < n$.
6. Use moment generating functions.
7. For the middle part, find the moment generating function of X_1^2 and use Theorem 7.52.
8. This is basically the same argument as in the random sum formula, Theorem 4.36. $M_S(t) = G_N(M_X(t))$.
9. $\text{var}(Z) = \sum a_m v_{mn} a_n$.
10. Let Z_i be the indicator function that A wins the ith game.

11. $M(s\sigma_1, t\sigma_2)\exp(s\mu_1 + t\mu_2)$.
12. To do the last part, show first that $\psi(t) = M(t)/M(-t)$ satisfies $\psi(t) = \psi(2^{-n}t)^{2^n}$. Show that $\psi(t) = 1 + o(t^2)$ as $t \to 0$, and deduce that $\psi(t) = 1$ by taking the limit above as $n \to \infty$. Hence, $M(t) = M(-t)$, and the original equation become $M(t) = M(\frac{1}{2}t)^4$. Repeat the procedure to obtain $M(t) = \exp(\frac{1}{2}t^2)$.
13. (b) Remember the result of Problem 7.7.8. The answer to the last part is XY, where Y has which distribution?
14. Remember Problem 4.5.9.
15. This is similar to Problem 7.7.13.
16. $e^{itx} = \cos tx + i \sin tx$.
17. Use the inversion theorem, Theorem 7.89, or remember the density function and characteristic function of the Cauchy distribution.
19. $\phi''(0) = 0$ if $\alpha > 2$. What does this imply about such a distribution?
22. $Q = p_1^m + p_2^m + \cdots + p_N^m$.
23. X_n has a gamma distribution.
24. (c) Let A, B be random variables and consider the moment generating function of a random variable Y that equals A with probability $\frac{1}{2}$ and B otherwise.
25. Put $a = b$ to obtain a lower bound for c_p, and use Jensen's inequality.
26. Let $0 < s \leq r$ and $t = r/s \geq 1$. Apply Jensen's inequality to $Y = Z^s$ to obtain $\mathbb{E}(Y^t) \geq \mathbb{E}(Y)^t$.

Chapter 8

1. $\mathbb{P}(Z_n \leq b) = (b/a)^n$ for $0 < b < a$.
3. Use Theorem 8.14.
5. The left-hand side equals $\mathbb{P}(X_1 + X_2 + \cdots + X_n \leq n)$ for appropriately distributed random variables.
6. (a) $2^{-n}T_n = \mathbb{P}(S_n > an)$ with S_n as in (8.41). (b) Use the Poisson distribution.
8. Adapt Example 8.19.
9. If $x + y > \epsilon$, then either $x > \frac{1}{2}\epsilon$ or $y > \frac{1}{2}\epsilon$.
10. Let $Z_i = X_i - Y_i$.
12. $|X| \leq a(1 - I) + MI$, where I is the indicator function of the event that $|X| \geq a$.
13. Combine the results of Problems 8.6.11 and 8.6.12.
14. Use the Cauchy–Schwarz inequality.
15. Moment generating functions may be useful for the first part. Y has a χ^2 distribution.
16. Express $\mathbb{P}(Z_n \leq x)$ in terms of the density function of Y_n and take the limit as $n \to \infty$. It is rather complicated.
17. $B_n(p) = \sum_k f(k/n)\binom{n}{k}p^k q^{n-k}$, where $q = 1 - p$. For the last part, note that $n^{-1}\sum X_i$ converges to p, so that, by continuity, $B_n(p)$ is close to $f(p)$ with large probability. The error probability is controlled using the boundedness of f.
18. The exponential distribution.
19. True, false, and true.
22. This is very similar to Problem 7.7.12.
24. Both mean and variance equal 1.

Chapter 9

1. Z_n has variance $\sigma^2\mu^{n-1}(\mu^n - 1)/(\mu - 1)$ if $\mu \neq 1$ and $n\sigma^2$ if $\mu = 1$.
4. $f(s) = e^{\lambda(s-1)}$.
6. $G_n(s) = 1 - \alpha^{(1-\beta^n)/(1-\beta)}(1-s)^{\beta^n}$.

Chapter 10

1. The first part is a variation of the result of Theorem 10.6. For the second part, follow the first walk for the first n steps and then return along the second walk, reversed.
2. Either solve the difference equation, or relate it to equation (10.26). The number of stages required is $N_1 + N_2 + \cdots + N_M$, where M is the number of moves (remember Theorem 10.28) and each N_i has a geometric distribution. Alternatively, solve the appropriate difference equation.
3. $\dfrac{1}{N}$.
4. There is only one solution to $p\theta^3 - \theta + q = 0$ with absolute value not exceeding 1, when $p \leq \frac{1}{3}$.
5. $e(a) = \dfrac{p}{(q-p)^2}(\rho^N - \rho^{N-a}) + \dfrac{a}{q-p}$ if $p \neq q$.
6. Use Theorem 10.12.
8. For the final part, the general argument in Problem 10.5.6 may be useful. See also Section 12.5.
10. Either condition on the value of D_{n-1}, or write $D_n^2 = X_n^2 + Y_n^2$, in the obvious notation.
11. The answer to the first part is $\left[\left(1 - \sqrt{1-s^2}\right)/s\right]^n$. The remaining part of the question is not dissimilar to Problem 10.5.9.
12. This is the three-dimensional version of Problem 10.5.8. See also Section 12.5.
15. Use the result of Problem 10.5.14, or the lack-of-memory property of the geometric distribution.

Chapter 11

1. Poisson, with parameter $\lambda(t-s)$.
2. Condition on N_s.
3. $\frac{1}{6}$.
4. To obtain the integral equation, condition on X_1. The solution of the integral equation is $g(t,x) = e^{-\lambda x}$.
5. Use the random sum formula, Theorem 4.36.
8. The time between the ith and $(i+1)$th birth has the exponential distribution with parameter λi.
10. Find $\mathbb{P}(T \leq t) = \mathbb{P}(L_t = 0)$ from the result of Theorem 11.53.
11. To find the mean $m(t)$, differentiate throughout with respect to z and set $z = 1$, to obtain

$$\left.\frac{\partial^2 \phi}{\partial z \partial t}\right|_{z=1} = \phi(1,t),$$

giving $m'(t) = 1$ since $m(t) = \partial\phi/\partial z\big|_{z=1}$. Hence, $m(t) = t$. The variance may be found similarly.

14. The following argument is not completely rigorous but is illuminating. Let $t \to \infty$ in the formula for $G(s,t)$ given in Problem 11.7.13 to obtain $G(s,t) \to \exp[\rho(s-1)]$ where $\rho = \theta/\mu$. This is the probability generating function of the Poisson distribution. Rewrite this in terms of characteristic functions and appeal to the continuity theorem for a watertight argument.
15. $p(t) = e^{-\lambda t}\cosh \lambda t = q(t)$. The time to the first change of state has the exponential distribution with parameter λ. Use independence for the last part.
16. Condition on the length of the service time.
22. (c) The mean total is $\frac{1}{2}t\mathbb{E}(N_t) = \frac{1}{2}\lambda t^2$, by either calculation or symmetry.

Chapter 12

1. $\pi_i = 1/N$ for $i \in S$.
2. Recall Example 12.15. The chain is reversible in equilibrium when $0 < \alpha\beta < 1$.
3. In 'physical' models of this type, one can start by looking for a solution to the detailed balance equations. In this case, we have $\pi_i = \binom{N}{i}^2 \Big/ \binom{2N}{N}$.
4. The chain is irreducible and $\mathbb{P}_0(T > n) = b_n$, where T is the first return time of 0. Recall Problem 2.6.6. The final answer is $\pi_j = b_j \Big/ \sum_i b_i$.
5. The Markov property holds by the lack-of-memory property of the geometric distribution (Problem 2.6.5). For the invariant distribution, look for a solution of the detailed balance equations (you may pick an easy equation to start).
6. Either solve the equation $\pi = \pi P$, or argue as follows. If the state is ijk, then i was the last book chosen and, of the books j and k, j was the last chosen. Therefore, $\pi_{ijk} = \alpha_i\alpha_j/(\alpha_j + \alpha_k)$.
7. Assume reversibility. By passing the π term along the product, we may see that $\pi_{i_1}p_{i_1,i_2}p_{i_2,i_3}\cdots p_{i_n,i_1} = p_{i_2,i_1}p_{i_3,i_2}\cdots p_{i_1,i_n}\pi_{i_1}$. For the converse, sum over the intermediate $j_2, \ldots, j_{n-1}$ to obtain $p_{j_1,j_n}(n-1)p_{j_n,j_1} = p_{j_1,j_n}p_{j_n,j_1}(n-1)$, and let $n \to \infty$.
8. Either slog it out, or consider a collapsed chain with two states, 1 and 0 (representing 'not 1'). Now use the result of Problem 12.13.2.
9. (a) 8, (b) 1, (c) 10. Use Theorem 12.57 and symmetry for the last part.
10. Check that Q and π are in detailed balance.
11. Use the fact that $U_n \geq x$ if and only if $W_{a(n)} \leq n$, where W_k is the time of the kth return to i, and $a(n) = \lceil (n/\mu) + x\sqrt{n\sigma^2/\mu^3} \rceil$. Now, $W_{a(n)}$ is the sum of $a(n)$ independent, identically distributed random variables, and $a(n)/n \to 1/\mu$.
12. $\frac{52}{1} + \frac{52}{2} + \cdots + \frac{52}{51} \leq 52(1 + \log 51)$. Use conditional probability and Markov's inequality for the last part.
13. The quick way is to look for a solution of the detailed balance equations.
14. Solve the equation $\pi = \pi P$ to find $\pi_k = \rho^k\pi_0$, where $\rho = p/q$. The chain is positive recurrent if and only if $p < q$. In order to distinguish between transience and null recurrence, consider a simple random walk on S at the times of leftward steps.
16. Bishops move diagonally. Starting at a corner, a bishop can reach 32 vertices, of which 14 have degree 7, 10 have degree 9, 6 have degree 11, and 2 have degree 13. Finally, 40.
17. (a) Solve recursively in terms of π_1, guess the answer, and prove by induction. Alternatively, read part (b).

Reading list

Charig, C. R., Webb, D. R., Payne, S. R., and Wickham, J. E. A. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shock wave lithotripsy. *British Med. J.*, **292**, 879–882.

Chung, K. L. (2001). *A Course in Probability Theory* (2nd edn). Academic Press, San Diego.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Chapman and Hall, London.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications* (2nd edn). Volume 2. John Wiley, New York.

Graham, R. L., Knuth, D. E., and Patashnik, O. (1994). *Concrete Mathematics* (2nd edn). Addison-Wesley, Reading, Mass.

Grimmett, G. R. and Stirzaker, D. R. (2001). *Probability and Random Processes* (3rd edn). Oxford University Press, Oxford.

Hall, M. (1967). *Combinatoral Theory*. Blaisdell, Waltham, Mass.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*. Academic Press, San Diego.

Norris, J. R. (1997). *Markov Chains*. Cambridge University Press, Cambridge.

Romik, D. (2000). Stirling's approximation for $n!$: the utimate short proof? *Amer. Math. Monthly*, **107**, 556–557.

Stirzaker, D. R. (1999). *Probability and Random Variables*. Cambridge University Press, Cambridge.

Stirzaker, D. R. (2003). *Elementary Probability* (2nd edn). Cambridge University Press, Cambridge.

Taylor, S. J. (1973). *Introduction to Measure and Integration*. Cambridge University Press, Cambridge.

Whittle, P. (2000). *Probability via Expectation* (4th edn). Springer, New York.

Index