OXFORD

Probability and Random Processes

GEOFFREY GRIMMETT and DAVID STIRZAKER

Third Edition

Probability and Random Processes

GEOFFREY R. GRIMMETT
Statistical Laboratory, University of Cambridge

and

DAVID R. STIRZAKER
Mathematical Institute, University of Oxford

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

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 in

Oxford New York

Athens Auckland Bangkok Bogotá Buenos Aires Cape Town
Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul Karachi
Kolkata Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
Paris São Paulo Shanghai Singapore Taipei Tokyo Toronto Warsaw

with associated companies in Berlin Ibadan

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First edition 1982
Second edition 1992
Third edition 2001

A catalogue record for this title is available from the British Library

Library of Congress Cataloging in Publication Data
Data available

ISBN 0 19 857223 9 [hardback]
ISBN 0 19 857222 0 [paperback]

10 9 8 7 6 5 4 3 2 1

Typeset by the authors
Printed in Great Britain
on acid-free paper by Biddles Ltd, Guildford & King's Lynn

Lastly, numbers are applicable even to such things as seem to be governed by no rule, I mean such as depend on chance: the quantity of probability and proportion of it in any two proposed cases being subject to calculation as much as anything else. Upon this depend the principles of game. We find sharpers know enough of this to cheat some men that would take it very ill to be thought bubbles; and one gamester exceeds another, as he has a greater sagacity and readiness in calculating his probability to win or lose in any particular case. To understand the theory of chance thoroughly, requires a great knowledge of numbers, and a pretty competent one of Algebra.

John Arbuthnot
An essay on the usefulness of mathematical learning
25 November 1700

To this may be added, that some of the problems about chance having a great appearance of simplicity, the mind is easily drawn into a belief, that their solution may be attained by the mere strength of natural good sense; which generally proving otherwise, and the mistakes occasioned thereby being not infrequent, it is presumed that a book of this kind, which teaches to distinguish truth from what seems so nearly to resemble it, will be looked on as a help to good reasoning.

Abraham de Moivre
The Doctrine of Chances
1717

Preface to the Third Edition

This book provides an extensive introduction to probability and random processes. It is intended for those working in the many and varied applications of the subject as well as for those studying more theoretical aspects. We hope it will be found suitable for mathematics undergraduates at all levels, as well as for graduate students and others with interests in these fields.

In particular, we aim:

- to give a rigorous introduction to probability theory while limiting the amount of measure theory in the early chapters;
- to discuss the most important random processes in some depth, with many examples;
- to include various topics which are suitable for undergraduate courses, but are not routinely taught;
- to impart to the beginner the flavour of more advanced work, thereby whetting the appetite for more.

The ordering and numbering of material in this third edition has for the most part been preserved from the second. However, a good many minor alterations and additions have been made in the pursuit of clearer exposition. Furthermore, we have included new sections on sampling and Markov chain Monte Carlo, coupling and its applications, geometrical probability, spatial Poisson processes, stochastic calculus and the Itô integral, Itô's formula and applications, including the Black–Scholes formula, networks of queues, and renewal–reward theorems and applications. In a mild manifestation of millennial mania, the number of exercises and problems has been increased to exceed 1000. These are not merely drill exercises, but complement and illustrate the text, or are entertaining, or (usually, we hope) both. In a companion volume *One Thousand Exercises in Probability* (Oxford University Press, 2001), we give worked solutions to almost all exercises and problems.

The basic layout of the book remains unchanged. Chapters 1–5 begin with the foundations of probability theory, move through the elementary properties of random variables, and finish with the weak law of large numbers and the central limit theorem; on route, the reader meets random walks, branching processes, and characteristic functions. This material is suitable for about two lecture courses at a moderately elementary level. The rest of the book is largely concerned with random processes. Chapter 6 deals with Markov chains, treating discrete-time chains in some detail (and including an easy proof of the ergodic theorem for chains with countably infinite state spaces) and treating continuous-time chains largely by example. Chapter 7 contains a general discussion of convergence, together with simple but rigorous

accounts of the strong law of large numbers, and martingale convergence. Each of these two chapters could be used as a basis for a lecture course. Chapters 8–13 are more fragmented and provide suitable material for about five shorter lecture courses on: stationary processes and ergodic theory; renewal processes; queues; martingales; diffusions and stochastic integration with applications to finance.

We thank those who have read and commented upon sections of this and earlier editions, and we make special mention of Dominic Welsh, Brian Davies, Tim Brown, Sean Collins, Stephen Suen, Geoff Eagleson, Harry Reuter, David Green, and Bernard Silverman for their contributions to the first edition.

Of great value in the preparation of the second and third editions were the detailed criticisms of Michel Dekking, Frank den Hollander, Torgny Lindvall, and the suggestions of Alan Bain, Erwin Bolthausen, Peter Clifford, Frank Kelly, Doug Kennedy, Colin McDiarmid, and Volker Priebe. Richard Buxton has helped us with classical matters, and Andy Burbanks with the design of the front cover, which depicts a favourite confluence of the authors.

This edition having been reset in its entirety, we would welcome help in thinning the errors should any remain after the excellent TEX-ing of Sarah Shea-Simonds and Julia Blackwell.

Cambridge and Oxford G. R. G.
April 2001 D. R. S.

Contents

4 Continuous random variables

5 Generating functions and their applications

6 Markov chains

1

Events and their probabilities

Summary. Any experiment involving randomness can be modelled as a probability space. Such a space comprises a set Ω of possible outcomes of the experiment, a set $\mathcal{F}$ of events, and a probability measure $\mathbb{P}$. The definition and basic properties of a probability space are explored, and the concepts of conditional probability and independence are introduced. Many examples involving modelling and calculation are included.

1.1 Introduction

Much of our life is based on the belief that the future is largely unpredictable. For example, games of chance such as dice or roulette would have few adherents if their outcomes were known in advance. We express this belief in chance behaviour by the use of words such as 'random' or 'probability', and we seek, by way of gaming and other experience, to assign quantitative as well as qualitative meanings to such usages. Our main acquaintance with statements about probability relies on a wealth of concepts, some more reasonable than others. A mathematical theory of probability will incorporate those concepts of chance which are expressed and implicit in common rational understanding. Such a theory will formalize these concepts as a collection of axioms, which should lead directly to conclusions in agreement with practical experimentation. This chapter contains the essential ingredients of this construction.

1.2 Events as sets

Many everyday statements take the form 'the chance (or probability) of A is p', where A is some event (such as 'the sun shining tomorrow', 'Cambridge winning the Boat Race', ...) and p is a number or adjective describing quantity (such as 'one-eighth', 'low', ...). The occurrence or non-occurrence of A depends upon the chain of circumstances involved. This chain is called an *experiment* or *trial*; the result of an experiment is called its *outcome*. In general, we cannot predict with certainty the outcome of an experiment in advance of its completion; we can only list the collection of possible outcomes.

(1) Definition. The set of all possible outcomes of an experiment is called the **sample space** and is denoted by Ω.

(2) Example. A coin is tossed. There are two possible outcomes, heads (denoted by H) and tails (denoted by T), so that $\Omega = \{\mathrm{H}, \mathrm{T}\}$. We may be interested in the possible occurrences of the following events:

(a) the outcome is a head;

(b) the outcome is either a head or a tail;

(c) the outcome is both a head and a tail (this seems very unlikely to occur);

(d) the outcome is not a head. ●

(3) Example. A die is thrown once. There are six possible outcomes depending on which of the numbers 1, 2, 3, 4, 5, or 6 is uppermost. Thus $\Omega = \{1, 2, 3, 4, 5, 6\}$. We may be interested in the following events:

(a) the outcome is the number 1;

(b) the outcome is an even number;

(c) the outcome is even but does not exceed 3;

(d) the outcome is not even. ●

We see immediately that each of the events of these examples can be specified as a subset A of the appropriate sample space Ω. In the first example they can be rewritten as

(a) $A = \{\mathrm{H}\}$, (b) $A = \{\mathrm{H}\} \cup \{\mathrm{T}\}$,

(c) $A = \{\mathrm{H}\} \cap \{\mathrm{T}\}$, (d) $A = \{\mathrm{H}\}^c$,

whilst those of the second example become

(a) $A = \{1\}$, (b) $A = \{2, 4, 6\}$,

(c) $A = \{2, 4, 6\} \cap \{1, 2, 3\}$, (d) $A = \{2, 4, 6\}^c$.

The *complement* of a subset A of Ω is denoted here and subsequently by A^c; from now on, subsets of Ω containing a single member, such as $\{\mathrm{H}\}$, will usually be written without the containing braces.

Henceforth we think of *events* as subsets of the sample space Ω. Whenever A and B are events in which we are interested, then we can reasonably concern ourselves also with the events $A \cup B$, $A \cap B$, and A^c, representing 'A or B', 'A and B', and 'not A' respectively. Events A and B are called *disjoint* if their intersection is the empty set $\varnothing$; $\varnothing$ is called the *impossible event*. The set Ω is called the *certain event*, since some member of Ω will certainly occur.

Thus events are subsets of Ω, but need all the subsets of Ω be events? The answer is *no*, but some of the reasons for this are too difficult to be discussed here. It suffices for us to think of the collection of events as a subcollection $\mathcal{F}$ of the set of all subsets of Ω. This subcollection should have certain properties in accordance with the earlier discussion:

(a) if $A, B \in \mathcal{F}$ then $A \cup B \in \mathcal{F}$ and $A \cap B \in \mathcal{F}$;

(b) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$;

(c) the empty set $\varnothing$ belongs to $\mathcal{F}$.

Any collection $\mathcal{F}$ of subsets of Ω which satisfies these three conditions is called a *field*. It follows from the properties of a field $\mathcal{F}$ that

$$\text{if} \quad A_1, A_2, \ldots, A_n \in \mathcal{F} \quad \text{then} \quad \bigcup_{i=1}^{n} A_i \in \mathcal{F};$$

Typical notation	Set jargon	Probability jargon
Ω	Collection of objects	Sample space
ω	Member of Ω	Elementary event, outcome
A	Subset of Ω	Event that some outcome in A occurs
A^c	Complement of A	Event that no outcome in A occurs
$A \cap B$	Intersection	Both A and B
$A \cup B$	Union	Either A or B or both
$A \setminus B$	Difference	A, but not B
$A \triangle B$	Symmetric difference	Either A or B, but not both
$A \subseteq B$	Inclusion	If A, then B
$\varnothing$	Empty set	Impossible event
Ω	Whole space	Certain event

Table 1.1. The jargon of set theory and probability theory.

that is to say, $\mathcal{F}$ is closed under finite unions and hence under finite intersections also (see Problem (1.8.3)). This is fine when Ω is a finite set, but we require slightly more to deal with the common situation when Ω is infinite, as the following example indicates.

(4) Example. A coin is tossed repeatedly until the first head turns up; we are concerned with the number of tosses before this happens. The set of all possible outcomes is the set $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$, where ω_i denotes the outcome when the first $i-1$ tosses are tails and the ith toss is a head. We may seek to assign a probability to the event A, that the first head occurs after an even number of tosses, that is, $A = \{\omega_2, \omega_4, \omega_6, \dots\}$. This is an infinite countable union of members of Ω and we require that such a set belong to $\mathcal{F}$ in order that we can discuss its probability. ●

Thus we also require that the collection of events be closed under the operation of taking countable unions. Any collection of subsets of Ω with these properties is called a *σ-field*.

(5) Definition. A collection $\mathcal{F}$ of subsets of Ω is called a **σ-field** if it satisfies the following conditions:

(a) $\varnothing \in \mathcal{F}$;

(b) if $A_1, A_2, \dots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$;

(c) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

It follows from Problem (1.8.3) that σ-fields are closed under the operation of taking countable intersections. Here are some examples of σ-fields.

(6) Example. The smallest σ-field associated with Ω is the collection $\mathcal{F} = \{\varnothing, \Omega\}$. ●

(7) Example. If A is any subset of Ω then $\mathcal{F} = \{\varnothing, A, A^c, \Omega\}$ is a σ-field. ●

(8) Example. The *power set* of Ω, which is written $\{0,1\}^{\Omega}$ and contains all subsets of Ω, is obviously a σ-field. For reasons beyond the scope of this book, when Ω is infinite, its power set is too large a collection for probabilities to be assigned reasonably to all its members. ●

To recapitulate, with any experiment we may associate a pair $(\Omega, \mathcal{F})$, where Ω is the set of all possible outcomes or *elementary events* and $\mathcal{F}$ is a σ-field of subsets of Ω which contains all the events in whose occurrences we may be interested; henceforth, to call a set A an *event* is equivalent to asserting that A belongs to the σ-field in question. We usually translate statements about combinations of events into set-theoretic jargon; for example, the event that both A and B occur is written as $A \cap B$. Table 1.1 is a translation chart.

Exercises for Section 1.2

1. Let $\{A_i : i \in I\}$ be a collection of sets. Prove 'De Morgan's Laws'†:

$$\Big(\bigcup_i A_i\Big)^c = \bigcap_i A_i^c, \qquad \Big(\bigcap_i A_i\Big)^c = \bigcup_i A_i^c.$$

2. Let A and B belong to some σ-field $\mathcal{F}$. Show that $\mathcal{F}$ contains the sets $A \cap B$, $A \setminus B$, and $A \triangle B$.

3. A conventional knock-out tournament (such as that at Wimbledon) begins with 2^n competitors and has n rounds. There are no play-offs for the positions 2, 3, ..., $2^n - 1$, and the initial table of draws is specified. Give a concise description of the sample space of all possible outcomes.

4. Let $\mathcal{F}$ be a σ-field of subsets of Ω and suppose that $B \in \mathcal{F}$. Show that $\mathcal{G} = \{A \cap B : A \in \mathcal{F}\}$ is a σ-field of subsets of B.

5. Which of the following are identically true? For those that are not, say when they are true.
(a) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$;
(b) $A \cap (B \cap C) = (A \cap B) \cap C$;
(c) $(A \cup B) \cap C = A \cup (B \cap C)$;
(d) $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$.

1.3 Probability

We wish to be able to discuss the likelihoods of the occurrences of events. Suppose that we repeat an experiment a large number N of times, keeping the initial conditions as equal as possible, and suppose that A is some event which may or may not occur on each repetition. Our experience of most scientific experimentation is that the proportion of times that A occurs settles down to some value as N becomes larger and larger; that is to say, writing $N(A)$ for the number of occurrences of A in the N trials, the ratio $N(A)/N$ appears to converge to a constant limit as N increases. We can think of the ultimate value of this ratio as being the probability $\mathbb{P}(A)$ that A occurs on any particular trial‡; it may happen that the empirical ratio does not behave in a coherent manner and our intuition fails us at this level, but we shall not discuss this here. In practice, N may be taken to be large but finite, and the ratio $N(A)/N$ may be taken as an approximation to $\mathbb{P}(A)$. Clearly, the ratio is a number between zero and one; if $A = \varnothing$ then $N(\varnothing) = 0$ and the ratio is 0, whilst if $A = \Omega$ then $N(\Omega) = N$ and the

†Augustus De Morgan is well known for having given the first clear statement of the principle of mathematical induction. He applauded probability theory with the words: "The tendency of our study is to substitute the satisfaction of mental exercise for the pernicious enjoyment of an immoral stimulus".

‡This superficial discussion of probabilities is inadequate in many ways; questioning readers may care to discuss the philosophical and empirical aspects of the subject amongst themselves (see Appendix III).

ratio is 1. Furthermore, suppose that A and B are two disjoint events, each of which may or may not occur at each trial. Then

$$N(A \cup B) = N(A) + N(B)$$

and so the ratio $N(A \cup B)/N$ is the sum of the two ratios $N(A)/N$ and $N(B)/N$. We now think of these ratios as representing the probabilities of the appropriate events. The above relations become

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B), \qquad \mathbb{P}(\varnothing) = 0, \qquad \mathbb{P}(\Omega) = 1.$$

This discussion suggests that the probability function $\mathbb{P}$ should be *finitely additive*, which is to say that

$$\text{if } A_1, A_2, \ldots, A_n \text{ are disjoint events, then } \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i);$$

a glance at Example (1.2.4) suggests the more extensive property that $\mathbb{P}$ be *countably additive*, in that the corresponding property should hold for countable collections $A_1, A_2, \ldots$ of disjoint events.

These relations are sufficient to specify the desirable properties of a probability function $\mathbb{P}$ applied to the set of events. Any such assignment of likelihoods to the members of $\mathcal{F}$ is called a *probability measure*. Some individuals refer informally to $\mathbb{P}$ as a 'probability distribution', especially when the sample space is finite or countably infinite; this practice is best avoided since the term 'probability distribution' is reserved for another purpose to be encountered in Chapter 2.

(1) Definition. A **probability measure** $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfying

(a) $\mathbb{P}(\varnothing) = 0, \quad \mathbb{P}(\Omega) = 1;$

(b) if $A_1, A_2, \ldots$ is a collection of disjoint members of $\mathcal{F}$, in that $A_i \cap A_j = \varnothing$ for all pairs i, j satisfying $i \neq j$, then

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

The triple $(\Omega, \mathcal{F}, \mathbb{P})$, comprising a set Ω, a σ-field $\mathcal{F}$ of subsets of Ω, and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$, is called a **probability space**.

A probability measure is a special example of what is called a *measure* on the pair $(\Omega, \mathcal{F})$. A measure is a function $\mu : \mathcal{F} \to [0, \infty)$ satisfying $\mu(\varnothing) = 0$ together with (b) above. A measure μ is a probability measure if $\mu(\Omega) = 1$.

We can associate a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with any experiment, and all questions associated with the experiment can be reformulated in terms of this space. It may seem natural to ask for the numerical value of the probability $\mathbb{P}(A)$ of some event A. The answer to such a question must be contained in the description of the experiment in question. For example, the assertion that a *fair* coin is tossed once is equivalent to saying that heads and tails have an equal probability of occurring; actually, this is the definition of fairness.

(2) Example. A coin, possibly biased, is tossed once. We can take $\Omega = \{\mathrm{H}, \mathrm{T}\}$ and $\mathcal{F} = \{\varnothing, \mathrm{H}, \mathrm{T}, \Omega\}$, and a possible probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ is given by

$$\mathbb{P}(\varnothing) = 0, \quad \mathbb{P}(\mathrm{H}) = p, \quad \mathbb{P}(\mathrm{T}) = 1 - p, \quad \mathbb{P}(\Omega) = 1,$$

where p is a fixed real number in the interval $[0, 1]$. If $p = \frac{1}{2}$, then we say that the coin is *fair*, or *unbiased*. ●

(3) Example. A die is thrown once. We can take $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = \{0, 1\}^{\Omega}$, and the probability measure $\mathbb{P}$ given by

$$\mathbb{P}(A) = \sum_{i \in A} p_i \quad \text{for any } A \subseteq \Omega,$$

where $p_1, p_2, \ldots, p_6$ are specified numbers from the interval $[0, 1]$ having unit sum. The probability that i turns up is p_i. The die is fair if $p_i = \frac{1}{6}$ for each i, in which case

$$\mathbb{P}(A) = \tfrac{1}{6}|A| \quad \text{for any } A \subseteq \Omega,$$

where $|A|$ denotes the cardinality of A. ●

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ denotes a typical probability space. We now give some of its simple but important properties.

(4) Lemma.

(a) $\mathbb{P}(A^{\mathrm{c}}) = 1 - \mathbb{P}(A)$,

(b) *if* $B \supseteq A$ *then* $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$,

(c) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$,

(d) *more generally, if* $A_1, A_2, \ldots, A_n$ *are events, then*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n)$$

where, for example, $\sum_{i<j}$ *sums over all unordered pairs* (i, j) *with* $i \neq j$.

Proof.

(a) $A \cup A^{\mathrm{c}} = \Omega$ and $A \cap A^{\mathrm{c}} = \varnothing$, so $\mathbb{P}(A \cup A^{\mathrm{c}}) = \mathbb{P}(A) + \mathbb{P}(A^{\mathrm{c}}) = 1$.

(b) $B = A \cup (B \setminus A)$. This is the union of disjoint sets and therefore

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A).$$

(c) $A \cup B = A \cup (B \setminus A)$, which is a disjoint union. Therefore, by (b),

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(A) + \mathbb{P}(B \setminus (A \cap B)) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \end{aligned}$$

(d) The proof is by induction on n, and is left as an *exercise* (see Exercise (1.3.4)). ■

In Lemma (4b), $B \setminus A$ denotes the set of members of B which are not in A. In order to write down the quantity $\mathbb{P}(B \setminus A)$, we require that $B \setminus A$ belongs to $\mathcal{F}$, the domain of $\mathbb{P}$; this is always true when A and B belong to $\mathcal{F}$, and to prove this was part of Exercise (1.2.2). Notice that each proof proceeded by expressing an event in terms of disjoint unions and then applying $\mathbb{P}$. It is sometimes easier to calculate the probabilities of intersections of events rather than their unions; part (d) of the lemma is useful then, as we shall discover soon. The next property of $\mathbb{P}$ is more technical, and says that $\mathbb{P}$ is a *continuous* set function; this property is essentially equivalent to the condition that $\mathbb{P}$ is countably additive rather than just finitely additive (see Problem (1.8.16) also).

(5) Lemma. *Let $A_1, A_2, \ldots$ be an increasing sequence of events, so that $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, and write A for their limit*:

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

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

Similarly, if $B_1, B_2, \ldots$ is a decreasing sequence of events, so that $B_1 \supseteq B_2 \supseteq B_3 \supseteq \cdots$, then

$$B = \bigcap_{i=1}^{\infty} B_i = \lim_{i \to \infty} B_i$$

satisfies $\mathbb{P}(B) = \lim_{i \to \infty} \mathbb{P}(B_i)$.

Proof. $A = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \cdots$ is the union of a disjoint family of events. Thus, by Definition (1),

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

To show the result for decreasing families of events, take complements and use the first part (*exercise*). ■

To recapitulate, statements concerning chance are implicitly related to experiments or trials, the outcomes of which are not entirely predictable. With any such experiment we can associate a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ the properties of which are consistent with our shared and reasonable conceptions of the notion of chance.

Here is some final jargon. An event A is called *null* if $\mathbb{P}(A) = 0$. If $\mathbb{P}(A) = 1$, we say that A occurs *almost surely*. Null events should not be confused with the impossible event $\varnothing$. Null events are happening all around us, even though they have zero probability; after all, what is the chance that a dart strikes any given point of the target at which it is thrown? That is, the impossible event is null, but null events need not be impossible.

Exercises for Section 1.3

1. Let A and B be events with probabilities $\mathbb{P}(A) = \frac{3}{4}$ and $\mathbb{P}(B) = \frac{1}{3}$. Show that $\frac{1}{12} \leq \mathbb{P}(A \cap B) \leq \frac{1}{3}$, and give examples to show that both extremes are possible. Find corresponding bounds for $\mathbb{P}(A \cup B)$.

2. A fair coin is tossed repeatedly. Show that, with probability one, a head turns up sooner or later. Show similarly that any given finite sequence of heads and tails occurs eventually with probability one. Explain the connection with Murphy's Law.

3. Six cups and saucers come in pairs: there are two cups and saucers which are red, two white, and two with stars on. If the cups are placed randomly onto the saucers (one each), find the probability that no cup is upon a saucer of the same pattern.

4. Let $A_1, A_2, \ldots, A_n$ be events where $n \geq 2$, and prove that

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n).$$

In each packet of Corn Flakes may be found a plastic bust of one of the last five Vice-Chancellors of Cambridge University, the probability that any given packet contains any specific Vice-Chancellor being $\frac{1}{5}$, independently of all other packets. Show that the probability that each of the last three Vice-Chancellors is obtained in a bulk purchase of six packets is $1 - 3(\frac{4}{5})^6 + 3(\frac{3}{5})^6 - (\frac{2}{5})^6$.

5. Let $A_r, r \geq 1$, be events such that $\mathbb{P}(A_r) = 1$ for all r. Show that $\mathbb{P}\big(\bigcap_{r=1}^{\infty} A_r\big) = 1$.

6. You are given that at least one of the events A_r, $1 \leq r \leq n$, is certain to occur, but certainly no more than two occur. If $\mathbb{P}(A_r) = p$, and $\mathbb{P}(A_r \cap A_s) = q$, $r \neq s$, show that $p \geq 1/n$ and $q \leq 2/n$.

7. You are given that at least one, but no more than three, of the events A_r, $1 \leq r \leq n$, occur, where $n \geq 3$. The probability of at least two occurring is $\frac{1}{2}$. If $\mathbb{P}(A_r) = p$, $\mathbb{P}(A_r \cap A_s) = q$, $r \neq s$, and $\mathbb{P}(A_r \cap A_s \cap A_t) = x$, $r < s < t$, show that $p \geq 3/(2n)$, and $q \leq 4/n$.

1.4 Conditional probability

Many statements about chance take the form 'if B occurs, then the probability of A is p', where B and A are events (such as 'it rains tomorrow' and 'the bus being on time' respectively) and p is a likelihood as before. To include this in our theory, we return briefly to the discussion about proportions at the beginning of the previous section. An experiment is repeated N times, and on each occasion we observe the occurrences or non-occurrences of two events A and B. Now, suppose we only take an interest in those outcomes for which B occurs; all other experiments are disregarded. In this smaller collection of trials the proportion of times that A occurs is $N(A \cap B)/N(B)$, since B occurs at each of them. However,

$$\frac{N(A \cap B)}{N(B)} = \frac{N(A \cap B)/N}{N(B)/N}.$$

If we now think of these ratios as probabilities, we see that the probability that A occurs, given that B occurs, should be reasonably defined as $\mathbb{P}(A \cap B)/\mathbb{P}(B)$.

Probabilistic intuition leads to the same conclusion. Given that an event B occurs, it is the case that A occurs if and only if $A \cap B$ occurs. Thus the conditional probability of A given B

should be proportional to $\mathbb{P}(A \cap B)$, which is to say that it equals $\alpha\mathbb{P}(A \cap B)$ for some constant $\alpha = \alpha(B)$. The conditional probability of Ω given B must equal 1, and thus $\alpha\mathbb{P}(\Omega \cap B) = 1$, yielding $\alpha = 1/\mathbb{P}(B)$.

We formalize these notions as follows.

(1) Definition. If $\mathbb{P}(B) > 0$ then the **conditional probability** that A occurs given that B occurs is defined to be

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

We denote this conditional probability by $\mathbb{P}(A \mid B)$, pronounced 'the probability of A given B', or sometimes 'the probability of A conditioned (or conditional) on B'.

(2) Example. Two fair dice are thrown. Given that the first shows 3, what is the probability that the total exceeds 6? The answer is obviously $\frac{1}{2}$, since the second must show 4, 5, or 6. However, let us labour the point. Clearly $\Omega = \{1, 2, 3, 4, 5, 6\}^2$, the set† of all ordered pairs (i, j) for $i, j \in \{1, 2, \ldots, 6\}$, and we can take $\mathcal{F}$ to be the set of all subsets of Ω, with $\mathbb{P}(A) = |A|/36$ for any $A \subseteq \Omega$. Let B be the event that the first die shows 3, and A be the event that the total exceeds 6. Then

$$B = \{(3, b) : 1 \leq b \leq 6\}, \quad A = \{(a, b) : a + b > 6\}, \quad A \cap B = \{(3, 4), (3, 5), (3, 6)\},$$

and

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{|A \cap B|}{|B|} = \frac{3}{6}. \qquad \bullet$$

(3) Example. A family has two children. What is the probability that both are boys, given that at least one is a boy? The older and younger child may each be male or female, so there are four possible combinations of sexes, which we assume to be equally likely. Hence we can represent the sample space in the obvious way as

$$\Omega = \{\mathrm{GG}, \mathrm{GB}, \mathrm{BG}, \mathrm{BB}\}$$

where $\mathbb{P}(\mathrm{GG}) = \mathbb{P}(\mathrm{BB}) = \mathbb{P}(\mathrm{GB}) = \mathbb{P}(\mathrm{BG}) = \frac{1}{4}$. From the definition of conditional probability,

$$\begin{aligned}\mathbb{P}(\mathrm{BB} \mid \text{one boy at least}) &= \mathbb{P}(\mathrm{BB} \mid \mathrm{GB} \cup \mathrm{BG} \cup \mathrm{BB}) \\ &= \frac{\mathbb{P}(\mathrm{BB} \cap (\mathrm{GB} \cup \mathrm{BG} \cup \mathrm{BB}))}{\mathbb{P}(\mathrm{GB} \cup \mathrm{BG} \cup \mathrm{BB})} \\ &= \frac{\mathbb{P}(\mathrm{BB})}{\mathbb{P}(\mathrm{GB} \cup \mathrm{BG} \cup \mathrm{BB})} = \frac{1}{3}.\end{aligned}$$

A popular but incorrect answer to the question is $\frac{1}{2}$. This is the correct answer to another question: for a family with two children, what is the probability that both are boys given that the younger is a boy? In this case,

$$\begin{aligned}\mathbb{P}(\mathrm{BB} \mid \text{younger is a boy}) &= \mathbb{P}(\mathrm{BB} \mid \mathrm{GB} \cup \mathrm{BB}) \\ &= \frac{\mathbb{P}(\mathrm{BB} \cap (\mathrm{GB} \cup \mathrm{BB}))}{\mathbb{P}(\mathrm{GB} \cup \mathrm{BB})} = \frac{\mathbb{P}(\mathrm{BB})}{\mathbb{P}(\mathrm{GB} \cup \mathrm{BB})} = \frac{1}{2}.\end{aligned}$$

†Remember that $A \times B = \{(a, b) : a \in A, \ b \in B\}$ and that $A \times A = A^2$.

The usual dangerous argument contains the assertion

$$\mathbb{P}(\text{BB} \mid \text{one child is a boy}) = \mathbb{P}(\text{other child is a boy}).$$

Why is this meaningless? [Hint: Consider the sample space.] ●

The next lemma is crucially important in probability theory. A family $B_1, B_2, \ldots, B_n$ of events is called a *partition* of the set Ω if

$$B_i \cap B_j = \varnothing \quad \text{when} \quad i \neq j, \quad \text{and} \quad \bigcup_{i=1}^{n} B_i = \Omega.$$

Each elementary event $\omega \in \Omega$ belongs to exactly one set in a partition of Ω.

(4) Lemma. *For any events A and B such that $0 < \mathbb{P}(B) < 1$,*

$$\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c).$$

More generally, let $B_1, B_2, \ldots, B_n$ be a partition of Ω such that $\mathbb{P}(B_i) > 0$ for all i. Then

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

Proof. $A = (A \cap B) \cup (A \cap B^c)$. This is a disjoint union and so

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \\ &= \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c). \end{aligned}$$

The second part is similar (see Problem (1.8.10)). ■

(5) Example. We are given two urns, each containing a collection of coloured balls. Urn I contains two white and three blue balls, whilst urn II contains three white and four blue balls. A ball is drawn at random from urn I and put into urn II, and then a ball is picked at random from urn II and examined. What is the probability that it is blue? We assume unless otherwise specified that a ball picked randomly from any urn is equally likely to be any of those present. The reader will be relieved to know that we no longer need to describe $(\Omega, \mathcal{F}, \mathbb{P})$ in detail; we are confident that we could do so if necessary. Clearly, the colour of the final ball depends on the colour of the ball picked from urn I. So let us 'condition' on this. Let A be the event that the final ball is blue, and let B be the event that the first one picked was blue. Then, by Lemma (4),

$$\mathbb{P}(A) = \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c).$$

We can easily find all these probabilities:

$$\begin{aligned} \mathbb{P}(A \mid B) &= \mathbb{P}(A \mid \text{urn II contains three white and five blue balls}) = \tfrac{5}{8}, \\ \mathbb{P}(A \mid B^c) &= \mathbb{P}(A \mid \text{urn II contains four white and four blue balls}) = \tfrac{1}{2}, \\ \mathbb{P}(B) &= \tfrac{3}{5}, \qquad \mathbb{P}(B^c) = \tfrac{2}{5}. \end{aligned}$$

Hence

$$\mathbb{P}(A) = \tfrac{5}{8} \cdot \tfrac{3}{5} + \tfrac{1}{2} \cdot \tfrac{2}{5} = \tfrac{23}{40}. \qquad \bullet$$

Unprepared readers may have been surprised by the sudden appearance of urns in this book. In the seventeenth and eighteenth centuries, lotteries often involved the drawing of slips from urns, and voting was often a matter of putting slips or balls into urns. In France today, *aller aux urnes* is synonymous with voting. It was therefore not unnatural for the numerous Bernoullis and others to model births, marriages, deaths, fluids, gases, and so on, using urns containing balls of varied hue.

(6) Example. Only two factories manufacture zoggles. 20 per cent of the zoggles from factory I and 5 per cent from factory II are defective. Factory I produces twice as many zoggles as factory II each week. What is the probability that a zoggle, randomly chosen from a week's production, is satisfactory? Clearly this satisfaction depends on the factory of origin. Let A be the event that the chosen zoggle is satisfactory, and let B be the event that it was made in factory I. Arguing as before,

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid B^c)\mathbb{P}(B^c) \\ &= \tfrac{4}{5} \cdot \tfrac{2}{3} + \tfrac{19}{20} \cdot \tfrac{1}{3} = \tfrac{51}{60}. \end{aligned}$$

If the chosen zoggle is defective, what is the probability that it came from factory I? In our notation this is just $\mathbb{P}(B \mid A^c)$. However,

$$\mathbb{P}(B \mid A^c) = \frac{\mathbb{P}(B \cap A^c)}{\mathbb{P}(A^c)} = \frac{\mathbb{P}(A^c \mid B)\mathbb{P}(B)}{\mathbb{P}(A^c)} = \frac{\frac{1}{5} \cdot \frac{2}{3}}{1 - \frac{51}{60}} = \frac{8}{9}. \qquad \bullet$$

This section is terminated with a cautionary example. It is not untraditional to perpetuate errors of logic in calculating conditional probabilities. Lack of unambiguous definitions and notation has led astray many probabilists, including even Boole, who was credited by Russell with the discovery of pure mathematics and by others for some of the logical foundations of computing. The well-known 'prisoners' paradox' also illustrates some of the dangers here.

(7) Example. Prisoners' paradox. In a dark country, three prisoners have been incarcerated without trial. Their warder tells them that the country's dictator has decided arbitrarily to free one of them and to shoot the other two, but he is not permitted to reveal to any prisoner the fate of that prisoner. Prisoner A knows therefore that his chance of survival is $\frac{1}{3}$. In order to gain information, he asks the warder to tell him in secret the name of some prisoner (but not himself) who will be killed, and the warder names prisoner B. What now is prisoner A's assessment of the chance that he will survive? Could it be $\frac{1}{2}$: after all, he knows now that the survivor will be either A or C, and he has no information about which? Could it be $\frac{1}{3}$: after all, according to the rules, at least one of B and C has to be killed, and thus the extra information cannot reasonably affect A's earlier calculation of the odds? What does the reader think about this? The resolution of the paradox lies in the situation when either response (B or C) is possible.

An alternative formulation of this paradox has become known as the Monty Hall problem, the controversy associated with which has been provoked by Marilyn vos Savant (and many others) in *Parade* magazine in 1990; see Exercise (1.4.5). $\bullet$

Exercises for Section 1.4

1. Prove that $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)\mathbb{P}(A)/\mathbb{P}(B)$ whenever $\mathbb{P}(A)\mathbb{P}(B) \neq 0$. Show that, if $\mathbb{P}(A \mid B) > \mathbb{P}(A)$, then $\mathbb{P}(B \mid A) > \mathbb{P}(B)$.

2. For events $A_1, A_2, \ldots, A_n$ satisfying $\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$, prove that

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

3. A man possesses five coins, two of which are double-headed, one is double-tailed, and two are normal. He shuts his eyes, picks a coin at random, and tosses it. What is the probability that the lower face of the coin is a head?

He opens his eyes and sees that the coin is showing heads; what is the probability that the lower face is a head?

He shuts his eyes again, and tosses the coin again. What is the probability that the lower face is a head?

He opens his eyes and sees that the coin is showing heads; what is the probability that the lower face is a head?

He discards this coin, picks another at random, and tosses it. What is the probability that it shows heads?

4. What do you think of the following 'proof' by Lewis Carroll that an urn cannot contain two balls of the same colour? Suppose that the urn contains two balls, each of which is either black or white; thus, in the obvious notation, $\mathbb{P}(\mathrm{BB}) = \mathbb{P}(\mathrm{BW}) = \mathbb{P}(\mathrm{WB}) = \mathbb{P}(\mathrm{WW}) = \frac{1}{4}$. We add a black ball, so that $\mathbb{P}(\mathrm{BBB}) = \mathbb{P}(\mathrm{BBW}) = \mathbb{P}(\mathrm{BWB}) = \mathbb{P}(\mathrm{BWW}) = \frac{1}{4}$. Next we pick a ball at random; the chance that the ball is black is (using conditional probabilities) $1 \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{4} = \frac{2}{3}$. However, if there is probability $\frac{2}{3}$ that a ball, chosen randomly from three, is black, then there must be two black and one white, which is to say that originally there was one black and one white ball in the urn.

5. The Monty Hall problem: goats and cars. (a) Cruel fate has made you a contestant in a game show; you have to choose one of three doors. One conceals a new car, two conceal old goats. You choose, but your chosen door is not opened immediately. Instead, the presenter opens another door to reveal a goat, and he offers you the opportunity to change your choice to the third door (unopened and so far unchosen). Let p be the (conditional) probability that the third door conceals the car. The value of p depends on the presenter's protocol. Devise protocols to yield the values $p = \frac{1}{2}$, $p = \frac{2}{3}$. Show that, for $\alpha \in [\frac{1}{2}, \frac{2}{3}]$, there exists a protocol such that $p = \alpha$. Are you well advised to change your choice to the third door?

(b) In a variant of this question, the presenter is permitted to open the first door chosen, and to reward you with whatever lies behind. If he chooses to open another door, then this door invariably conceals a goat. Let p be the probability that the unopened door conceals the car, conditional on the presenter having chosen to open a second door. Devise protocols to yield the values $p = 0$, $p = 1$, and deduce that, for any $\alpha \in [0, 1]$, there exists a protocol with $p = \alpha$.

6. The prosecutor's fallacy†. Let G be the event that an accused is guilty, and T the event that some testimony is true. Some lawyers have argued on the assumption that $\mathbb{P}(G \mid T) = \mathbb{P}(T \mid G)$. Show that this holds if and only if $\mathbb{P}(G) = \mathbb{P}(T)$.

7. Urns. There are n urns of which the rth contains $r - 1$ red balls and $n - r$ magenta balls. You pick an urn at random and remove two balls at random without replacement. Find the probability that:

(a) the second ball is magenta;

(b) the second ball is magenta, given that the first is magenta.

---

†The prosecution made this error in the famous Dreyfus case of 1894.

1.5 Independence

In general, the occurrence of some event B changes the probability that another event A occurs, the original probability $\mathbb{P}(A)$ being replaced by $\mathbb{P}(A \mid B)$. If the probability remains unchanged, that is to say $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, then we call A and B 'independent'. This is well defined only if $\mathbb{P}(B) > 0$. Definition (1.4.1) of conditional probability leads us to the following.

(1) Definition. Events A and B are called **independent** if

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

More generally, a family $\{A_i : i \in I\}$ is called **independent** if

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

for all finite subsets J of I.

Remark. A common student error is to make the fallacious statement that A and B are independent if $A \cap B = \varnothing$.

If the family $\{A_i : i \in I\}$ has the property that

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \qquad \text{for all } i \neq j$$

then it is called *pairwise independent*. Pairwise-independent families are not necessarily independent, as the following example shows.

(2) Example. Suppose $\Omega = \{abc, acb, cab, cba, bca, bac, aaa, bbb, ccc\}$, and each of the nine elementary events in Ω occurs with equal probability $\frac{1}{9}$. Let A_k be the event that the kth letter is a. It is left as an *exercise* to show that the family $\{A_1, A_2, A_3\}$ is pairwise independent but not independent. ●

(3) Example (1.4.6) revisited. The events A and B of this example are clearly dependent because $\mathbb{P}(A \mid B) = \frac{4}{5}$ and $\mathbb{P}(A) = \frac{51}{60}$. ●

(4) Example. Choose a card at random from a pack of 52 playing cards, each being picked with equal probability $\frac{1}{52}$. We claim that the suit of the chosen card is independent of its rank. For example,

$$\mathbb{P}(\text{king}) = \tfrac{4}{52}, \quad \mathbb{P}(\text{king} \mid \text{spade}) = \tfrac{1}{13}.$$

Alternatively,

$$\mathbb{P}(\text{spade king}) = \tfrac{1}{52} = \tfrac{1}{4} \cdot \tfrac{1}{13} = \mathbb{P}(\text{spade})\mathbb{P}(\text{king}).$$ ●

Let C be an event with $\mathbb{P}(C) > 0$. To the conditional probability measure $\mathbb{P}(\cdot \mid C)$ corresponds the idea of *conditional independence*. Two events A and B are called *conditionally independent given* C if

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C); \tag{5}$$

there is a natural extension to families of events. [However, note Exercise (1.5.5).]

Exercises for Section 1.5

1. Let A and B be independent events; show that A^c, B are independent, and deduce that A^c, B^c are independent.

2. We roll a die n times. Let A_{ij} be the event that the ith and jth rolls produce the same number. Show that the events $\{A_{ij} : 1 \le i < j \le n\}$ are pairwise independent but not independent.

3. A fair coin is tossed repeatedly. Show that the following two statements are equivalent:
(a) the outcomes of different tosses are independent,
(b) for any given finite sequence of heads and tails, the chance of this sequence occurring in the first m tosses is 2^{-m}, where m is the length of the sequence.

4. Let $\Omega = \{1, 2, \ldots, p\}$ where p is prime, $\mathcal{F}$ be the set of all subsets of Ω, and $\mathbb{P}(A) = |A|/p$ for all $A \in \mathcal{F}$. Show that, if A and B are independent events, then at least one of A and B is either $\varnothing$ or Ω.

5. Show that the conditional independence of A and B given C neither implies, nor is implied by, the independence of A and B. For which events C is it the case that, for all A and B, the events A and B are independent if and only if they are conditionally independent given C?

6. Safe or sorry? Some form of prophylaxis is said to be 90 per cent effective at prevention during one year's treatment. If the degrees of effectiveness in different years are independent, show that the treatment is more likely than not to fail within 7 years.

7. Families. Jane has three children, each of which is equally likely to be a boy or a girl independently of the others. Define the events:

$$A = \{\text{all the children are of the same sex}\},$$
$$B = \{\text{there is at most one boy}\},$$
$$C = \{\text{the family includes a boy and a girl}\}.$$

(a) Show that A is independent of B, and that B is independent of C.
(b) Is A independent of C?
(c) Do these results hold if boys and girls are not equally likely?
(d) Do these results hold if Jane has four children?

8. Galton's paradox. You flip three fair coins. At least two are alike, and it is an evens chance that the third is a head or a tail. Therefore $\mathbb{P}(\text{all alike}) = \frac{1}{2}$. Do you agree?

9. Two fair dice are rolled. Show that the event that their sum is 7 is independent of the score shown by the first die.

1.6 Completeness and product spaces

This section should be omitted at the first reading, but we shall require its contents later. It contains only a sketch of complete probability spaces and product spaces; the reader should look elsewhere for a more detailed treatment (see Billingsley 1995). We require the following result.

(1) Lemma. *If $\mathcal{F}$ and $\mathcal{G}$ are two σ-fields of subsets of Ω then their intersection $\mathcal{F} \cap \mathcal{G}$ is a σ-field also. More generally, if $\{\mathcal{F}_i : i \in I\}$ is a family of σ-fields of subsets of Ω then $\mathcal{G} = \bigcap_{i \in I} \mathcal{F}_i$ is a σ-field also.*

The proof is not difficult and is left as an *exercise.* Note that the union $\mathcal{F} \cup \mathcal{G}$ may not be a σ-field, although it may be extended to a unique smallest σ-field written $\sigma(\mathcal{F} \cup \mathcal{G})$, as follows. Let $\{\mathcal{G}_i : i \in I\}$ be the collection of all σ-fields which contain both $\mathcal{F}$ and $\mathcal{G}$ as subsets; this collection is non-empty since it contains the set of all subsets of Ω. Then $\mathcal{G} = \bigcap_{i \in I} \mathcal{G}_i$ is the unique smallest σ-field which contains $\mathcal{F} \cup \mathcal{G}$.

(A) Completeness. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Any event A which has zero probability, that is $\mathbb{P}(A) = 0$, is called *null.* It may seem reasonable to suppose that any subset B of a null set A will itself be null, but this may be without meaning since B may not be an event, and thus $\mathbb{P}(B)$ may not be defined.

(2) Definition. A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called **complete** if all subsets of null sets are events.

Any incomplete space can be completed thus. Let $\mathcal{N}$ be the collection of all subsets of null sets in $\mathcal{F}$ and let $\mathcal{G} = \sigma(\mathcal{F} \cup \mathcal{N})$ be the smallest σ-field which contains all sets in $\mathcal{F}$ and $\mathcal{N}$. It can be shown that the domain of $\mathbb{P}$ may be extended in an obvious way from $\mathcal{F}$ to $\mathcal{G}$; $(\Omega, \mathcal{G}, \mathbb{P})$ is called the *completion* of $(\Omega, \mathcal{F}, \mathbb{P})$.

(B) Product spaces. The probability spaces discussed in this chapter have usually been constructed around the outcomes of one experiment, but instances occur naturally when we need to combine the outcomes of several independent experiments into one space (see Examples (1.2.4) and (1.4.2)). How should we proceed in general?

Suppose two experiments have associated probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ respectively. The sample space of the pair of experiments, considered jointly, is the collection $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \ \omega_2 \in \Omega_2\}$ of ordered pairs. The appropriate σ-field of events is more complicated to construct. Certainly it should contain all subsets of $\Omega_1 \times \Omega_2$ of the form $A_1 \times A_2 = \{(a_1, a_2) : a_1 \in A_1, \ a_2 \in A_2\}$ where A_1 and A_2 are typical members of $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively. However, the family of all such sets, $\mathcal{F}_1 \times \mathcal{F}_2 = \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, \ A_2 \in \mathcal{F}_2\}$, is not in general a σ-field. By the discussion after (1), there exists a unique smallest σ-field $\mathcal{G} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ of subsets of $\Omega_1 \times \Omega_2$ which contains $\mathcal{F}_1 \times \mathcal{F}_2$. All we require now is a suitable probability function on $(\Omega_1 \times \Omega_2, \mathcal{G})$. Let $\mathbb{P}_{12} : \mathcal{F}_1 \times \mathcal{F}_2 \to [0, 1]$ be given by:

$$\textbf{(3)} \qquad \mathbb{P}_{12}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2) \quad \text{for } A_1 \in \mathcal{F}_1, \ A_2 \in \mathcal{F}_2.$$

It can be shown that the domain of $\mathbb{P}_{12}$ can be extended from $\mathcal{F}_1 \times \mathcal{F}_2$ to the whole of $\mathcal{G} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$. The ensuing probability space $(\Omega_1 \times \Omega_2, \mathcal{G}, \mathbb{P}_{12})$ is called the *product space* of $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$. Products of larger numbers of spaces are constructed similarly. The measure $\mathbb{P}_{12}$ is sometimes called the 'product measure' since its defining equation (3) assumed that two experiments are independent. There are of course many other measures that can be applied to $(\Omega_1 \times \Omega_2, \mathcal{G})$.

In many simple cases this technical discussion is unnecessary. Suppose that Ω_1 and Ω_2 are finite, and that their σ-fields contain all their subsets; this is the case in Examples (1.2.4) and (1.4.2). Then $\mathcal{G}$ contains all subsets of $\Omega_1 \times \Omega_2$.

1.7 Worked examples

Here are some more examples to illustrate the ideas of this chapter. The reader is now equipped to try his or her hand at a substantial number of those problems which exercised the pioneers in probability. These frequently involved experiments having equally likely outcomes, such as dealing whist hands, putting balls of various colours into urns and taking them out again, throwing dice, and so on. In many such instances, the reader will be pleasantly surprised to find that it is not necessary to write down $(\Omega, \mathcal{F}, \mathbb{P})$ explicitly, but only to think of Ω as being a collection $\{\omega_1, \omega_2, \ldots, \omega_N\}$ of possibilities, each of which may occur with probability $1/N$. Thus, $\mathbb{P}(A) = |A|/N$ for any $A \subseteq \Omega$. The basic tools used in such problems are as follows.

(a) Combinatorics: remember that the number of permutations of n objects is $n!$ and that the number of ways of choosing r objects from n is $\binom{n}{r}$.

(b) Set theory: to obtain $\mathbb{P}(A)$ we can compute $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ or we can partition A by conditioning on events B_i, and then use Lemma (1.4.4).

(c) Use of independence.

(1) Example. Consider a series of hands dealt at bridge. Let A be the event that in a given deal each player has one ace. Show that the probability that A occurs at least once in seven deals is approximately $\frac{1}{2}$.

Solution. The number of ways of dealing 52 cards into four equal hands is $52!/(13!)^4$. There are 4! ways of distributing the aces so that each hand holds one, and there are $48!/(12!)^4$ ways of dealing the remaining cards. Thus

$$\mathbb{P}(A) = \frac{4!\,48!/(12!)^4}{52!/(13!)^4} \simeq \frac{1}{10}.$$

Now let B_i be the event that A occurs for the first time on the ith deal. Clearly $B_i \cap B_j = \varnothing$, $i \neq j$. Thus

$$\mathbb{P}(A \text{ occurs in seven deals}) = \mathbb{P}(B_1 \cup \cdots \cup B_7) = \sum_1^7 \mathbb{P}(B_i) \quad \text{using Definition (1.3.1).}$$

Since successive deals are independent, we have

$$\begin{aligned}\mathbb{P}(B_i) &= \mathbb{P}\big(A^c \text{ occurs on deal 1},\ A^c \text{ occurs on deal 2},\\ &\qquad \ldots,\ A^c \text{ occurs on deal } i-1,\ A \text{ occurs on deal } i\big)\\ &= \mathbb{P}(A^c)^{i-1}\mathbb{P}(A) \quad \text{using Definition (1.5.1)}\\ &\simeq \left(1 - \tfrac{1}{10}\right)^{i-1} \tfrac{1}{10}.\end{aligned}$$

Thus

$$\mathbb{P}(A \text{ occurs in seven deals}) = \sum_1^7 \mathbb{P}(B_i) \simeq \sum_1^7 \left(\tfrac{9}{10}\right)^{i-1} \tfrac{1}{10} \simeq \tfrac{1}{2}.$$

Can you see an easier way of obtaining this answer? ●

(2) Example. There are two roads from A to B and two roads from B to C. Each of the four roads has probability p of being blocked by snow, independently of all the others. What is the probability that there is an open road from A to C?

Solution.

$$\begin{aligned}\mathbb{P}(\text{open road}) &= \mathbb{P}\big((\text{open road from A to B}) \cap (\text{open road from B to C})\big)\\ &= \mathbb{P}(\text{open road from A to B})\mathbb{P}(\text{open road from B to C})\end{aligned}$$

using the independence. However, p is the same for all roads; thus, using Lemma (1.3.4),

$$\begin{aligned}\mathbb{P}(\text{open road}) &= \big(1 - \mathbb{P}(\text{no road from A to B})\big)^2\\ &= \big\{1 - \mathbb{P}\big((\text{first road blocked}) \cap (\text{second road blocked})\big)\big\}^2\\ &= \big\{1 - \mathbb{P}(\text{first road blocked})\mathbb{P}(\text{second road blocked})\big\}^2\end{aligned}$$

using the independence. Thus

$$\mathbb{P}(\text{open road}) = (1 - p^2)^2. \tag{3}$$

Further suppose that there is also a direct road from A to C, which is independently blocked with probability p. Then, by Lemma (1.4.4) and equation (3),

$$\begin{aligned}\mathbb{P}(\text{open road}) &= \mathbb{P}(\text{open road} \mid \text{direct road blocked}) \cdot p\\ &\quad + \mathbb{P}(\text{open road} \mid \text{direct road open}) \cdot (1 - p)\\ &= (1 - p^2)^2 \cdot p + 1 \cdot (1 - p).\end{aligned}$$

●

(4) Example. Symmetric random walk (or 'Gambler's ruin'). A man is saving up to buy a new Jaguar at a cost of N units of money. He starts with k units where $0 < k < N$, and tries to win the remainder by the following gamble with his bank manager. He tosses a fair coin repeatedly; if it comes up heads then the manager pays him one unit, but if it comes up tails then he pays the manager one unit. He plays this game repeatedly until one of two events occurs: either he runs out of money and is bankrupted or he wins enough to buy the Jaguar. What is the probability that he is ultimately bankrupted?

Solution. This is one of many problems the solution to which proceeds by the construction of a linear difference equation subject to certain boundary conditions. Let A denote the event that he is eventually bankrupted, and let B be the event that the first toss of the coin shows heads. By Lemma (1.4.4),

$$\mathbb{P}_k(A) = \mathbb{P}_k(A \mid B)\mathbb{P}(B) + \mathbb{P}_k(A \mid B^c)\mathbb{P}(B^c), \tag{5}$$

where $\mathbb{P}_k$ denotes probabilities calculated relative to the starting point k. We want to find $\mathbb{P}_k(A)$. Consider $\mathbb{P}_k(A \mid B)$. If the first toss is a head then his capital increases to $k + 1$ units and the game starts afresh from a different starting point. Thus $\mathbb{P}_k(A \mid B) = \mathbb{P}_{k+1}(A)$ and similarly $\mathbb{P}_k(A \mid B^c) = \mathbb{P}_{k-1}(A)$. So, writing $p_k = \mathbb{P}_k(A)$, (5) becomes

$$p_k = \tfrac{1}{2}(p_{k+1} + p_{k-1}) \quad \text{if} \quad 0 < k < N, \tag{6}$$

which is a linear difference equation subject to the boundary conditions $p_0 = 1$, $p_N = 0$. The analytical solution to such equations is routine, and we shall return later to the general

method of solution. In this case we can proceed directly. We put $b_k = p_k - p_{k-1}$ to obtain $b_k = b_{k-1}$ and hence $b_k = b_1$ for all k. Thus

$$p_k = b_1 + p_{k-1} = 2b_1 + p_{k-2} = \cdots = kb_1 + p_0$$

is the general solution to (6). The boundary conditions imply that $p_0 = 1, b_1 = -1/N$, giving

$$\mathbb{P}_k(A) = 1 - \frac{k}{N}. \tag{7}$$

As the price of the Jaguar rises, that is as $N \to \infty$, ultimate bankruptcy becomes very likely. This is the problem of the 'symmetric random walk with two absorbing barriers' to which we shall return in more generality later. ●

Remark. Our experience of student calculations leads us to stress that probabilities lie between zero and one; any calculated probability which violates this must be incorrect.

(8) Example. Testimony. A court is investigating the possible occurrence of an unlikely event T. The reliability of two independent witnesses called Alf and Bob is known to the court: Alf tells the truth with probability α and Bob with probability β, and there is no collusion between the two of them. Let A and B be the events that Alf and Bob assert (respectively) that T occurred, and let $\tau = \mathbb{P}(T)$. What is the probability that T occurred given that both Alf and Bob declare that T occurred?

Solution. We are asked to calculate $\mathbb{P}(T \mid A \cap B)$, which is equal to $\mathbb{P}(T \cap A \cap B)/\mathbb{P}(A \cap B)$. Now $\mathbb{P}(T \cap A \cap B) = \mathbb{P}(A \cap B \mid T)\mathbb{P}(T)$ and

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

We have from the independence of the witnesses that A and B are conditionally independent given either T or T^c. Therefore

$$\begin{aligned}\mathbb{P}(A \cap B \mid T) &= \mathbb{P}(A \mid T)\mathbb{P}(B \mid T) = \alpha\beta,\\ \mathbb{P}(A \cap B \mid T^c) &= \mathbb{P}(A \mid T^c)\mathbb{P}(B \mid T^c) = (1-\alpha)(1-\beta),\end{aligned}$$

so that

$$\mathbb{P}(T \mid A \cap B) = \frac{\alpha\beta\tau}{\alpha\beta\tau + (1-\alpha)(1-\beta)(1-\tau)}.$$

As an example, suppose that $\alpha = \beta = \frac{9}{10}$ and $\tau = 1/1000$. Then $\mathbb{P}(T \mid A \cap B) = 81/1080$, which is somewhat small as a basis for a judicial conclusion.

This calculation may be informative. However, it is generally accepted that such an application of the axioms of probability is inappropriate to questions of truth and belief. ●

(9) Example. Zoggles revisited. A new process for the production of zoggles is invented, and both factories of Example (1.4.6) install extra production lines using it. The new process is cheaper but produces fewer reliable zoggles, only 75 per cent of items produced in this new way being reliable.

Factory I fails to implement its new production line efficiently, and only 10 per cent of its output is made in this manner. Factory II does better: it produces 20 per cent of its output by the new technology, and now produces twice as many zoggles in all as Factory I.

Is the new process beneficial to the consumer?

Solution. Both factories now produce a higher proportion of unreliable zoggles than before, and so it might seem at first sight that there is an increased proportion of unreliable zoggles on the market.

Let A be the event that a randomly chosen zoggle is satisfactory, B the event that it came from factory I, and C the event that it was made by the new method. Then

$$\begin{aligned}\mathbb{P}(A) &= \tfrac{1}{3}\mathbb{P}(A \mid B) + \tfrac{2}{3}\mathbb{P}(A \mid B^c) \\ &= \tfrac{1}{3}\left(\tfrac{1}{10}\mathbb{P}(A \mid B \cap C) + \tfrac{9}{10}\mathbb{P}(A \mid B \cap C^c)\right) \\ &\quad + \tfrac{2}{3}\left(\tfrac{1}{5}\mathbb{P}(A \mid B^c \cap C) + \tfrac{4}{5}\mathbb{P}(A \mid B^c \cap C^c)\right) \\ &= \tfrac{1}{3}\left(\tfrac{1}{10}\cdot\tfrac{3}{4} + \tfrac{9}{10}\cdot\tfrac{4}{5}\right) + \tfrac{2}{3}\left(\tfrac{1}{5}\cdot\tfrac{3}{4} + \tfrac{4}{5}\cdot\tfrac{19}{20}\right) = \tfrac{523}{600} > \tfrac{51}{60},\end{aligned}$$

so that the proportion of satisfactory zoggles has been increased. ●

(10) Example. Simpson's paradox†. A doctor has performed clinical trials to determine the relative efficacies of two drugs, with the following results.

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

Which drug is the better? Here are two conflicting responses.

1. Drug I was given to 2020 people, of whom 219 were cured. The success rate was 219/2020, which is much smaller than the corresponding figure, 1010/2200, for drug II. Therefore drug II is better than drug I.
2. Amongst women the success rates of the drugs are 1/10 and 1/20, and amongst men 19/20 and 1/2. Drug I wins in both cases.

This well-known statistical paradox may be reformulated in the following more general way. Given three events A, B, C, it is possible to allocate probabilities such that

(11) $$\mathbb{P}(A \mid B \cap C) > \mathbb{P}(A \mid B^c \cap C) \quad \text{and} \quad \mathbb{P}(A \mid B \cap C^c) > \mathbb{P}(A \mid B^c \cap C^c)$$

but

(12) $$\mathbb{P}(A \mid B) < \mathbb{P}(A \mid B^c).$$

†This paradox, named after Simpson (1951), was remarked by Yule in 1903. The nomenclature is an instance of Stigler's law of eponymy: "No law, theorem, or discovery is named after its originator". This law applies to many eponymous statements in this book, including the law itself. As remarked by A. N. Whitehead, "Everything of importance has been said before, by somebody who did not discover it".

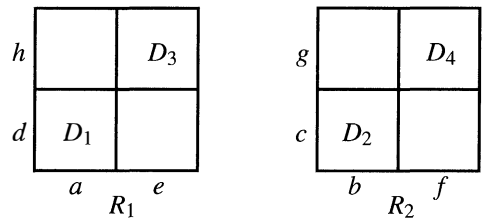

Figure 1.1. Two unions of rectangles illustrating Simpson's paradox.

We may think of A as the event that treatment is successful, B as the event that drug I is given to a randomly chosen individual, and C as the event that this individual is female. The above inequalities imply that B is preferred to B^c when C occurs and when C^c occurs, but B^c is preferred to B overall.

Setting

$$\begin{aligned} a &= \mathbb{P}(A \cap B \cap C), & b &= \mathbb{P}(A^c \cap B \cap C), \\ c &= \mathbb{P}(A \cap B^c \cap C), & d &= \mathbb{P}(A^c \cap B^c \cap C), \\ e &= \mathbb{P}(A \cap B \cap C^c), & f &= \mathbb{P}(A^c \cap B \cap C^c), \\ g &= \mathbb{P}(A \cap B^c \cap C^c), & h &= \mathbb{P}(A^c \cap B^c \cap C^c), \end{aligned}$$

and expanding (11)–(12), we arrive at the (equivalent) inequalities

$$\textbf{(13)} \qquad ad > bc, \quad eh > fg, \quad (a+e)(d+h) < (b+f)(c+g),$$

subject to the conditions $a, b, c, \ldots, h \geq 0$ and $a + b + c + \cdots + h = 1$. Inequalities (13) are equivalent to the existence of two rectangles R_1 and R_2, as in Figure 1.1, satisfying

$$\text{area}\,(D_1) > \text{area}\,(D_2), \quad \text{area}\,(D_3) > \text{area}\,(D_4), \quad \text{area}\,(R_1) < \text{area}\,(R_2).$$

Many such rectangles may be found, by inspection, as for example those with $a = \frac{3}{30}, b = \frac{1}{30}$, $c = \frac{8}{30}, d = \frac{3}{30}, e = \frac{3}{30}, f = \frac{8}{30}, g = \frac{1}{30}, h = \frac{3}{30}$. Similar conclusions are valid for finer partitions $\{C_i : i \in I\}$ of the sample space, though the corresponding pictures are hardèr to draw.

Simpson's paradox has arisen many times in practical situations. There are many well-known cases, including the admission of graduate students to the University of California at Berkeley and a clinical trial comparing treatments for kidney stones. ●

(14) Example. False positives. A rare disease affects one person in 10^5. A test for the disease shows positive with probability $\frac{99}{100}$ when applied to an ill person, and with probability $\frac{1}{100}$ when applied to a healthy person. What is the probability that you have the disease given that the test shows positive?

Solution. In the obvious notation,

$$\begin{aligned} \mathbb{P}(\text{ill} \mid +) &= \frac{\mathbb{P}(+ \mid \text{ill})\mathbb{P}(\text{ill})}{\mathbb{P}(+ \mid \text{ill})\mathbb{P}(\text{ill}) + \mathbb{P}(+ \mid \text{healthy})\mathbb{P}(\text{healthy})} \\ &= \frac{\frac{99}{100} \cdot 10^{-5}}{\frac{99}{100} \cdot 10^{-5} + \frac{1}{100}(1 - 10^{-5})} = \frac{99}{99 + 10^5 - 1} \simeq \frac{1}{1011}. \end{aligned}$$

The chance of being ill is rather small. Indeed it is more likely that the test was incorrect. ●

Exercises for Section 1.7

1. There are two roads from A to B and two roads from B to C. Each of the four roads is blocked by snow with probability p, independently of the others. Find the probability that there is an open road from A to B given that there is no open route from A to C.

If, in addition, there is a direct road from A to C, this road being blocked with probability p independently of the others, find the required conditional probability.

2. Calculate the probability that a hand of 13 cards dealt from a normal shuffled pack of 52 contains exactly two kings and one ace. What is the probability that it contains exactly one ace given that it contains exactly two kings?

3. A symmetric random walk takes place on the integers $0, 1, 2, \ldots, N$ with absorbing barriers at 0 and N, starting at k. Show that the probability that the walk is never absorbed is zero.

4. The so-called 'sure thing principle' asserts that if you prefer x to y given C, and also prefer x to y given C^c, then you surely prefer x to y. Agreed?

5. A pack contains m cards, labelled $1, 2, \ldots, m$. The cards are dealt out in a random order, one by one. Given that the label of the kth card dealt is the largest of the first k cards dealt, what is the probability that it is also the largest in the pack?

1.8 Problems

1. A traditional fair die is thrown twice. What is the probability that:
(a) a six turns up exactly once?
(b) both numbers are odd?
(c) the sum of the scores is 4?
(d) the sum of the scores is divisible by 3?

2. A fair coin is thrown repeatedly. What is the probability that on the nth throw:
(a) a head appears for the first time?
(b) the numbers of heads and tails to date are equal?
(c) exactly two heads have appeared altogether to date?
(d) at least two heads have appeared to date?

3. Let $\mathcal{F}$ and $\mathcal{G}$ be σ-fields of subsets of Ω.
(a) Use elementary set operations to show that $\mathcal{F}$ is closed under countable intersections; that is, if $A_1, A_2, \ldots$ are in $\mathcal{F}$, then so is $\bigcap_i A_i$.
(b) Let $\mathcal{H} = \mathcal{F} \cap \mathcal{G}$ be the collection of subsets of Ω lying in both $\mathcal{F}$ and $\mathcal{G}$. Show that $\mathcal{H}$ is a σ-field.
(c) Show that $\mathcal{F} \cup \mathcal{G}$, the collection of subsets of Ω lying in either $\mathcal{F}$ or $\mathcal{G}$, is not necessarily a σ-field.

4. Describe the underlying probability spaces for the following experiments:
(a) a biased coin is tossed three times;
(b) two balls are drawn without replacement from an urn which originally contained two ultramarine and two vermilion balls;
(c) a biased coin is tossed repeatedly until a head turns up.

5. Show that the probability that *exactly* one of the events A and B occurs is

$$\mathbb{P}(A) + \mathbb{P}(B) - 2\mathbb{P}(A \cap B).$$

6. Prove that $\mathbb{P}(A \cup B \cup C) = 1 - \mathbb{P}(A^c \mid B^c \cap C^c)\mathbb{P}(B^c \mid C^c)\mathbb{P}(C^c)$.

7. (a) If A is independent of itself, show that $\mathbb{P}(A)$ is 0 or 1.
(b) If $\mathbb{P}(A)$ is 0 or 1, show that A is independent of all events B.

8. Let $\mathcal{F}$ be a σ-field of subsets of Ω, and suppose $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfies: (i) $\mathbb{P}(\Omega) = 1$, and (ii) $\mathbb{P}$ is additive, in that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ whenever $A \cap B = \varnothing$. Show that $\mathbb{P}(\varnothing) = 0$.

9. Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $B \in \mathcal{F}$ satisfies $\mathbb{P}(B) > 0$. Let $\mathbb{Q} : \mathcal{F} \to [0, 1]$ be defined by $\mathbb{Q}(A) = \mathbb{P}(A \mid B)$. Show that $(\Omega, \mathcal{F}, \mathbb{Q})$ is a probability space. If $C \in \mathcal{F}$ and $\mathbb{Q}(C) > 0$, show that $\mathbb{Q}(A \mid C) = \mathbb{P}(A \mid B \cap C)$; discuss.

10. Let $B_1, B_2, \ldots$ be a partition of the sample space Ω, each B_i having positive probability, and show that

$$\mathbb{P}(A) = \sum_{j=1}^{\infty} \mathbb{P}(A \mid B_j)\mathbb{P}(B_j).$$

11. Prove **Boole's inequalities**:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i), \qquad \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} \mathbb{P}(A_i^c).$$

12. Prove that

$$\mathbb{P}\left(\bigcap_{1}^{n} A_i\right) = \sum_{i} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cup A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cup A_j \cup A_k) - \cdots - (-1)^n \mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n).$$

13. Let $A_1, A_2, \ldots, A_n$ be events, and let N_k be the event that exactly k of the A_i occur. Prove the result sometimes referred to as **Waring's theorem**:

$$\mathbb{P}(N_k) = \sum_{i=0}^{n-k} (-1)^i \binom{k+i}{k} S_{k+i}, \text{ where } S_j = \sum_{i_1 < i_2 < \cdots < i_j} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_j}).$$

Use this result to find an expression for the probability that a purchase of six packets of Corn Flakes yields exactly three distinct busts (see Exercise (1.3.4)).

14. Prove **Bayes's formula**: if $A_1, A_2, \ldots, A_n$ is a partition of Ω, each A_i having positive probability, then

$$\mathbb{P}(A_j \mid B) = \frac{\mathbb{P}(B \mid A_j)\mathbb{P}(A_j)}{\sum_1^n \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}.$$

15. A random number N of dice is thrown. Let A_i be the event that $N = i$, and assume that $\mathbb{P}(A_i) = 2^{-i}$, $i \geq 1$. The sum of the scores is S. Find the probability that:
(a) $N = 2$ given $S = 4$;
(b) $S = 4$ given N is even;
(c) $N = 2$, given that $S = 4$ and the first die showed 1;
(d) the largest number shown by any die is r, where S is unknown.

16. Let $A_1, A_2, \ldots$ be a sequence of events. Define

$$B_n = \bigcup_{m=n}^{\infty} A_m, \quad C_n = \bigcap_{m=n}^{\infty} A_m.$$

Clearly $C_n \subseteq A_n \subseteq B_n$. The sequences $\{B_n\}$ and $\{C_n\}$ are decreasing and increasing respectively with limits

$$\lim B_n = B = \bigcap_n B_n = \bigcap_n \bigcup_{m \geq n} A_m, \qquad \lim C_n = C = \bigcup_n C_n = \bigcup_n \bigcap_{m \geq n} A_m.$$

The events B and C are denoted $\limsup_{n\to\infty} A_n$ and $\liminf_{n\to\infty} A_n$ respectively. Show that

(a) $B = \{\omega \in \Omega : \omega \in A_n$ for infinitely many values of $n\}$,

(b) $C = \{\omega \in \Omega : \omega \in A_n$ for all but finitely many values of $n\}$.

We say that the sequence $\{A_n\}$ converges to a limit $A = \lim A_n$ if B and C are the same set A. Suppose that $A_n \to A$ and show that

(c) A is an event, in that $A \in \mathcal{F}$,

(d) $\mathbb{P}(A_n) \to \mathbb{P}(A)$.

17. In Problem (1.8.16) above, show that B and C are independent whenever B_n and C_n are independent for all n. Deduce that if this holds and furthermore $A_n \to A$, then $\mathbb{P}(A)$ equals either zero or one.

18. Show that the assumption that $\mathbb{P}$ is *countably* additive is equivalent to the assumption that $\mathbb{P}$ is continuous. That is to say, show that if a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfies $\mathbb{P}(\varnothing) = 0$, $\mathbb{P}(\Omega) = 1$, and $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ whenever $A, B \in \mathcal{F}$ and $A \cap B = \varnothing$, then $\mathbb{P}$ is countably additive (in the sense of satisfying Definition (1.3.1b)) if and only if $\mathbb{P}$ is continuous (in the sense of Lemma (1.3.5)).

19. Anne, Betty, Chloë, and Daisy were all friends at school. Subsequently each of the $\binom{4}{2} = 6$ subpairs meet up; at each of the six meetings the pair involved quarrel with some fixed probability p, or become firm friends with probability $1 - p$. Quarrels take place independently of each other. In future, if any of the four hears a rumour, then she tells it to her firm friends only. If Anne hears a rumour, what is the probability that:

(a) Daisy hears it?

(b) Daisy hears it if Anne and Betty have quarrelled?

(c) Daisy hears it if Betty and Chloë have quarrelled?

(d) Daisy hears it if she has quarrelled with Anne?

20. A biased coin is tossed repeatedly. Each time there is a probability p of a head turning up. Let p_n be the probability that an even number of heads has occurred after n tosses (zero is an even number). Show that $p_0 = 1$ and that $p_n = p(1 - p_{n-1}) + (1 - p)p_{n-1}$ if $n \geq 1$. Solve this difference equation.

21. A biased coin is tossed repeatedly. Find the probability that there is a run of r heads in a row before there is a run of s tails, where r and s are positive integers.

22. A bowl contains twenty cherries, exactly fifteen of which have had their stones removed. A greedy pig eats five whole cherries, picked at random, without remarking on the presence or absence of stones. Subsequently, a cherry is picked randomly from the remaining fifteen.

(a) What is the probability that this cherry contains a stone?

(b) Given that this cherry contains a stone, what is the probability that the pig consumed at least one stone?

23. The '**ménages**' problem poses the following question. Some consider it to be desirable that men and women alternate when seated at a circular table. If n couples are seated randomly according to this rule, show that the probability that nobody sits next to his or her partner is

$$\frac{1}{n!} \sum_{k=0}^{n} (-1)^k \frac{2n}{2n-k} \binom{2n-k}{k} (n-k)!$$

You may find it useful to show first that the number of ways of selecting k non-overlapping pairs of adjacent seats is $\binom{2n-k}{k} 2n(2n-k)^{-1}$.

24. An urn contains b blue balls and r red balls. They are removed at random and not replaced. Show that the probability that the first red ball drawn is the $(k+1)$th ball drawn equals $\binom{r+b-k-1}{r-1} \Big/ \binom{r+b}{b}$. Find the probability that the last ball drawn is red.

25. An urn contains a azure balls and c carmine balls, where $ac \neq 0$. Balls are removed at random and discarded until the first time that a ball (B, say) is removed having a different colour from its predecessor. The ball B is now replaced and the procedure restarted. This process continues until the last ball is drawn from the urn. Show that this last ball is equally likely to be azure or carmine.

26. Protocols. A pack of four cards contains one spade, one club, and the two red aces. You deal two cards faces downwards at random in front of a truthful friend. She inspects them and tells you that one of them is the ace of hearts. What is the chance that the other card is the ace of diamonds? Perhaps $\frac{1}{3}$?

Suppose that your friend's protocol was:

(a) with no red ace, say "no red ace",

(b) with the ace of hearts, say "ace of hearts",

(c) with the ace of diamonds but not the ace of hearts, say "ace of diamonds".

Show that the probability in question is $\frac{1}{3}$.

Devise a possible protocol for your friend such that the probability in question is zero.

27. Eddington's controversy. Four witnesses, A, B, C, and D, at a trial each speak the truth with probability $\frac{1}{3}$ independently of each other. In their testimonies, A claimed that B denied that C declared that D lied. What is the (conditional) probability that D told the truth? [This problem seems to have appeared first as a parody in a university magazine of the 'typical' Cambridge Philosophy Tripos question.]

28. The probabilistic method. 10 per cent of the surface of a sphere is coloured blue, the rest is red. Show that, irrespective of the manner in which the colours are distributed, it is possible to inscribe a cube in S with all its vertices red.

29. Repulsion. The event A is said to be repelled by the event B if $\mathbb{P}(A \mid B) < \mathbb{P}(A)$, and to be attracted by B if $\mathbb{P}(A \mid B) > \mathbb{P}(A)$. Show that if B attracts A, then A attracts B, and B^c repels A.

If A attracts B, and B attracts C, does A attract C?

30. Birthdays. If m students born on independent days in 1991 are attending a lecture, show that the probability that at least two of them share a birthday is $p = 1 - (365)!/\{(365-m)!\,365^m\}$. Show that $p > \frac{1}{2}$ when $m = 23$.

31. Lottery. You choose r of the first n positive integers, and a lottery chooses a random subset L of the same size. What is the probability that:

(a) L includes no consecutive integers?

(b) L includes exactly one pair of consecutive integers?

(c) the numbers in L are drawn in increasing order?

(d) your choice of numbers is the same as L?

(e) there are exactly k of your numbers matching members of L?

32. Bridge. During a game of bridge, you are dealt at random a hand of thirteen cards. With an obvious notation, show that $\mathbb{P}(4S, 3H, 3D, 3C) \simeq 0.026$ and $\mathbb{P}(4S, 4H, 3D, 2C) \simeq 0.018$. However if suits are not specified, so numbers denote the shape of your hand, show that $\mathbb{P}(4, 3, 3, 3) \simeq 0.11$ and $\mathbb{P}(4, 4, 3, 2) \simeq 0.22$.

33. Poker. During a game of poker, you are dealt a five-card hand at random. With the convention that aces may count high or low, show that:

$$\mathbb{P}(\text{1 pair}) \simeq 0.423, \quad \mathbb{P}(\text{2 pairs}) \simeq 0.0475, \quad \mathbb{P}(\text{3 of a kind}) \simeq 0.021,$$
$$\mathbb{P}(\text{straight}) \simeq 0.0039, \quad \mathbb{P}(\text{flush}) \simeq 0.0020, \quad \mathbb{P}(\text{full house}) \simeq 0.0014,$$
$$\mathbb{P}(\text{4 of a kind}) \simeq 0.00024, \quad \mathbb{P}(\text{straight flush}) \simeq 0.000015.$$

34. Poker dice. There are five dice each displaying 9, 10, J, Q, K, A. Show that, when rolled:

$$\begin{array}{lll} \mathbb{P}(\text{1 pair}) \simeq 0.46, & \mathbb{P}(\text{2 pairs}) \simeq 0.23, & \mathbb{P}(\text{3 of a kind}) \simeq 0.15, \\ \mathbb{P}(\text{no 2 alike}) \simeq 0.093, & \mathbb{P}(\text{full house}) \simeq 0.039, & \mathbb{P}(\text{4 of a kind}) \simeq 0.019, \\ \mathbb{P}(\text{5 of a kind}) \simeq 0.0008. & & \end{array}$$

35. You are lost in the National Park of **Bandrika**†. Tourists comprise two-thirds of the visitors to the park, and give a correct answer to requests for directions with probability $\frac{3}{4}$. (Answers to repeated questions are independent, even if the question and the person are the same.) If you ask a Bandrikan for directions, the answer is always false.

(a) You ask a passer-by whether the exit from the Park is East or West. The answer is East. What is the probability this is correct?

(b) You ask the same person again, and receive the same reply. Show the probability that it is correct is $\frac{1}{2}$.

(c) You ask the same person again, and receive the same reply. What is the probability that it is correct?

(d) You ask for the fourth time, and receive the answer East. Show that the probability it is correct is $\frac{27}{70}$.

(e) Show that, had the fourth answer been West instead, the probability that East is nevertheless correct is $\frac{9}{10}$.

36. Mr Bayes goes to Bandrika. Tom is in the same position as you were in the previous problem, but he has reason to believe that, with probability ϵ, East is the correct answer. Show that:

(a) whatever answer first received, Tom continues to believe that East is correct with probability ϵ,

(b) if the first two replies are the same (that is, either WW or EE), Tom continues to believe that East is correct with probability ϵ,

(c) after three like answers, Tom will calculate as follows, in the obvious notation:

$$\mathbb{P}(\text{East correct} \mid \text{EEE}) = \frac{9\epsilon}{11 - 2\epsilon}, \quad \mathbb{P}(\text{East correct} \mid \text{WWW}) = \frac{11\epsilon}{9 + 2\epsilon}.$$

Evaluate these when $\epsilon = \frac{9}{20}$.

37. Bonferroni's inequality. Show that

$$\mathbb{P}\Big(\bigcup_{r=1}^{n} A_r\Big) \geq \sum_{r=1}^{n} \mathbb{P}(A_r) - \sum_{r<k} \mathbb{P}(A_r \cap A_k).$$

38. Kounias's inequality. Show that

$$\mathbb{P}\Big(\bigcup_{r=1}^{n} A_r\Big) \leq \min_k \left\{ \sum_{r=1}^{n} \mathbb{P}(A_r) - \sum_{r:r\neq k} \mathbb{P}(A_r \cap A_k) \right\}.$$

39. The n passengers for a Bell-Air flight in an airplane with n seats have been told their seat numbers. They get on the plane one by one. The first person sits in the wrong seat. Subsequent passengers sit in their assigned seats whenever they find them available, or otherwise in a randomly chosen empty seat. What is the probability that the last passenger finds his seat free?

---

†A fictional country made famous in the Hitchcock film 'The Lady Vanishes'.

2

Random variables and their distributions

Summary. Quantities governed by randomness correspond to functions on the probability space called random variables. The value taken by a random variable is subject to chance, and the associated likelihoods are described by a function called the distribution function. Two important classes of random variables are discussed, namely discrete variables and continuous variables. The law of averages, known also as the law of large numbers, states that the proportion of successes in a long run of independent trials converges to the probability of success in any one trial. This result provides a mathematical basis for a philosophical view of probability based on repeated experimentation. Worked examples involving random variables and their distributions are included, and the chapter terminates with sections on random vectors and on Monte Carlo simulation.

2.1 Random variables

We shall not always be interested in an experiment itself, but rather in some consequence of its random outcome. For example, many gamblers are more concerned with their losses than with the games which give rise to them. Such consequences, when real valued, may be thought of as functions which map Ω into the real line $\mathbb{R}$, and these functions are called 'random† variables'.

(1) Example. A fair coin is tossed twice: $\Omega = \{\text{HH, HT, TH, TT}\}$. For $\omega \in \Omega$, let $X(\omega)$ be the number of heads, so that

$$X(\text{HH}) = 2, \quad X(\text{HT}) = X(\text{TH}) = 1, \quad X(\text{TT}) = 0.$$

Now suppose that a gambler wagers his fortune of £1 on the result of this experiment. He gambles cumulatively so that his fortune is doubled each time a head appears, and is annihilated on the appearance of a tail. His subsequent fortune W is a random variable given by

$$W(\text{HH}) = 4, \quad W(\text{HT}) = W(\text{TH}) = W(\text{TT}) = 0. \qquad \bullet$$

†Derived from the Old French word *randon* meaning 'haste'.

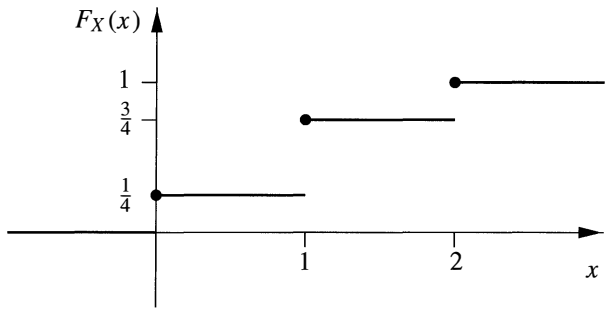

Figure 2.1. The distribution function F_X of the random variable X of Examples (1) and (5).

After the experiment is done and the outcome $\omega \in \Omega$ is known, a random variable $X : \Omega \to \mathbb{R}$ takes some value. In general this numerical value is more likely to lie in certain subsets of $\mathbb{R}$ than in certain others, depending on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the function X itself. We wish to be able to describe the distribution of the likelihoods of possible values of X. Example (1) above suggests that we might do this through the function $f : \mathbb{R} \to [0, 1]$ defined by

$$f(x) = \text{probability that } X \text{ is equal to } x,$$

but this turns out to be inappropriate in general. Rather, we use the *distribution function* $F : \mathbb{R} \to \mathbb{R}$ defined by

$$F(x) = \text{probability that } X \text{ does not exceed } x.$$

More rigorously, this is

(2) $$F(x) = \mathbb{P}(A(x))$$

where $A(x) \subseteq \Omega$ is given by $A(x) = \{\omega \in \Omega : X(\omega) \leq x\}$. However, $\mathbb{P}$ is a function on the collection $\mathcal{F}$ of events; we cannot discuss $\mathbb{P}(A(x))$ unless $A(x)$ belongs to $\mathcal{F}$, and so we are led to the following definition.

(3) Definition. A random variable is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$. Such a function is said to be **$\mathcal{F}$-measurable**.

If you so desire, you may pay no attention to the technical condition in the definition and think of random variables simply as functions mapping Ω into $\mathbb{R}$. We shall always use upper-case letters, such as X, Y, and Z, to represent generic random variables, whilst lower-case letters, such as x, y, and z, will be used to represent possible numerical values of these variables. Do not confuse this notation in your written work.

Every random variable has a distribution function, given by (2); distribution functions are *very* important and useful.

(4) Definition. The **distribution function** of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ given by $F(x) = \mathbb{P}(X \leq x)$.

This is the obvious abbreviation of equation (2). Events written as $\{\omega \in \Omega : X(\omega) \leq x\}$ are commonly abbreviated to $\{\omega : X(\omega) \leq x\}$ or $\{X \leq x\}$. We write F_X where it is necessary to emphasize the role of X.

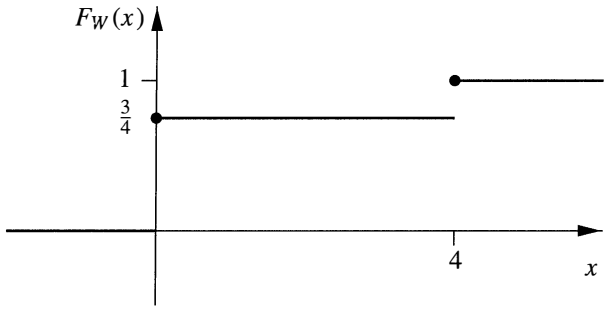

Figure 2.2. The distribution function F_W of the random variable W of Examples (1) and (5).

(5) Example (1) revisited. The distribution function F_X of X is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{4} & \text{if } 0 \le x < 1, \\ \frac{3}{4} & \text{if } 1 \le x < 2, \\ 1 & \text{if } x \ge 2, \end{cases}$$

and is sketched in Figure 2.1. The distribution function F_W of W is given by

$$F_W(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{3}{4} & \text{if } 0 \le x < 4, \\ 1 & \text{if } x \ge 4, \end{cases}$$

and is sketched in Figure 2.2. This illustrates the important point that the distribution function of a random variable X tells us about the values taken by X and their relative likelihoods, rather than about the sample space and the collection of events. ●

(6) Lemma. *A distribution function F has the following properties*:

(a) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$,

(b) *if $x < y$ then $F(x) \le F(y)$,*

(c) *F is right-continuous, that is, $F(x + h) \to F(x)$ as $h \downarrow 0$.*

Proof.

(a) Let $B_n = \{\omega \in \Omega : X(\omega) \le -n\} = \{X \le -n\}$. The sequence $B_1, B_2, \ldots$ is decreasing with the empty set as limit. Thus, by Lemma (1.3.5), $\mathbb{P}(B_n) \to \mathbb{P}(\varnothing) = 0$. The other part is similar.

(b) Let $A(x) = \{X \le x\}$, $A(x, y) = \{x < X \le y\}$. Then $A(y) = A(x) \cup A(x, y)$ is a disjoint union, and so by Definition (1.3.1),

$$\mathbb{P}(A(y)) = \mathbb{P}(A(x)) + \mathbb{P}(A(x, y))$$

giving

$$F(y) = F(x) + \mathbb{P}(x < X \le y) \ge F(x).$$

(c) This is an *exercise*. Use Lemma (1.3.5). ■

Actually, this lemma characterizes distribution functions. That is to say, F is the distribution function of some random variable if and only if it satisfies (6a), (6b), and (6c).

For the time being we can forget all about probability spaces and concentrate on random variables and their distribution functions. The distribution function F of X contains a great deal of information about X.

(7) Example. Constant variables. The simplest random variable takes a constant value on the whole domain Ω. Let $c \in \mathbb{R}$ and define $X : \Omega \to \mathbb{R}$ by

$$X(\omega) = c \quad \text{for all} \quad \omega \in \Omega.$$

The distribution function $F(x) = \mathbb{P}(X \leq x)$ is the step function

$$F(x) = \begin{cases} 0 & x < c, \\ 1 & x \geq c. \end{cases}$$

Slightly more generally, we call X *constant* (*almost surely*) if there exists $c \in \mathbb{R}$ such that $\mathbb{P}(X = c) = 1$. ●

(8) Example. Bernoulli variables. Consider Example (1.3.2). Let $X : \Omega \to \mathbb{R}$ be given by

$$X(H) = 1, \quad X(T) = 0.$$

Then X is the simplest non-trivial random variable, having two possible values, 0 and 1. Its distribution function $F(x) = \mathbb{P}(X \leq x)$ is

$$F(x) = \begin{cases} 0 & x < 0, \\ 1 - p & 0 \leq x < 1, \\ 1 & x \geq 1. \end{cases}$$

X is said to have the *Bernoulli distribution* sometimes denoted Bern(p). ●

(9) Example. Indicator functions. A particular class of Bernoulli variables is very useful in probability theory. Let A be an event and let $I_A : \Omega \to \mathbb{R}$ be the *indicator function* of A; that is,

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

Then I_A is a Bernoulli random variable taking the values 1 and 0 with probabilities $\mathbb{P}(A)$ and $\mathbb{P}(A^c)$ respectively. Suppose $\{B_i : i \in I\}$ is a family of disjoint events with $A \subseteq \bigcup_{i \in I} B_i$. Then

$$\textbf{(10)} \qquad I_A = \sum_i I_{A \cap B_i},$$

an identity which is often useful. ●

(11) Lemma. *Let F be the distribution function of X. Then*

(a) $\mathbb{P}(X > x) = 1 - F(x)$,

(b) $\mathbb{P}(x < X \leq y) = F(y) - F(x)$,

(c) $\mathbb{P}(X = x) = F(x) - \lim_{y \uparrow x} F(y)$.

Proof. (a) and (b) are *exercises*.

(c) Let $B_n = \{x - 1/n < X \leq x\}$ and use the method of proof of Lemma (6). ■

Note one final piece of jargon for future use. A random variable X with distribution function F is said to have two 'tails' given by

$$T_1(x) = \mathbb{P}(X > x) = 1 - F(x), \quad T_2(x) = \mathbb{P}(X \leq x) = F(-x),$$

where x is large and positive. We shall see later that the rates at which the T_i decay to zero as $x \to \infty$ have a substantial effect on the existence or non-existence of certain associated quantities called the 'moments' of the distribution.

Exercises for Section 2.1

1. Let X be a random variable on a given probability space, and let $a \in \mathbb{R}$. Show that

(i) aX is a random variable,

(ii) $X - X = 0$, the random variable taking the value 0 always, and $X + X = 2X$.

2. A random variable X has distribution function F. What is the distribution function of $Y = aX + b$, where a and b are real constants?

3. A fair coin is tossed n times. Show that, under reasonable assumptions, the probability of exactly k heads is $\binom{n}{k}(\frac{1}{2})^n$. What is the corresponding quantity when heads appears with probability p on each toss?

4. Show that if F and G are distribution functions and $0 \leq \lambda \leq 1$ then $\lambda F + (1-\lambda)G$ is a distribution function. Is the product FG a distribution function?

5. Let F be a distribution function and r a positive integer. Show that the following are distribution functions:

(a) $F(x)^r$,

(b) $1 - \{1 - F(x)\}^r$,

(c) $F(x) + \{1 - F(x)\}\log\{1 - F(x)\}$,

(d) $\{F(x) - 1\}e + \exp\{1 - F(x)\}$.

2.2 The law of averages

We may recall the discussion in Section 1.3 of repeated experimentation. In each of N repetitions of an experiment, we observe whether or not a given event A occurs, and we write $N(A)$ for the total number of occurrences of A. One possible philosophical underpinning of probability theory requires that the proportion $N(A)/N$ settles down as $N \to \infty$ to some limit interpretable as the 'probability of A'. Is our theory to date consistent with such a requirement?

With this question in mind, let us suppose that $A_1, A_2, \ldots$ is a sequence of independent events having equal probability $\mathbb{P}(A_i) = p$, where $0 < p < 1$; such an assumption requires of

course the existence of a corresponding probability space $(\Omega, \mathcal{F}, \mathbb{P})$, but we do not plan to get bogged down in such matters here. We think of A_i as being the event 'that A occurs on the ith experiment'. We write $S_n = \sum_{i=1}^{n} I_{A_i}$, the sum of the indicator functions of $A_1, A_2, \ldots, A_n$; S_n is a random variable which counts the number of occurrences of A_i for $1 \leq i \leq n$ (certainly S_n is a function of Ω, since it is the sum of such functions, and it is left as an *exercise* to show that S_n is $\mathcal{F}$-measurable). The following result concerning the ratio $n^{-1}S_n$ was proved by James Bernoulli before 1692.

(1) Theorem. *It is the case that $n^{-1}S_n$ converges to p as $n \to \infty$ in the sense that, for all $\epsilon > 0$,*

$$\mathbb{P}\big(p - \epsilon \leq n^{-1}S_n \leq p + \epsilon\big) \to 1 \quad \text{as} \quad n \to \infty.$$

There are certain technicalities involved in the study of the convergence of random variables (see Chapter 7), and this is the reason for the careful statement of the theorem. For the time being, we encourage the reader to interpret the theorem as asserting simply that the proportion $n^{-1}S_n$ of times that the events $A_1, A_2, \ldots, A_n$ occur converges as $n \to \infty$ to their common probability p. We shall see later how important it is to be careful when making such statements.

Interpreted in terms of tosses of a fair coin, the theorem implies that the proportion of heads is (with large probability) near to $\frac{1}{2}$. As a caveat regarding the difficulties inherent in studying the convergence of random variables, we remark that it is *not* true that, in a 'typical' sequence of tosses of a fair coin, heads outnumber tails about one-half of the time.

Proof. Suppose that we toss a coin repeatedly, and heads occurs on each toss with probability p. The random variable S_n has the same probability distribution as the number H_n of heads which occur during the first n tosses, which is to say that $\mathbb{P}(S_n = k) = \mathbb{P}(H_n = k)$ for all k. It follows that, for small positive values of ϵ,

$$\mathbb{P}\left(\frac{1}{n}S_n \geq p + \epsilon\right) = \sum_{k \geq n(p+\epsilon)} \mathbb{P}(H_n = k).$$

We have from Exercise (2.1.3) that

$$\mathbb{P}(H_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for} \quad 0 \leq k \leq n,$$

and hence

$$\mathbf{(2)} \qquad \mathbb{P}\left(\frac{1}{n}S_n \geq p + \epsilon\right) = \sum_{k=m}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

where $m = \lceil n(p+\epsilon) \rceil$, the least integer not less than $n(p+\epsilon)$. The following argument is standard in probability theory. Let $\lambda > 0$ and note that $e^{\lambda k} \geq e^{\lambda n(p+\epsilon)}$ if $k \geq m$. Writing $q = 1 - p$, we have that

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}S_n \geq p + \epsilon\right) &\leq \sum_{k=m}^{n} e^{\lambda[k - n(p+\epsilon)]} \binom{n}{k} p^k q^{n-k} \\
&\leq e^{-\lambda n\epsilon} \sum_{k=0}^{n} \binom{n}{k} (pe^{\lambda q})^k (qe^{-\lambda p})^{n-k} \\
&= e^{-\lambda n\epsilon} (pe^{\lambda q} + qe^{-\lambda p})^n,
\end{aligned}$$

by the binomial theorem. It is a simple *exercise* to show that $e^x \le x + e^{x^2}$ for $x \in \mathbb{R}$. With the aid of this inequality, we obtain

$$\mathbf{(3)} \qquad \mathbb{P}\left(\frac{1}{n}S_n \ge p + \epsilon\right) \le e^{-\lambda n\epsilon}\left[pe^{\lambda^2 q^2} + qe^{\lambda^2 p^2}\right]^n \le e^{\lambda^2 n - \lambda n\epsilon}.$$

We can pick λ to minimize the right-hand side, namely $\lambda = \frac{1}{2}\epsilon$, giving

$$\mathbf{(4)} \qquad \mathbb{P}\left(\frac{1}{n}S_n \ge p + \epsilon\right) \le e^{-\frac{1}{4}n\epsilon^2} \quad \text{for} \quad \epsilon > 0,$$

an inequality that is known as 'Bernstein's inequality'. It follows immediately that $\mathbb{P}(n^{-1}S_n \ge p + \epsilon) \to 0$ as $n \to \infty$. An exactly analogous argument shows that $\mathbb{P}(n^{-1}S_n \le p - \epsilon) \to 0$ as $n \to \infty$, and thus the theorem is proved. ■

Bernstein's inequality (4) is rather powerful, asserting that the chance that S_n exceeds its mean by a quantity of order n tends to zero *exponentially fast* as $n \to \infty$; such an inequality is known as a 'large-deviation estimate'. We may use the inequality to prove rather more than the conclusion of the theorem. Instead of estimating the chance that, for a specific value of n, S_n lies between $n(p - \epsilon)$ and $n(p + \epsilon)$, let us estimate the chance that this occurs *for all large n*. Writing $A_n = \{p - \epsilon \le n^{-1}S_n \le p + \epsilon\}$, we wish to estimate $\mathbb{P}(\bigcap_{n=m}^{\infty} A_n)$. Now the complement of this intersection is the event $\bigcup_{n=m}^{\infty} A_n^c$, and the probability of this union satisfies, by the inequalities of Boole and Bernstein,

$$\mathbf{(5)} \qquad \mathbb{P}\left(\bigcup_{n=m}^{\infty} A_n^c\right) \le \sum_{n=m}^{\infty} \mathbb{P}(A_n^c) \le \sum_{n=m}^{\infty} 2e^{-\frac{1}{4}n\epsilon^2} \to 0 \quad \text{as} \quad m \to \infty,$$

giving that, as required,

$$\mathbf{(6)} \qquad \mathbb{P}\left(p - \epsilon \le \frac{1}{n}S_n \le p + \epsilon \text{ for all } n \ge m\right) \to 1 \quad \text{as} \quad m \to \infty.$$

Exercises for Section 2.2

1. You wish to ask each of a large number of people a question to which the answer "yes" is embarrassing. The following procedure is proposed in order to determine the embarrassed fraction of the population. As the question is asked, a coin is tossed out of sight of the questioner. If the answer would have been "no" and the coin shows heads, then the answer "yes" is given. Otherwise people respond truthfully. What do you think of this procedure?

2. A coin is tossed repeatedly and heads turns up on each toss with probability p. Let H_n and T_n be the numbers of heads and tails in n tosses. Show that, for $\epsilon > 0$,

$$\mathbb{P}\left(2p - 1 - \epsilon \le \frac{1}{n}(H_n - T_n) \le 2p - 1 + \epsilon\right) \to 1 \qquad \text{as } n \to \infty.$$

3. Let $\{X_r : r \ge 1\}$ be observations which are independent and identically distributed with unknown distribution function F. Describe and justify a method for estimating $F(x)$.

2.3 Discrete and continuous variables

Much of the study of random variables is devoted to distribution functions, characterized by Lemma (2.1.6). The general theory of distribution functions and their applications is quite difficult and abstract and is best omitted at this stage. It relies on a rigorous treatment of the construction of the Lebesgue–Stieltjes integral; this is sketched in Section 5.6. However, things become much easier if we are prepared to restrict our attention to certain subclasses of random variables specified by properties which make them tractable. We shall consider in depth the collection of 'discrete' random variables and the collection of 'continuous' random variables.

(1) Definition. The random variable X is called **discrete** if it takes values in some countable subset $\{x_1, x_2, \dots\}$, only, of $\mathbb{R}$. The discrete random variable X has **(probability) mass function** $f : \mathbb{R} \to [0, 1]$ given by $f(x) = \mathbb{P}(X = x)$.

We shall see that the distribution function of a discrete variable has jump discontinuities at the values $x_1, x_2, \dots$ and is constant in between; such a distribution is called *atomic*. this contrasts sharply with the other important class of distribution functions considered here.

(2) Definition. The random variable X is called **continuous** if its distribution function can be expressed as

$$F(x) = \int_{-\infty}^{x} f(u)\, du \qquad x \in \mathbb{R},$$

for some integrable function $f : \mathbb{R} \to [0, \infty)$ called the **(probability) density function** of X.

The distribution function of a continuous random variable is certainly continuous (actually it is 'absolutely continuous'). For the moment we are concerned only with discrete variables and continuous variables. There is another sort of random variable, called 'singular', for a discussion of which the reader should look elsewhere. A common example of this phenomenon is based upon the Cantor ternary set (see Grimmett and Welsh 1986, or Billingsley 1995). Other variables are 'mixtures' of discrete, continuous, and singular variables. Note that the word 'continuous' is a misnomer when used in this regard: in describing X as continuous, we are referring to a property of its distribution function rather than of the random variable (function) X itself.

(3) Example. Discrete variables. The variables X and W of Example (2.1.1) take values in the sets $\{0, 1, 2\}$ and $\{0, 4\}$ respectively; they are both discrete. ●

(4) Example. Continuous variables. A straight rod is flung down at random onto a horizontal plane and the angle ω between the rod and true north is measured. The result is a number in $\Omega = [0, 2\pi)$. Never mind about $\mathcal{F}$ for the moment; we can suppose that $\mathcal{F}$ contains all nice subsets of Ω, including the collection of open subintervals such as (a, b), where $0 \le a < b < 2\pi$. The implicit symmetry suggests the probability measure $\mathbb{P}$ which satisfies $\mathbb{P}((a, b)) = (b - a)/(2\pi)$; that is to say, the probability that the angle lies in some interval is directly proportional to the length of the interval. Here are two random variables X and Y:

$$X(\omega) = \omega, \quad Y(\omega) = \omega^2.$$

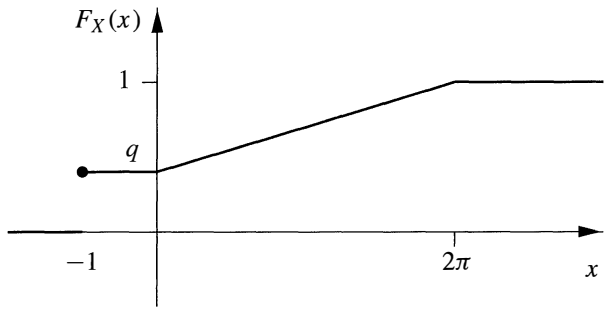

Figure 2.3. The distribution function F_X of the random variable X in Example (5).

Notice that Y is a function of X in that $Y = X^2$. The distribution functions of X and Y are

$$F_X(x) = \begin{cases} 0 & x \le 0, \\ x/(2\pi) & 0 \le x < 2\pi, \\ 1 & x \ge 2\pi, \end{cases} \qquad F_Y(y) = \begin{cases} 0 & y \le 0, \\ \sqrt{y}/(2\pi) & 0 \le y < 4\pi^2, \\ 1 & y \ge 4\pi^2. \end{cases}$$

To see this, let $0 \le x < 2\pi$ and $0 \le y < 4\pi^2$. Then

$$\begin{aligned} F_X(x) &= \mathbb{P}\big(\{\omega \in \Omega : 0 \le X(\omega) \le x\}\big) \\ &= \mathbb{P}\big(\{\omega \in \Omega : 0 \le \omega \le x\}\big) = x/(2\pi), \\ F_Y(y) &= \mathbb{P}\big(\{\omega : Y(\omega) \le y\}\big) \\ &= \mathbb{P}\big(\{\omega : \omega^2 \le y\}\big) = \mathbb{P}\big(\{\omega : 0 \le \omega \le \sqrt{y}\}\big) = \mathbb{P}(X \le \sqrt{y}) \\ &= \sqrt{y}/(2\pi). \end{aligned}$$

The random variables X and Y are continuous because

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du, \qquad F_Y(y) = \int_{-\infty}^{y} f_Y(u)\,du,$$

where

$$f_X(u) = \begin{cases} 1/(2\pi) & \text{if } 0 \le u \le 2\pi, \\ 0 & \text{otherwise,} \end{cases}$$

$$f_Y(u) = \begin{cases} u^{-\frac{1}{2}}/(4\pi) & \text{if } 0 \le u \le 4\pi^2, \\ 0 & \text{otherwise.} \end{cases}$$

●

(5) Example. A random variable which is neither continuous nor discrete. A coin is tossed, and a head turns up with probability $p\,(= 1-q)$. If a head turns up then a rod is flung on the ground and the angle measured as in Example (4). Then $\Omega = \{\mathrm{T}\} \cup \{(\mathrm{H}, x) : 0 \le x < 2\pi\}$, in the obvious notation. Let $X : \Omega \to \mathbb{R}$ be given by

$$X(\mathrm{T}) = -1, \qquad X((\mathrm{H}, x)) = x.$$

The random variable X takes values in $\{-1\} \cup [0, 2\pi)$ (see Figure 2.3 for a sketch of its distribution function). We say that X is continuous with the exception of a 'point mass (or *atom*) at -1'. ●

Exercises for Section 2.3

1. Let X be a random variable with distribution function F, and let $a = (a_m : -\infty < m < \infty)$ be a strictly increasing sequence of real numbers satisfying $a_{-m} \to -\infty$ and $a_m \to \infty$ as $m \to \infty$. Define $G(x) = \mathbb{P}(X \leq a_m)$ when $a_{m-1} \leq x < a_m$, so that G is the distribution function of a discrete random variable. How does the function G behave as the sequence a is chosen in such a way that $\sup_m |a_m - a_{m-1}|$ becomes smaller and smaller?

2. Let X be a random variable and let $g : \mathbb{R} \to \mathbb{R}$ be continuous and strictly increasing. Show that $Y = g(X)$ is a random variable.

3. Let X be a random variable with distribution function

$$\mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x & \text{if } 0 < x \leq 1, \\ 1 & \text{if } x > 1. \end{cases}$$

Let F be a distribution function which is continuous and strictly increasing. Show that $Y = F^{-1}(X)$ is a random variable having distribution function F. Is it necessary that F be continuous and/or strictly increasing?

4. Show that, if f and g are density functions, and $0 \leq \lambda \leq 1$, then $\lambda f + (1 - \lambda)g$ is a density. Is the product fg a density function?

5. Which of the following are density functions? Find c and the corresponding distribution function F for those that are.

(a) $f(x) = \begin{cases} cx^{-d} & x > 1, \\ 0 & \text{otherwise.} \end{cases}$

(b) $f(x) = ce^x(1 + e^x)^{-2}$, $x \in \mathbb{R}$.

2.4 Worked examples

(1) Example. Darts. A dart is flung at a circular target of radius 3. We can think of the hitting point as the outcome of a random experiment; we shall suppose for simplicity that the player is guaranteed to hit the target somewhere. Setting the centre of the target at the origin of $\mathbb{R}^2$, we see that the sample space of this experiment is

$$\Omega = \{(x, y) : x^2 + y^2 < 9\}.$$

Never mind about the collection $\mathcal{F}$ of events. Let us suppose that, roughly speaking, the probability that the dart lands in some region A is proportional to its area $|A|$. Thus

(2) $$\mathbb{P}(A) = |A|/(9\pi).$$

The scoring system is as follows. The target is partitioned by three concentric circles C_1, C_2, and C_3, centered at the origin with radii 1, 2, and 3. These circles divide the target into three annuli A_1, A_2, and A_3, where

$$A_k = \{(x, y) : k - 1 \leq \sqrt{x^2 + y^2} < k\}.$$

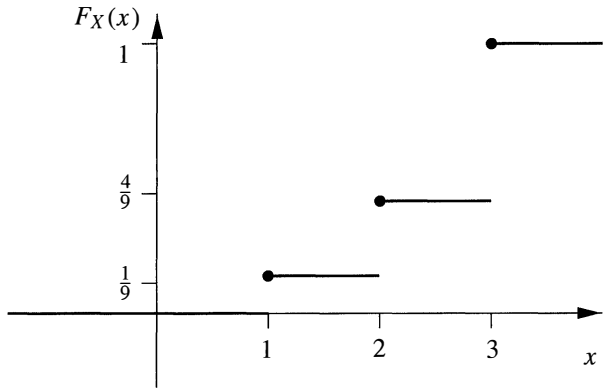

Figure 2.4. The distribution function F_X of X in Example (1).

We suppose that the player scores an amount k if and only if the dart hits A_k. The resulting score X is the random variable given by

$$X(\omega) = k \quad \text{whenever} \quad \omega \in A_k.$$

What is its distribution function?
Solution. Clearly

$$\mathbb{P}(X = k) = \mathbb{P}(A_k) = |A_k|/(9\pi) = \tfrac{1}{9}(2k - 1), \quad \text{for} \quad k = 1, 2, 3,$$

and so the distribution function of X is given by

$$F_X(r) = \mathbb{P}(X \leq r) = \begin{cases} 0 & \text{if } r < 1, \\ \frac{1}{9}\lfloor r \rfloor^2 & \text{if } 1 \leq r < 3, \\ 1 & \text{if } r \geq 3, \end{cases}$$

where $\lfloor r \rfloor$ denotes the largest integer not larger than r (see Figure 2.4). ●

(3) Example. Continuation of (1). Let us consider a revised method of scoring in which the player scores an amount equal to the distance between the hitting point ω and the centre of the target. This time the score Y is a random variable given by

$$Y(\omega) = \sqrt{x^2 + y^2}, \quad \text{if} \quad \omega = (x, y).$$

What is the distribution function of Y?
Solution. For any real r let C_r denote the disc with centre $(0, 0)$ and radius r, that is

$$C_r = \{(x, y) : x^2 + y^2 \leq r\}.$$

Then

$$F_Y(r) = \mathbb{P}(Y \leq r) = \mathbb{P}(C_r) = \tfrac{1}{9}r^2 \qquad \text{if} \qquad 0 \leq r \leq 3.$$

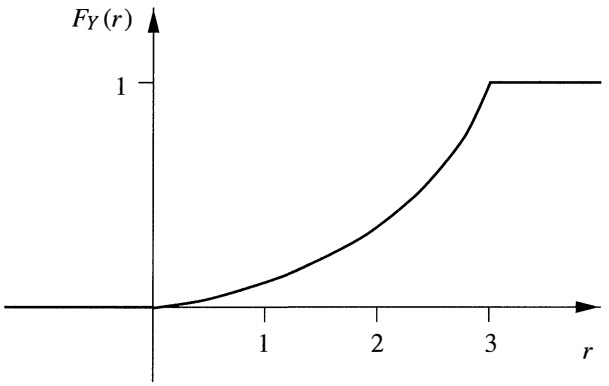

Figure 2.5. The distribution function F_Y of Y in Example (3).

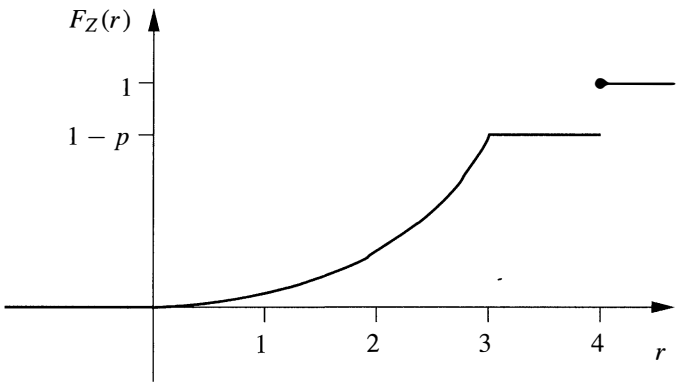

Figure 2.6. The distribution function F_Z of Z in Example (4).

This distribution function is sketched in Figure 2.5. ●

(4) Example. Continuation of (1). Now suppose that the player fails to hit the target with fixed probability p; if he is successful then we suppose that the distribution of the hitting point is described by equation (2). His score is specified as follows. If he hits the target then he scores an amount equal to the distance between the hitting point and the centre; if he misses then he scores 4. What is the distribution function of his score Z?

Solution. Clearly Z takes values in the interval $[0, 4]$. Use Lemma (1.4.4) to see that

$$\begin{aligned} F_Z(r) &= \mathbb{P}(Z \le r) \\ &= \mathbb{P}(Z \le r \mid \text{hits target})\mathbb{P}(\text{hits target}) + \mathbb{P}(Z \le r \mid \text{misses target})\mathbb{P}(\text{misses target}) \\ &= \begin{cases} 0 & \text{if } r < 0, \\ (1-p)F_Y(r) & \text{if } 0 \le r < 4, \\ 1 & \text{if } r \ge 4, \end{cases} \end{aligned}$$

where F_Y is given in Example (3) (see Figure 2.6 for a sketch of F_Z). ●

Exercises for Section 2.4

1. Let X be a random variable with a continuous distribution function F. Find expressions for the distribution functions of the following random variables:

(a) X^2, (b) $\sqrt{X}$,
(c) $\sin X$, (d) $G^{-1}(X)$,
(e) $F(X)$, (f) $G^{-1}(F(X))$,

where G is a continuous and strictly increasing function.

2. Truncation. Let X be a random variable with distribution function F, and let $a < b$. Sketch the distribution functions of the 'truncated' random variables Y and Z given by

$$Y = \begin{cases} a & \text{if } X < a, \\ X & \text{if } a \le X \le b, \\ b & \text{if } X > b, \end{cases} \qquad Z = \begin{cases} X & \text{if } |X| \le b, \\ 0 & \text{if } |X| > b. \end{cases}$$

Indicate how these distribution functions behave as $a \to -\infty$, $b \to \infty$.

2.5 Random vectors

Suppose that X and Y are random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Their distribution functions, F_X and F_Y, contain information about their associated probabilities. But how may we encapsulate information about their properties *relative to each other*? The key is to think of X and Y as being the components of a 'random vector' (X, Y) taking values in $\mathbb{R}^2$, rather than being unrelated random variables each taking values in $\mathbb{R}$.

(1) Example. Tontine is a scheme wherein subscribers to a common fund each receive an annuity from the fund during his or her lifetime, this annuity increasing as the other subscribers die. When all the subscribers are dead, the fund passes to the French government (this was the case in the first such scheme designed by Lorenzo Tonti around 1653). The performance of the fund depends on the lifetimes $L_1, L_2, \ldots, L_n$ of the subscribers (as well as on their wealths), and we may record these as a vector $(L_1, L_2, \ldots, L_n)$ of random variables. ●

(2) Example. Darts. A dart is flung at a conventional dartboard. The point of striking determines a distance R from the centre, an angle Θ with the upward vertical (measured clockwise, say), and a score S. With this experiment we may associate the random vector (R, Θ, S), and we note that S is a function of the pair (R, Θ). ●

(3) Example. Coin tossing. Suppose that we toss a coin n times, and set X_i equal to 0 or 1 depending on whether the ith toss results in a tail or a head. We think of the vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ as describing the result of this composite experiment. The total number of heads is the sum of the entries in $\mathbf{X}$. ●

An individual random variable X has a distribution function F_X defined by $F_X(x) = \mathbb{P}(X \le x)$ for $x \in \mathbb{R}$. The corresponding 'joint' distribution function of a random vector $(X_1, X_2, \ldots, X_n)$ is the quantity $\mathbb{P}(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n)$, a function of n real variables $x_1, x_2, \ldots, x_n$. In order to aid the notation, we introduce an ordering of vectors of

real numbers: for vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ we write $\mathbf{x} \leq \mathbf{y}$ if $x_i \leq y_i$ for each $i = 1, 2, \ldots, n$.

(4) Definition. The **joint distribution function** of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is the function $F_{\mathbf{X}} : \mathbb{R}^n \to [0, 1]$ given by $F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^n$.

As before, the expression $\{\mathbf{X} \leq \mathbf{x}\}$ is an abbreviation for the event $\{\omega \in \Omega : \mathbf{X}(\omega) \leq \mathbf{x}\}$. Joint distribution functions have properties similar to those of ordinary distribution functions. For example, Lemma (2.1.6) becomes the following.

(5) Lemma. *The joint distribution function $F_{X,Y}$ of the random vector (X, Y) has the following properties*:

(a) $\lim_{x,y\to-\infty} F_{X,Y}(x, y) = 0$, $\lim_{x,y\to\infty} F_{X,Y}(x, y) = 1$,

(b) *if* $(x_1, y_1) \leq (x_2, y_2)$ *then* $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$,

(c) $F_{X,Y}$ *is continuous from above, in that*

$$F_{X,Y}(x + u, y + v) \to F_{X,Y}(x, y) \quad \text{as} \quad u, v \downarrow 0.$$

We state this lemma for a random vector with only two components X and Y, but the corresponding result for n components is valid also. The proof of the lemma is left as an *exercise*. Rather more is true. It may be seen without great difficulty that

$$\lim_{y\to\infty} F_{X,Y}(x, y) = F_X(x) \ (= \mathbb{P}(X \leq x)) \tag{6}$$

and similarly

$$\lim_{x\to\infty} F_{X,Y}(x, y) = F_Y(y) \ (= \mathbb{P}(Y \leq y)). \tag{7}$$

This more refined version of part (a) of the lemma tells us that we may recapture the individual distribution functions of X and Y from a knowledge of their joint distribution function. The converse is false: it is not generally possible to calculate $F_{X,Y}$ from a knowledge of F_X and F_Y alone. The functions F_X and F_Y are called the 'marginal' distribution functions of $F_{X,Y}$.

(8) Example. A schoolteacher asks each member of his or her class to flip a fair coin twice and to record the outcomes. The diligent pupil D does this and records a pair (X_D, Y_D) of outcomes. The lazy pupil L flips the coin only once and writes down the result twice, recording thus a pair (X_L, Y_L) where $X_L = Y_L$. Clearly X_D, Y_D, X_L, and Y_L are random variables with the same distribution functions. However, the pairs (X_D, Y_D) and (X_L, Y_L) have different *joint* distribution functions. In particular, $\mathbb{P}(X_D = Y_D = \text{heads}) = \frac{1}{4}$ since only one of the four possible pairs of outcomes contains heads only, whereas $\mathbb{P}(X_L = Y_L = \text{heads}) = \frac{1}{2}$. ●

Once again there are two classes of random vectors which are particularly interesting: the 'discrete' and the 'continuous'.

(9) Definition. The random variables X and Y on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are called **(jointly) discrete** if the vector (X, Y) takes values in some countable subset of $\mathbb{R}^2$ only. The jointly discrete random variables X, Y have **joint (probability) mass function** $f : \mathbb{R}^2 \to [0, 1]$ given by $f(x, y) = \mathbb{P}(X = x, Y = y)$.

(10) Definition. The random variables X and Y on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are called (**jointly**) **continuous** if their joint distribution function can be expressed as

$$F_{X,Y}(x, y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f(u, v)\, du\, dv \qquad x, y \in \mathbb{R},$$

for some integrable function $f : \mathbb{R}^2 \to [0, \infty)$ called the **joint (probability) density function** of the pair (X, Y).

We shall return to such questions in later chapters. Meanwhile here are two concrete examples.

(11) Example. Three-sided coin. We are provided with a special three-sided coin, each toss of which results in one of the possibilities H (heads), T (tails), E (edge), each having probability $\frac{1}{3}$. Let H_n, T_n, and E_n be the numbers of such outcomes in n tosses of the coin. The vector (H_n, T_n, E_n) is a vector of random variables satisfying $H_n + T_n + E_n = n$. If the outcomes of different tosses have no influence on each other, it is not difficult to see that

$$\mathbb{P}\big((H_n, T_n, E_n) = (h, t, e)\big) = \frac{n!}{h!\,t!\,e!}\left(\frac{1}{3}\right)^n$$

for any triple (h, t, e) of non-negative integers with sum n. The random variables H_n, T_n, E_n are (jointly) discrete and are said to have (jointly) the *trinomial* distribution. ●

(12) Example. Darts. Returning to the flung dart of Example (2), let us assume that no region of the dartboard is preferred unduly over any other region of equal area. It may then be shown (see Example (2.4.3)) that

$$\mathbb{P}(R \le r) = \frac{r^2}{\rho^2}, \quad \mathbb{P}(\Theta \le \theta) = \frac{\theta}{2\pi}, \quad \text{for} \quad 0 \le r \le \rho,\ 0 \le \theta \le 2\pi,$$

where ρ is the radius of the board, and furthermore

$$\mathbb{P}(R \le r,\ \Theta \le \theta) = \mathbb{P}(R \le r)\mathbb{P}(\Theta \le \theta).$$

It follows that

$$F_{R,\Theta}(r, \theta) = \int_{u=0}^{r} \int_{v=0}^{\theta} f(u, v)\, du\, dv$$

where

$$f(u, v) = \frac{u}{\pi\rho^2}, \quad 0 \le u \le \rho,\ 0 \le v \le 2\pi.$$

The pair (R, Θ) is (jointly) continuous. ●

Exercises for Section 2.5

1. A fair coin is tossed twice. Let X be the number of heads, and let W be the indicator function of the event $\{X = 2\}$. Find $\mathbb{P}(X = x, W = w)$ for all appropriate values of x and w.

2. Let X be a Bernoulli random variable, so that $\mathbb{P}(X = 0) = 1 - p, \mathbb{P}(X = 1) = p$. Let $Y = 1 - X$ and $Z = XY$. Find $\mathbb{P}(X = x, Y = y)$ and $\mathbb{P}(X = x, Z = z)$ for $x, y, z \in \{0, 1\}$.

3. The random variables X and Y have joint distribution function

$$F_{X,Y}(x, y) = \begin{cases} 0 & \text{if } x < 0, \\ (1 - e^{-x})\left(\frac{1}{2} + \frac{1}{\pi}\tan^{-1} y\right) & \text{if } x \geq 0. \end{cases}$$

Show that X and Y are (jointly) continuously distributed.

4. Let X and Y have joint distribution function F. Show that

$$\mathbb{P}(a < X \leq b, c < Y \leq d) = F(b, d) - F(a, d) - F(b, c) + F(a, c)$$

whenever $a < b$ and $c < d$.

5. Let X, Y be discrete random variables taking values in the integers, with joint mass function f. Show that, for integers x, y,

$$\begin{aligned} f(x, y) = \mathbb{P}(X \geq x, Y \leq y) - \mathbb{P}(X \geq x + 1, Y \leq y) \\ - \mathbb{P}(X \geq x, Y \leq y - 1) + \mathbb{P}(X \geq x + 1, Y \leq y - 1). \end{aligned}$$

Hence find the joint mass function of the smallest and largest numbers shown in r rolls of a fair die.

6. Is the function $F(x, y) = 1 - e^{-xy}, 0 \leq x, y < \infty$, the joint distribution function of some pair of random variables?

2.6 Monte Carlo simulation

It is presumably the case that the physical shape of a coin is one of the major factors relevant to whether or not it will fall with heads uppermost. In principle, the shape of the coin may be determined by direct examination, and hence we may arrive at an estimate for the chance of heads. Unfortunately, such a calculation would be rather complicated, and it is easier to estimate this chance by simulation, which is to say that we may toss the coin many times and record the proportion of successes. Similarly, roulette players are well advised to observe the behaviour of the wheel with care in advance of placing large bets, in order to discern its peculiarities (unfortunately, casinos are now wary of such observation, and change their wheels at regular intervals).

Here is a related question. Suppose that we know that our coin is fair (so that the chance of heads is $\frac{1}{2}$ on each toss), and we wish to know the chance that a sequence of 50 tosses contains a run of outcomes of the form HTHHT. In principle, this probability may be calculated explicitly and exactly. If we require only an estimate of its value, then another possibility is to simulate the experiment: toss the coin $50N$ times for some N, divide the result into N runs of 50, and find the proportion of such runs which contain HTHHT.

It is not unusual in real life for a specific calculation to be possible in principle but extremely difficult in practice, often owing to limitations on the operating speed or the size of the memory of a computer. Simulation can provide a way around such a problem. Here are some examples.

(1) Example. Gambler's ruin revisited. The gambler of Example (1.7.4) eventually won his Jaguar after a long period devoted to tossing coins, and he has now decided to save up for a yacht. His bank manager has suggested that, in order to speed things up, the stake on each gamble should not remain constant but should vary as a certain prescribed function of the gambler's current fortune. The gambler would like to calculate the chance of winning the yacht in advance of embarking on the project, but he finds himself incapable of doing so.

Fortunately, he has kept a record of the extremely long sequence of heads and tails encountered in his successful play for the Jaguar. He calculates his sequence of hypothetical fortunes based on this information, until the point when this fortune reaches either zero or the price of the yacht. He then starts again, and continues to repeat the procedure until he has completed it a total of N times, say. He estimates the probability that he will actually win the yacht by the proportion of the N calculations which result in success.

Can you see why this method will make him overconfident? He might do better to retoss the coins. ●

(2) Example. A dam. It is proposed to build a dam in order to regulate the water supply, and in particular to prevent seasonal flooding downstream. How high should the dam be? Dams are expensive to construct, and some compromise between cost and risk is necessary. It is decided to build a dam which is just high enough to ensure that the chance of a flood of some given extent within ten years is less than 10^{-2}, say. No one knows exactly how high such a dam need be, and a young probabilist proposes the following scheme. Through examination of existing records of rainfall and water demand we may arrive at an acceptable model for the pattern of supply and demand. This model includes, for example, estimates for the distributions of rainfall on successive days over long periods. With the aid of a computer, the 'real world' situation is simulated many times in order to study the likely consequences of building dams of various heights. In this way we may arrive at an accurate estimate of the height required. ●

(3) Example. Integration. Let $g : [0, 1] \to [0, 1]$ be a continuous but nowhere differentiable function. How may we calculate its integral $I = \int_0^1 g(x)\,dx$? The following experimental technique is known as the 'hit or miss Monte Carlo technique'.

Let (X, Y) be a random vector having the uniform distribution on the unit square. That is, we assume that $\mathbb{P}\big((X, Y) \in A\big) = |A|$, the area of A, for any nice subset A of the unit square $[0, 1]^2$; we leave the assumption of niceness somewhat up in the air for the moment, and shall return to such matters in Chapter 4. We declare (X, Y) to be 'successful' if $Y \leq g(X)$. The chance that (X, Y) is successful equals I, the area under the curve $y = g(x)$. We now repeat this experiment a large number N of times, and calculate the proportion of times that the experiment is successful. Following the law of averages, Theorem (2.2.1), we may use this value as an estimate of I.

Clearly it is desirable to know the accuracy of this estimate. This is a harder problem to which we shall return later. ●

Simulation is a dangerous game, and great caution is required in interpreting the results. There are two major reasons for this. First, a computer simulation is limited by the degree to which its so-called 'pseudo-random number generator' may be trusted. It has been said for example that the summon-according-to-birthday principle of conscription to the United States armed forces may have been marred by a pseudo-random number generator with a bias

for some numbers over others. Secondly, in estimating a given quantity, one may in some circumstances have little or no idea how many repetitions are necessary in order to achieve an estimate within a specified accuracy.

We have made no remark about the methods by which computers calculate 'pseudo-random numbers'. Needless to say they do not flip coins, but rely instead on operations of sufficient numerical complexity that the outcome, although deterministic, is apparently unpredictable except by an exact repetition of the calculation.

These techniques were named in honour of Monte Carlo by Metropolis, von Neumann, and Ulam, while they were involved in the process of building bombs at Los Alamos in the 1940s.

2.7 Problems

1. Each toss of a coin results in a head with probability p. The coin is tossed until the first head appears. Let X be the total number of tosses. What is $\mathbb{P}(X > m)$? Find the distribution function of the random variable X.

2. (a) Show that any discrete random variable may be written as a linear combination of indicator variables.

(b) Show that any random variable may be expressed as the limit of an increasing sequence of discrete random variables.

(c) Show that the limit of any increasing convergent sequence of random variables is a random variable.

3. (a) Show that, if X and Y are random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then so are $X + Y$, XY, and $\min\{X, Y\}$.

(b) Show that the set of all random variables on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ constitutes a vector space over the reals. If Ω is finite, write down a basis for this space.

4. Let X have distribution function

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{2}x & \text{if } 0 \le x \le 2, \\ 1 & \text{if } x > 2, \end{cases}$$

and let $Y = X^2$. Find

(a) $\mathbb{P}(\frac{1}{2} \le X \le \frac{3}{2})$, (b) $\mathbb{P}(1 \le X < 2)$,
(c) $\mathbb{P}(Y \le X)$, (d) $\mathbb{P}(X \le 2Y)$,
(e) $\mathbb{P}(X + Y \le \frac{3}{4})$, (f) the distribution function of $Z = \sqrt{X}$.

5. Let X have distribution function

$$F(x) = \begin{cases} 0 & \text{if } x < -1, \\ 1 - p & \text{if } -1 \le x < 0, \\ 1 - p + \frac{1}{2}xp & \text{if } 0 \le x \le 2, \\ 1 & \text{if } x > 2. \end{cases}$$

Sketch this function, and find: (a) $\mathbb{P}(X = -1)$, (b) $\mathbb{P}(X = 0)$, (c) $\mathbb{P}(X \ge 1)$.

6. Buses arrive at ten minute intervals starting at noon. A man arrives at the bus stop a random number X minutes after noon, where X has distribution function

$$\mathbb{P}(X \le x) = \begin{cases} 0 & \text{if } x < 0, \\ x/60 & \text{if } 0 \le x \le 60, \\ 1 & \text{if } x > 60. \end{cases}$$

What is the probability that he waits less than five minutes for a bus?

7. Airlines find that each passenger who reserves a seat fails to turn up with probability $\frac{1}{10}$ independently of the other passengers. So Teeny Weeny Airlines always sell 10 tickets for their 9 seat aeroplane while Blockbuster Airways always sell 20 tickets for their 18 seat aeroplane. Which is more often over-booked?

8. A fairground performer claims the power of telekinesis. The crowd throws coins and he wills them to fall heads up. He succeeds five times out of six. What chance would he have of doing at least as well if he had no supernatural powers?

9. Express the distribution functions of

$$X^+ = \max\{0, X\}, \quad X^- = -\min\{0, X\}, \quad |X| = X^+ + X^-, \quad -X,$$

in terms of the distribution function F of the random variable X.

10. Show that $F_X(x)$ is continuous at $x = x_0$ if and only if $\mathbb{P}(X = x_0) = 0$.

11. The real number m is called a *median* of the distribution function F whenever $\lim_{y \uparrow m} F(y) \leq \frac{1}{2} \leq F(m)$. Show that every distribution function F has at least one median, and that the set of medians of F is a closed interval of $\mathbb{R}$.

12. Show that it is not possible to weight two dice in such a way that the sum of the two numbers shown by these loaded dice is equally likely to take any value between 2 and 12 (inclusive).

13. A function $d : S \times S \to \mathbb{R}$ is called a *metric* on S if:
(i) $d(s, t) = d(t, s) \geq 0$ for all $s, t \in S$,
(ii) $d(s, t) = 0$ if and only if $s = t$, and
(iii) $d(s, t) \leq d(s, u) + d(u, t)$ for all $s, t, u \in S$.
(a) **Lévy metric.** Let F and G be distribution functions and define the *Lévy metric*

$$d_{\mathrm{L}}(F, G) = \inf\Big\{\epsilon > 0 : G(x - \epsilon) - \epsilon \leq F(x) \leq G(x + \epsilon) + \epsilon \text{ for all } x\Big\}.$$

Show that d_{L} is indeed a metric on the space of distribution functions.
(b) **Total variation distance.** Let X and Y be integer-valued random variables, and let

$$d_{\mathrm{TV}}(X, Y) = \sum_k \big|\mathbb{P}(X = k) - \mathbb{P}(Y = k)\big|.$$

Show that d_{TV} satisfies (i) and (iii) with S the space of integer-valued random variables, and that $d_{\mathrm{TV}}(X, Y) = 0$ if and only if $\mathbb{P}(X = Y) = 1$. Thus d_{TV} is a metric on the space of equivalence classes of S with equivalence relation given by $X \sim Y$ if $\mathbb{P}(X = Y) = 1$. We call d_{TV} the *total variation distance*.

Show that

$$d_{\mathrm{TV}}(X, Y) = 2 \sup_{A \subseteq \mathbb{Z}} \big|\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)\big|.$$

14. Ascertain in the following cases whether or not F is the joint distribution function of some pair (X, Y) of random variables. If your conclusion is affirmative, find the distribution functions of X and Y separately.

(a)
$$F(x, y) = \begin{cases} 1 - e^{-x-y} & \text{if } x, y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

(b)
$$F(x, y) = \begin{cases} 1 - e^{-x} - xe^{-y} & \text{if } 0 \leq x \leq y, \\ 1 - e^{-y} - ye^{-y} & \text{if } 0 \leq y \leq x, \\ 0 & \text{otherwise.} \end{cases}$$

15. It is required to place in order n books $B_1, B_2, \ldots, B_n$ on a library shelf in such a way that readers searching from left to right waste as little time as possible on average. Assuming that each reader requires book B_i with probability p_i, find the ordering of the books which minimizes $\mathbb{P}(T \geq k)$ for all k, where T is the (random) number of titles examined by a reader before discovery of the required book.

16. Transitive coins. Three coins each show heads with probability $\frac{3}{5}$ and tails otherwise. The first counts 10 points for a head and 2 for a tail, the second counts 4 points for both head and tail, and the third counts 3 points for a head and 20 for a tail.

You and your opponent each choose a coin; you cannot choose the same coin. Each of you tosses your coin and the person with the larger score wins $£10^{10}$. Would you prefer to be the first to pick a coin or the second?

17. Before the development of radar and inertial navigation, flying to isolated islands (for example, from Los Angeles to Hawaii) was somewhat 'hit or miss'. In heavy cloud or at night it was necessary to fly by dead reckoning, and then to search the surface. With the aid of a radio, the pilot had a good idea of the correct great circle along which to search, but could not be sure which of the two directions along this great circle was correct (since a strong tailwind could have carried the plane over its target). When you are the pilot, you calculate that you can make n searches before your plane will run out of fuel. On each search you will discover the island with probability p (if it is indeed in the direction of the search) independently of the results of other searches; you estimate initially that there is probability α that the island is ahead of you. What policy should you adopt in deciding the directions of your various searches in order to maximize the probability of locating the island?

18. Eight pawns are placed randomly on a chessboard, no more than one to a square. What is the probability that:

(a) they are in a straight line (do not forget the diagonals)?

(b) no two are in the same row or column?

19. Which of the following are distribution functions? For those that are, give the corresponding density function f.

(a) $F(x) = \begin{cases} 1 - e^{-x^2} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$

(b) $F(x) = \begin{cases} e^{-1/x} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$

(c) $F(x) = e^x/(e^x + e^{-x})$, $x \in \mathbb{R}$.

(d) $F(x) = e^{-x^2} + e^x/(e^x + e^{-x})$, $x \in \mathbb{R}$.

20. (a) If U and V are jointly continuous, show that $\mathbb{P}(U = V) = 0$.

(b) Let X be uniformly distributed on $(0, 1)$, and let $Y = X$. Then X and Y are continuous, and $\mathbb{P}(X = Y) = 1$. Is there a contradiction here?

3

Discrete random variables

Summary. The distribution of a discrete random variable may be specified via its probability mass function. The key notion of independence for discrete random variables is introduced. The concept of expectation, or mean value, is defined for discrete variables, leading to a definition of the variance and the moments of a discrete random variable. Joint distributions, conditional distributions, and conditional expectation are introduced, together with the ideas of covariance and correlation. The Cauchy–Schwarz inequality is presented. The analysis of sums of random variables leads to the convolution formula for mass functions. Random walks are studied in some depth, including the reflection principle, the ballot theorem, the hitting time theorem, and the arc sine laws for visits to the origin and for sojourn times.

3.1 Probability mass functions

Recall that a random variable X is *discrete* if it takes values only in some countable set $\{x_1, x_2, \dots\}$. Its distribution function $F(x) = \mathbb{P}(X \le x)$ is a jump function; just as important as its distribution function is its mass function.

(1) Definition. The (**probability**) **mass function**†of a discrete random variable X is the function $f : \mathbb{R} \to [0, 1]$ given by $f(x) = \mathbb{P}(X = x)$.

The distribution and mass functions are related by

$$F(x) = \sum_{i : x_i \le x} f(x_i), \qquad f(x) = F(x) - \lim_{y \uparrow x} F(y).$$

(2) Lemma. *The probability mass function $f : \mathbb{R} \to [0, 1]$ satisfies*:
(a) *the set of x such that $f(x) \neq 0$ is countable,*
(b) $\sum_i f(x_i) = 1$, *where $x_1, x_2, \dots$ are the values of x such that $f(x) \neq 0$.*

Proof. The proof is obvious. ■

This lemma characterizes probability mass functions.

†Some refer loosely to the mass function of X as its distribution.

(3) Example. Binomial distribution. A coin is tossed n times, and a head turns up each time with probability p $(= 1 - q)$. Then $\Omega = \{\mathrm{H}, \mathrm{T}\}^n$. The total number X of heads takes values in the set $\{0, 1, 2, \ldots, n\}$ and is a discrete random variable. Its probability mass function $f(x) = \mathbb{P}(X = x)$ satisfies

$$f(x) = 0 \quad \text{if} \quad x \notin \{0, 1, 2, \ldots, n\}.$$

Let $0 \leq k \leq n$, and consider $f(k)$. Exactly $\binom{n}{k}$ points in Ω give a total of k heads; each of these points occurs with probability $p^k q^{n-k}$, and so

$$f(k) = \binom{n}{k} p^k q^{n-k} \quad \text{if} \quad 0 \leq k \leq n.$$

The random variable X is said to have the *binomial distribution* with parameters n and p, written bin(n, p). It is the sum $X = Y_1 + Y_2 + \cdots + Y_n$ of n Bernoulli variables (see Example (2.1.8)). ●

(4) Example. Poisson distribution. If a random variable X takes values in the set $\{0, 1, 2, \ldots\}$ with mass function

$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \ldots,$$

where $\lambda > 0$, then X is said to have the *Poisson distribution* with parameter λ. ●

Exercises for Section 3.1

1. For what values of the constant C do the following define mass functions on the positive integers $1, 2, \ldots$?

(a) Geometric: $f(x) = C2^{-x}$.
(b) Logarithmic: $f(x) = C2^{-x}/x$.
(c) Inverse square: $f(x) = Cx^{-2}$.
(d) 'Modified' Poisson: $f(x) = C2^x/x!$.

2. For a random variable X having (in turn) each of the four mass functions of Exercise (1), find:

(i) $\mathbb{P}(X > 1)$,
(ii) the most probable value of X,
(iii) the probability that X is even.

3. We toss n coins, and each one shows heads with probability p, independently of each of the others. Each coin which shows heads is tossed again. What is the mass function of the number of heads resulting from the second round of tosses?

4. Let S_k be the set of positive integers whose base-10 expansion contains exactly k elements (so that, for example, $1024 \in S_4$). A fair coin is tossed until the first head appears, and we write T for the number of tosses required. We pick a random element, N say, from S_T, each such element having equal probability. What is the mass function of N?

5. Log-convexity. (a) Show that, if X is a binomial or Poisson random variable, then the mass function $f(k) = \mathbb{P}(X = k)$ has the property that $f(k-1)f(k+1) \leq f(k)^2$.
(b) Show that, if $f(k) = 90/(\pi k)^4$, $k \geq 1$, then $f(k-1)f(k+1) \geq f(k)^2$.
(c) Find a mass function f such that $f(k)^2 = f(k-1)f(k+1)$, $k \geq 1$.

3.2 Independence

Remember that events A and B are called 'independent' if the occurrence of A does not change the subsequent probability of B occurring. More rigorously, A and B are independent if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Similarly, we say that discrete variables X and Y are 'independent' if the numerical value of X does not affect the distribution of Y. With this in mind we make the following definition.

(1) Definition. Discrete variables X and Y are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all x and y.

Suppose X takes values in the set $\{x_1, x_2, \dots\}$ and Y takes values in the set $\{y_1, y_2, \dots\}$. Let

$$A_i = \{X = x_i\}, \qquad B_j = \{Y = y_j\}.$$

Notice (see Problem (2.7.2)) that X and Y are linear combinations of the indicator variables I_{A_i}, I_{B_j}, in that

$$X = \sum_i x_i I_{A_i} \qquad \text{and} \qquad Y = \sum_j y_j I_{B_j}.$$

The random variables X and Y are independent if and only if A_i and B_j are independent for all pairs i, j. A similar definition holds for collections $\{X_1, X_2, \dots, X_n\}$ of discrete variables.

(2) Example. Poisson flips. A coin is tossed once and heads turns up with probability $p = 1 - q$. Let X and Y be the numbers of heads and tails respectively. It is no surprise that X and Y are not independent. After all,

$$\mathbb{P}(X = Y = 1) = 0, \qquad \mathbb{P}(X = 1)\mathbb{P}(Y = 1) = p(1 - p).$$

Suppose now that the coin is tossed a random number N of times, where N has the Poisson distribution with parameter λ. It is a remarkable fact that the resulting numbers X and Y of heads and tails *are* independent, since

$$\begin{aligned}\mathbb{P}(X = x,\ Y = y) &= \mathbb{P}\big(X = x, Y = y \,\big|\, N = x + y\big)\mathbb{P}(N = x + y) \\ &= \binom{x+y}{x} p^x q^y \frac{\lambda^{x+y}}{(x+y)!} e^{-\lambda} = \frac{(\lambda p)^x (\lambda p)^y}{x!\, y!} e^{-\lambda}.\end{aligned}$$

However, by Lemma (1.4.4),

$$\begin{aligned}\mathbb{P}(X = x) &= \sum_{n \geq x} \mathbb{P}(X = x \mid N = n)\mathbb{P}(N = n) \\ &= \sum_{n \geq x} \binom{n}{x} p^x q^{n-x} \frac{\lambda^n}{n!} e^{-\lambda} = \frac{(\lambda p)^x}{x!} e^{-\lambda p};\end{aligned}$$

a similar result holds for Y, and so

$$\mathbb{P}(X = x,\ Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y). \qquad \bullet$$

If X is a random variable and $g : \mathbb{R} \to \mathbb{R}$, then $Z = g(X)$, defined by $Z(\omega) = g(X(\omega))$, is a random variable also. We shall need the following.

(3) Theorem. *If X and Y are independent and $g, h : \mathbb{R} \to \mathbb{R}$, then $g(X)$ and $h(Y)$ are independent also.*

Proof. *Exercise.* See Problem (3.11.1). ■

More generally, we say that a family $\{X_i : i \in I\}$ of (discrete) random variables is *independent* if the events $\{X_i = x_i\}$, $i \in I$, are independent for all possible choices of the set $\{x_i : i \in I\}$ of the values of the X_i. That is to say, $\{X_i : i \in I\}$ is an independent family if and only if

$$\mathbb{P}(X_i = x_i \text{ for all } i \in J) = \prod_{i \in J} \mathbb{P}(X_i = x_i)$$

for all sets $\{x_i : i \in I\}$ and for all finite subsets J of I. The conditional independence of a family of random variables, given an event C, is defined similarly to the conditional independence of events; see equation (1.5.5).

Independent families of random variables are very much easier to study than dependent families, as we shall see soon. Note that pairwise-independent families are not necessarily independent.

Exercises for Section 3.2

1. Let X and Y be independent random variables, each taking the values -1 or 1 with probability $\frac{1}{2}$, and let $Z = XY$. Show that X, Y, and Z are pairwise independent. Are they independent?

2. Let X and Y be independent random variables taking values in the positive integers and having the same mass function $f(x) = 2^{-x}$ for $x = 1, 2, \ldots$. Find:

(a) $\mathbb{P}(\min\{X, Y\} \leq x)$, (b) $\mathbb{P}(Y > X)$,
(c) $\mathbb{P}(X = Y)$, (d) $\mathbb{P}(X \geq kY)$, for a given positive integer k,
(e) $\mathbb{P}(X \text{ divides } Y)$, (f) $\mathbb{P}(X = rY)$, for a given positive rational r.

3. Let X_1, X_2, X_3 be independent random variables taking values in the positive integers and having mass functions given by $\mathbb{P}(X_i = x) = (1 - p_i)p_i^{x-1}$ for $x = 1, 2, \ldots$, and $i = 1, 2, 3$.

(a) Show that

$$\mathbb{P}(X_1 < X_2 < X_3) = \frac{(1 - p_1)(1 - p_2)p_2 p_3^2}{(1 - p_2 p_3)(1 - p_1 p_2 p_3)}.$$

(b) Find $\mathbb{P}(X_1 \leq X_2 \leq X_3)$.

4. Three players, A, B, and C, take turns to roll a die; they do this in the order ABCABCA....

(a) Show that the probability that, of the three players, A is the first to throw a 6, B the second, and C the third, is 216/1001.

(b) Show that the probability that the first 6 to appear is thrown by A, the second 6 to appear is thrown by B, and the third 6 to appear is thrown by C, is 46656/753571.

5. Let X_r, $1 \leq r \leq n$, be independent random variables which are symmetric about 0; that is, X_r and $-X_r$ have the same distributions. Show that, for all x, $\mathbb{P}(S_n \geq x) = \mathbb{P}(S_n \leq -x)$ where $S_n = \sum_{r=1}^{n} X_r$.

Is the conclusion necessarily true without the assumption of independence?

3.3 Expectation

Let $x_1, x_2, \ldots, x_N$ be the numerical outcomes of N repetitions of some experiment. The average of these outcomes is

$$m = \frac{1}{N} \sum_i x_i.$$

In advance of performing these experiments we can represent their outcomes by a sequence $X_1, X_2, \ldots, X_N$ of random variables, and we shall suppose that these variables are discrete with a common mass function f. Then, roughly speaking (see the beginning of Section 1.3), for each possible value x, about $Nf(x)$ of the X_i will take that value x. So the average m is about

$$m \simeq \frac{1}{N} \sum_x xNf(x) = \sum_x xf(x)$$

where the summation here is over all possible values of the X_i. This average is called the 'expectation' or 'mean value' of the underlying distribution with mass function f.

(1) Definition. The **mean value**, or **expectation**, or **expected value** of the random variable X with mass function f is defined to be

$$\mathbb{E}(X) = \sum_{x:f(x)>0} xf(x)$$

whenever this sum is absolutely convergent.

We require *absolute* convergence in order that $\mathbb{E}(X)$ be unchanged by reordering the x_i. We can, for notational convenience, write $\mathbb{E}(X) = \sum_x xf(x)$. This appears to be an uncountable sum; however, all but countably many of its contributions are zero. If the numbers $f(x)$ are regarded as masses $f(x)$ at points x then $\mathbb{E}(X)$ is just the position of the centre of gravity; we can speak of X as having an 'atom' or 'point mass' of size $f(x)$ at x. We sometimes omit the parentheses and simply write $\mathbb{E}X$.

(2) Example (2.1.5) revisited. The random variables X and W of this example have mean values

$$\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x) = 0 \cdot \tfrac{1}{4} + 1 \cdot \tfrac{1}{2} + 2 \cdot \tfrac{1}{4} = 1,$$
$$\mathbb{E}(W) = \sum_x x\mathbb{P}(W = x) = 0 \cdot \tfrac{3}{4} + 4 \cdot \tfrac{1}{4} = 1. \quad \bullet$$

If X is a random variable and $g : \mathbb{R} \to \mathbb{R}$, then $Y = g(X)$, given formally by $Y(\omega) = g(X(\omega))$, is a random variable also. To calculate its expectation we need first to find its probability mass function f_Y. This process can be complicated, and it is avoided by the following lemma (called by some the 'law of the unconscious statistician'!).

(3) Lemma. *If X has mass function f and $g : \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}(g(X)) = \sum_x g(x)f(x)$$

whenever this sum is absolutely convergent.

Proof. This is Problem (3.11.3). ■

(4) Example. Suppose that X takes values $-2, -1, 1, 3$ with probabilities $\frac{1}{4}, \frac{1}{8}, \frac{1}{4}, \frac{3}{8}$ respectively. The random variable $Y = X^2$ takes values $1, 4, 9$ with probabilities $\frac{3}{8}, \frac{1}{4}, \frac{3}{8}$ respectively, and so

$$\mathbb{E}(Y) = \sum_x x\mathbb{P}(Y = x) = 1 \cdot \tfrac{3}{8} + 4 \cdot \tfrac{1}{4} + 9 \cdot \tfrac{3}{8} = \tfrac{19}{4}.$$

Alternatively, use the law of the unconscious statistician to find that

$$\mathbb{E}(Y) = \mathbb{E}(X^2) = \sum_x x^2\mathbb{P}(X = x) = 4 \cdot \tfrac{1}{4} + 1 \cdot \tfrac{1}{8} + 1 \cdot \tfrac{1}{4} + 9 \cdot \tfrac{3}{8} = \tfrac{19}{4}.$$ ●

Lemma (3) provides a method for calculating the 'moments' of a distribution; these are defined as follows.

(5) Definition. If k is a positive integer, the kth **moment** m_k of X is defined to be $m_k = \mathbb{E}(X^k)$. The kth **central moment** σ_k is $\sigma_k = \mathbb{E}((X - m_1)^k)$.

The two moments of most use are $m_1 = \mathbb{E}(X)$ and $\sigma_2 = \mathbb{E}((X - \mathbb{E}X)^2)$, called the *mean* (or *expectation*) and *variance* of X. These two quantities are measures of the mean and dispersion of X; that is, m_1 is the average value of X, and σ_2 measures the amount by which X tends to deviate from this average. The mean m_1 is often denoted μ, and the variance of X is often denoted $\operatorname{var}(X)$. The positive square root $\sigma = \sqrt{\operatorname{var}(X)}$ is called the *standard deviation*, and in this notation $\sigma_2 = \sigma^2$. The central moments $\{\sigma_i\}$ can be expressed in terms of the ordinary moments $\{m_i\}$. For example, $\sigma_1 = 0$ and

$$\begin{aligned}\sigma_2 &= \sum_x (x - m_1)^2 f(x) \\ &= \sum_x x^2 f(x) - 2m_1 \sum_x x f(x) + m_1^2 \sum_x f(x) \\ &= m_2 - m_1^2,\end{aligned}$$

which may be written as

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

Remark. Experience with student calculations of variances causes us to stress the following elementary fact: *variances cannot be negative*. We sometimes omit the parentheses and write simply $\operatorname{var} X$. The expression $\mathbb{E}(X)^2$ means $(\mathbb{E}(X))^2$ and must not be confused with $\mathbb{E}(X^2)$.

(6) Example. Bernoulli variables. Let X be a Bernoulli variable, taking the value 1 with probability p $(= 1 - q)$. Then

$$\mathbb{E}(X) = \sum_x xf(x) = 0 \cdot q + 1 \cdot p = p,$$
$$\mathbb{E}(X^2) = \sum_x x^2 f(x) = 0 \cdot q + 1 \cdot p = p,$$
$$\text{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = pq.$$

Thus the indicator variable I_A has expectation $\mathbb{P}(A)$ and variance $\mathbb{P}(A)\mathbb{P}(A^c)$. ●

(7) Example. Binomial variables. Let X be bin(n, p). Then

$$\mathbb{E}(X) = \sum_{k=0}^{n} kf(k) = \sum_{k=0}^{n} k\binom{n}{k} p^k q^{n-k}.$$

To calculate this, differentiate the identity

$$\sum_{k=0}^{n} \binom{n}{k} x^k = (1 + x)^n,$$

multiply by x to obtain

$$\sum_{k=0}^{n} k\binom{n}{k} x^k = nx(1 + x)^{n-1},$$

and substitute $x = p/q$ to obtain $\mathbb{E}(X) = np$. A similar argument shows that the variance of X is given by $\text{var}(X) = npq$. ●

We can think of the process of calculating expectations as a linear operator on the space of random variables.

(8) Theorem. *The expectation operator* $\mathbb{E}$ *has the following properties*:
(a) *if* $X \geq 0$ *then* $\mathbb{E}(X) \geq 0$,
(b) *if* $a, b \in \mathbb{R}$ *then* $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$,
(c) *the random variable* 1, *taking the value* 1 *always, has expectation* $\mathbb{E}(1) = 1$.

Proof. (a) and (c) are obvious.
(b) Let $A_x = \{X = x\}$, $B_y = \{Y = y\}$. Then

$$aX + bY = \sum_{x,y} (ax + by) I_{A_x \cap B_y}$$

and the solution of the first part of Problem (3.11.3) shows that

$$\mathbb{E}(aX + bY) = \sum_{x,y} (ax + by)\mathbb{P}(A_x \cap B_y).$$

However,

$$\sum_y \mathbb{P}(A_x \cap B_y) = \mathbb{P}\left(A_x \cap \left(\bigcup_y B_y\right)\right) = \mathbb{P}(A_x \cap \Omega) = \mathbb{P}(A_x)$$

and similarly $\sum_x \mathbb{P}(A_x \cap B_y) = \mathbb{P}(B_y)$, which gives

$$\begin{aligned}\mathbb{E}(aX + bY) &= \sum_x ax \sum_y \mathbb{P}(A_x \cap B_y) + \sum_y by \sum_x \mathbb{P}(A_x \cap B_y) \\ &= a\sum_x x\mathbb{P}(A_x) + b\sum_y y\mathbb{P}(B_y) \\ &= a\mathbb{E}(X) + b\mathbb{E}(Y).\end{aligned}$$

■

Remark. It is not in general true that $\mathbb{E}(XY)$ is the same as $\mathbb{E}(X)\mathbb{E}(Y)$.

(9) Lemma. *If X and Y are independent then* $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Proof. Let A_x and B_y be as in the proof of (8). Then

$$XY = \sum_{x,y} xy I_{A_x \cap B_y}$$

and so

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

■

(10) Definition. X and Y are called **uncorrelated** if $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Lemma (9) asserts that independent variables are uncorrelated. The converse is not true, as Problem (3.11.16) indicates.

(11) Theorem. *For random variables X and Y,*

(a) $\operatorname{var}(aX) = a^2 \operatorname{var}(X)$ *for* $a \in \mathbb{R}$,

(b) $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y)$ *if X and Y are uncorrelated.*

Proof. (a) Using the linearity of $\mathbb{E}$,

$$\begin{aligned}\operatorname{var}(aX) &= \mathbb{E}\{(aX - \mathbb{E}(aX))^2\} = \mathbb{E}\{a^2(X - \mathbb{E}X)^2\} \\ &= a^2\mathbb{E}\{(X - \mathbb{E}X)^2\} = a^2 \operatorname{var}(X).\end{aligned}$$

(b) We have when X and Y are uncorrelated that

$$\begin{aligned}\operatorname{var}(X + Y) &= \mathbb{E}\{(X + Y - \mathbb{E}(X + Y))^2\} \\ &= \mathbb{E}\big[(X - \mathbb{E}X)^2 + 2(XY - \mathbb{E}(X)\mathbb{E}(Y)) + (Y - \mathbb{E}Y)^2\big] \\ &= \operatorname{var}(X) + 2[\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)] + \operatorname{var}(Y) \\ &= \operatorname{var}(X) + \operatorname{var}(Y).\end{aligned}$$

■

Theorem (11a) shows that the variance operator 'var' is *not* a linear operator, even when it is applied only to uncorrelated variables.

Sometimes the sum $S = \sum xf(x)$ does not converge absolutely, and the mean of the distribution does not exist. If $S = -\infty$ or $S = +\infty$, then we can sometimes speak of the mean as taking these values also. Of course, there exist distributions which do not have a mean value.

(12) Example. A distribution without a mean. Let X have mass function

$$f(k) = Ak^{-2} \quad \text{for} \quad k = \pm 1, \pm 2, \ldots$$

where A is chosen so that $\sum f(k) = 1$. The sum $\sum_k kf(k) = A \sum_{k \neq 0} k^{-1}$ does not converge absolutely, because both the positive and the negative parts diverge. ●

This is a suitable opportunity to point out that we can base probability theory upon the expectation operator $\mathbb{E}$ rather than upon the probability measure $\mathbb{P}$. After all, our intuitions about the notion of 'average' are probably just as well developed as those about quantitative chance. Roughly speaking, the way we proceed is to postulate axioms, such as (a), (b), and (c) of Theorem (8), for a so-called 'expectation operator' $\mathbb{E}$ acting on a space of 'random variables'. The probability of an event can then be recaptured by defining $\mathbb{P}(A) = \mathbb{E}(I_A)$. Whittle (2000) is an able advocate of this approach.

This method can be easily and naturally adapted to deal with probabilistic questions in quantum theory. In this major branch of theoretical physics, questions arise which cannot be formulated entirely within the usual framework of probability theory. However, there still exists an expectation operator $\mathbb{E}$, which is applied to linear operators known as observables (such as square matrices) rather than to random variables. There does not exist a sample space Ω, and nor therefore are there any indicator functions, but nevertheless there exist analogues of other concepts in probability theory. For example, the *variance* of an operator X is defined by $\text{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. Furthermore, it can be shown that $\mathbb{E}(X) = \text{tr}(UX)$ where tr denotes *trace* and U is a non-negative definite operator with unit trace.

(13) Example. Wagers. Historically, there has been confusion amongst probabilists between the price that an individual may be willing to pay in order to play a game, and her expected return from this game. For example, I conceal £2 in one hand and nothing in the other, and then invite a friend to pay a fee which entitles her to choose a hand at random and keep the contents. Other things being equal (my friend is neither a compulsive gambler, nor particularly busy), it would seem that £1 would be a 'fair' fee to ask, since £1 is the expected return to the player. That is to say, faced with a modest (but random) gain, then a fair 'entrance fee' would seem to be the expected value of the gain. However, suppose that I conceal $£2^{10}$ in one hand and nothing in the other; what now is a 'fair' fee? Few persons of modest means can be expected to offer $£2^9$ for the privilege of playing. There is confusion here between fairness and reasonableness: we do not generally treat large payoffs or penalties in the same way as small ones, even though the relative odds may be unquestionable. The customary resolution of this paradox is to introduce the notion of 'utility'. Writing $u(x)$ for the 'utility' to an individual of £x, it would be fairer to charge a fee of $\frac{1}{2}(u(0) + u(2^{10}))$ for the above prospect. Of course, different individuals have different utility functions, although such functions have presumably various features in common: $u(0) = 0$, u is non-decreasing, $u(x)$ is near to x for small positive x, and u is concave, so that in particular $u(x) \leq xu(1)$ when $x \geq 1$.

The use of expectation to assess a 'fair fee' may be convenient but is sometimes inappropriate. For example, a more suitable criterion in the finance market would be absence of arbitrage; see Exercise (3.3.7) and Section 13.10. And, in a rather general model of financial markets, there is a criterion commonly expressed as 'no free lunch with vanishing risk'. ●

Exercises for Section 3.3

1. Is it generally true that $\mathbb{E}(1/X) = 1/\mathbb{E}(X)$? Is it ever true that $\mathbb{E}(1/X) = 1/\mathbb{E}(X)$?

2. Coupons. Every package of some intrinsically dull commodity includes a small and exciting plastic object. There are c different types of object, and each package is equally likely to contain any given type. You buy one package each day.

(a) Find the mean number of days which elapse between the acquisitions of the jth new type of object and the $(j+1)$th new type.

(b) Find the mean number of days which elapse before you have a full set of objects.

3. Each member of a group of n players rolls a die.

(a) For any pair of players who throw the same number, the group scores 1 point. Find the mean and variance of the total score of the group.

(b) Find the mean and variance of the total score if any pair of players who throw the same number scores that number.

4. St Petersburg paradox†**.** A fair coin is tossed repeatedly. Let T be the number of tosses until the first head. You are offered the following prospect, which you may accept on payment of a fee. If $T = k$, say, then you will receive $£2^k$. What would be a 'fair' fee to ask of you?

5. Let X have mass function

$$f(x) = \begin{cases} \{x(x+1)\}^{-1} & \text{if } x = 1, 2, \ldots, \\ 0 & \text{otherwise,} \end{cases}$$

and let $\alpha \in \mathbb{R}$. For what values of α is it the case‡ that $\mathbb{E}(X^\alpha) < \infty$?

6. Show that $\text{var}(a + X) = \text{var}(X)$ for any random variable X and constant a.

7. Arbitrage. Suppose you find a warm-hearted bookmaker offering payoff odds of $\pi(k)$ against the kth horse in an n-horse race where $\sum_{k=1}^{n} \{\pi(k) + 1\}^{-1} < 1$. Show that you can distribute your bets in such a way as to ensure you win.

8. You roll a conventional fair die repeatedly. If it shows 1, you must stop, but you may choose to stop at any prior time. Your score is the number shown by the die on the final roll. What stopping strategy yields the greatest expected score? What strategy would you use if your score were the square of the final roll?

9. Continuing with Exercise (8), suppose now that you lose c points from your score each time you roll the die. What strategy maximizes the expected final score if $c = \frac{1}{3}$? What is the best strategy if $c = 1$?

---

†This problem was mentioned by Nicholas Bernoulli in 1713, and Daniel Bernoulli wrote about the question for the Academy of St Petersburg.

‡If α is not integral, than $\mathbb{E}(X^\alpha)$ is called the *fractional moment of order* α of X. A point concerning notation: for real α and complex $x = re^{i\theta}$, x^α should be interpreted as $r^\alpha e^{i\theta\alpha}$, so that $|x^\alpha| = r^\alpha$. In particular, $\mathbb{E}(|X^\alpha|) = \mathbb{E}(|X|^\alpha)$.

3.4 Indicators and matching

This section contains light entertainment, in the guise of some illustrations of the uses of indicator functions. These were defined in Example (2.1.9) and have appeared occasionally since. Recall that

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \in A^c, \end{cases}$$

and $\mathbb{E}I_A = \mathbb{P}(A)$.

(1) Example. Proofs of Lemma (1.3.4c, d). Note that

$$I_A + I_{A^c} = I_{A \cup A^c} = I_\Omega = 1$$

and that $I_{A \cap B} = I_A I_B$. Thus

$$\begin{aligned} I_{A \cup B} &= 1 - I_{(A \cup B)^c} = 1 - I_{A^c \cap B^c} \\ &= 1 - I_{A^c} I_{B^c} = 1 - (1 - I_A)(1 - I_B) \\ &= I_A + I_B - I_A I_B. \end{aligned}$$

Take expectations to obtain

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

More generally, if $B = \bigcup_{i=1}^n A_i$ then

$$I_B = 1 - \prod_{i=1}^n (1 - I_{A_i});$$

multiply this out and take expectations to obtain

$$\textbf{(2)} \qquad \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n+1} \mathbb{P}(A_1 \cap \cdots \cap A_n).$$

This very useful identity is known as the *inclusion–exclusion formula.* ●

(3) Example. Matching problem. A number of melodramatic applications of (2) are available, of which the following is typical. A secretary types n different letters together with matching envelopes, drops the pile down the stairs, and then places the letters randomly in the envelopes. Each arrangement is equally likely, and we ask for the probability that exactly r are in their correct envelopes. Rather than using (2), we shall proceed directly by way of indicator functions. (Another approach is presented in Exercise (3.4.9).)

Solution. Let $L_1, L_2, \ldots, L_n$ denote the letters. Call a letter *good* if it is correctly addressed and *bad* otherwise; write X for the number of good letters. Let A_i be the event that L_i is good, and let I_i be the indicator function of A_i. Let $j_1, \ldots, j_r, k_{r+1}, \ldots, k_n$ be a permutation of the numbers $1, 2, \ldots, n$, and define

$$\textbf{(4)} \qquad S = \sum_\pi I_{j_1} \cdots I_{j_r} (1 - I_{k_{r+1}}) \cdots (1 - I_{k_n})$$

where the sum is taken over all such permutations π. Then

$$S = \begin{cases} 0 & \text{if } X \neq r, \\ r!\,(n-r)! & \text{if } X = r. \end{cases}$$

To see this, let $L_{i_1}, \ldots, L_{i_m}$ be the good letters. If $m \neq r$ then each summand in (4) equals 0. If $m = r$ then the summand in (4) equals 1 if and only if $j_1, \ldots, j_r$ is a permutation of $i_1, \ldots, i_r$ and $k_{r+1}, \ldots, k_n$ is a permutation of the remaining numbers; there are $r!\,(n-r)!$ such pairs of permutations. It follows that I, given by

$$I = \frac{1}{r!\,(n-r)!}S, \tag{5}$$

is the indicator function of the event $\{X = r\}$ that exactly r letters are good. We take expectations of (4) and multiply out to obtain

$$\mathbb{E}(S) = \sum_{\pi}\sum_{s=0}^{n-r}(-1)^s\binom{n-r}{s}\mathbb{E}(I_{j_1}\cdots I_{j_r}I_{k_{r+1}}\cdots I_{k_{r+s}})$$

by a symmetry argument. However,

$$\mathbb{E}(I_{j_1}\cdots I_{j_r}I_{k_{r+1}}\cdots I_{k_{r+s}}) = \frac{(n-r-s)!}{n!} \tag{6}$$

since there are $n!$ possible permutations, only $(n-r-s)!$ of which allocate $L_{i_1}, \ldots, L_{j_r}$, $L_{k_{r+1}}, \ldots, L_{k_{r+s}}$ to their correct envelopes. We combine (4), (5), and (6) to obtain

$$\begin{aligned}
\mathbb{P}(X = r) = \mathbb{E}(I) &= \frac{1}{r!\,(n-r)!}\mathbb{E}(S) \\
&= \frac{1}{r!\,(n-r)!}\sum_{s=0}^{n-r}(-1)^s\binom{n-r}{s}n!\frac{(n-r-s)!}{n!} \\
&= \frac{1}{r!}\sum_{s=0}^{n-r}(-1)^s\frac{1}{s!} \\
&= \frac{1}{r!}\left(\frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^{n-r}}{(n-r)!}\right) \quad \text{for } r \leq n-2 \text{ and } n \geq 2.
\end{aligned}$$

In particular, as the number n of letters tends to infinity, we obtain the possibly surprising result that the probability that no letter is put into its correct envelope approaches e^{-1}. It is left as an *exercise* to prove this without using indicators. ●

(7) Example. Reliability. When you telephone your friend in Cambridge, your call is routed through the telephone network in a way which depends on the current state of the traffic. For example, if all lines into the Ascot switchboard are in use, then your call may go through the switchboard at Newmarket. Sometimes you may fail to get through at all, owing to a combination of faulty and occupied equipment in the system. We may think of the network as comprising nodes joined by edges, drawn as 'graphs' in the manner of the examples of

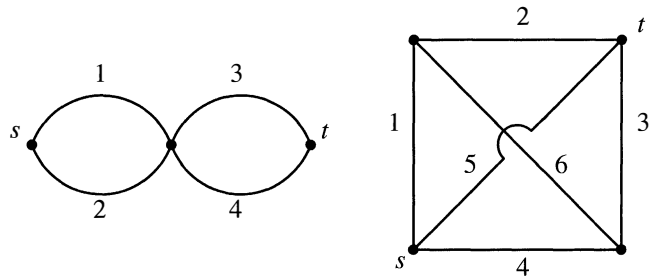

Figure 3.1. Two networks with source s and sink t.

Figure 3.1. In each of these examples there is a designated 'source' s and 'sink' t, and we wish to find a path through the network from s to t which uses available channels. As a simple model for such a system in the presence of uncertainty, we suppose that each edge e is 'working' with probability p_e, independently of all other edges. We write $\mathbf{p}$ for the vector of edge probabilities p_e, and define the *reliability* $R(\mathbf{p})$ of the network to be the probability that there is a path from s to t using only edges which are working. Denoting the network by G, we write $R_G(\mathbf{p})$ for $R(\mathbf{p})$ when we wish to emphasize the role of G.

We have encountered questions of reliability already. In Example (1.7.2) we were asked for the reliability of the first network in Figure 3.1 and in Problem (1.8.19) of the second, assuming on each occasion that the value of p_e does not depend on the choice of e.

Let us write

$$X_e = \begin{cases} 1 & \text{if edge } e \text{ is working,} \\ 0 & \text{otherwise,} \end{cases}$$

the indicator function of the event that e is working, so that X_e takes the values 0 and 1 with probabilities $1 - p_e$ and p_e respectively. Each realization X of the X_e either includes a working connection from s to t or does not. Thus, there exists a *structure function* ζ taking values 0 and 1 such that

$$\zeta(X) = \begin{cases} 1 & \text{if such a working connection exists,} \\ 0 & \text{otherwise;} \end{cases} \tag{8}$$

thus $\zeta(X)$ is the indicator function of the event that a working connection exists. It is immediately seen that $R(\mathbf{p}) = \mathbb{E}(\zeta(X))$. The function ζ may be expressed as

$$\zeta(X) = 1 - \prod_{\pi} I_{\{\pi \text{ not working}\}} = 1 - \prod_{\pi} \left(1 - \prod_{e \in \pi} X_e\right) \tag{9}$$

where π is a typical path in G from s to t, and we say that π is working if and only if every edge in π is working.

For instance, in the case of the first example of Figure 3.1, there are four different paths from s to t. Numbering the edges as indicated, we have that the structure function is given by

$$\zeta(X) = 1 - (1 - X_1X_3)(1 - X_1X_4)(1 - X_2X_3)(1 - X_2X_4). \tag{10}$$

As an *exercise*, expand this and take expectations to calculate the reliability of the network when $p_e = p$ for all edges e. ●

(11) Example. The probabilistic method†. Probability may be used to derive non-trivial results not involving probability. Here is an example. There are 17 fenceposts around the perimeter of a field, exactly 5 of which are rotten. Show that, irrespective of which these 5 are, there necessarily exists a run of 7 consecutive posts at least 3 of which are rotten.

Our solution involves probability. We label the posts $1, 2, \ldots, 17$, and let I_k be the indicator function that post k is rotten. Let R_k be the number of rotten posts amongst those labelled $k+1, k+2, \ldots, k+7$, all taken modulo 17. We now pick a random post labelled K, each being equally likely. We have that

$$\mathbb{E}(R_K) = \sum_{k=1}^{17} \frac{1}{17}(I_{k+1} + I_{k+2} + \cdots + I_{k+7}) = \sum_{j=1}^{17} \frac{7}{17} I_j = \frac{7}{17} \cdot 5.$$

Now $\frac{35}{17} > 2$, implying that $\mathbb{P}(R_K > 2) > 0$. Since R_K is integer valued, it must be the case that $\mathbb{P}(R_K \geq 3) > 0$, implying that $R_k \geq 3$ for some k. ●

Exercises for Section 3.4

1. A biased coin is tossed n times, and heads shows with probability p on each toss. A *run* is a sequence of throws which result in the same outcome, so that, for example, the sequence HHTHTTH contains five runs. Show that the expected number of runs is $1 + 2(n-1)p(1-p)$. Find the variance of the number of runs.

2. An urn contains n balls numbered $1, 2, \ldots, n$. We remove k balls at random (without replacement) and add up their numbers. Find the mean and variance of the total.

3. Of the $2n$ people in a given collection of n couples, exactly m die. Assuming that the m have been picked at random, find the mean number of surviving couples. This problem was formulated by Daniel Bernoulli in 1768.

4. Urn R contains n red balls and urn B contains n blue balls. At each stage, a ball is selected at random from each urn, and they are swapped. Show that the mean number of red balls in urn R after stage k is $\frac{1}{2}n\{1 + (1 - 2/n)^k\}$. This 'diffusion model' was described by Daniel Bernoulli in 1769.

5. Consider a square with diagonals, with distinct source and sink. Each edge represents a component which is working correctly with probability p, independently of all other components. Write down an expression for the Boolean function which equals 1 if and only if there is a working path from source to sink, in terms of the indicator functions X_i of the events {edge i is working} as i runs over the set of edges. Hence calculate the reliability of the network.

6. A system is called a 'k out of n' system if it contains n components and it works whenever k or more of these components are working. Suppose that each component is working with probability p, independently of the other components, and let X_c be the indicator function of the event that component c is working. Find, in terms of the X_c, the indicator function of the event that the system works, and deduce the reliability of the system.

7. The probabilistic method. Let $G = (V, E)$ be a finite graph. For any set W of vertices and any edge $e \in E$, define the indicator function

$$I_W(e) = \begin{cases} 1 & \text{if } e \text{ connects } W \text{ and } W^c, \\ 0 & \text{otherwise.} \end{cases}$$

Set $N_W = \sum_{e \in E} I_W(e)$. Show that there exists $W \subseteq V$ such that $N_W \geq \frac{1}{2}|E|$.

†Generally credited to Erdős.

8. A total of n bar magnets are placed end to end in a line with random independent orientations. Adjacent like poles repel, ends with opposite polarities join to form blocks. Let X be the number of blocks of joined magnets. Find $\mathbb{E}(X)$ and $\text{var}(X)$.

9. Matching. (a) Use the inclusion–exclusion formula (3.4.2) to derive the result of Example (3.4.3), namely: in a random permutation of the first n integers, the probability that exactly r retain their original positions is

$$\frac{1}{r!}\left(\frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^{n-r}}{(n-r)!}\right).$$

(b) Let d_n be the number of derangements of the first n integers (that is, rearrangements with no integers in their original positions). Show that $d_{n+1} = nd_n + nd_{n-1}$ for $n \geq 2$. Deduce the result of part (a).

3.5 Examples of discrete variables

(1) Bernoulli trials. A random variable X takes values 1 and 0 with probabilities p and q $(= 1 - p)$, respectively. Sometimes we think of these values as representing the 'success' or the 'failure' of a trial. The mass function is

$$f(0) = 1 - p, \qquad f(1) = p,$$

and it follows that $\mathbb{E}X = p$ and $\text{var}(X) = p(1 - p)$. ●

(2) Binomial distribution. We perform n independent Bernoulli trials $X_1, X_2, \ldots, X_n$ and count the total number of successes $Y = X_1 + X_2 + \cdots + X_n$. As in Example (3.1.3), the mass function of Y is

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

Application of Theorems (3.3.8) and (3.3.11) yields immediately

$$\mathbb{E}Y = np, \qquad \text{var}(Y) = np(1 - p);$$

the method of Example (3.3.7) provides a more lengthy derivation of this. ●

(3) Trinomial distribution. More generally, suppose we conduct n trials, each of which results in one of three outcomes (red, white, or blue, say), where red occurs with probability p, white with probability q, and blue with probability $1 - p - q$. The probability of r reds, w whites, and $n - r - w$ blues is

$$\frac{n!}{r!\, w!\, (n-r-w)!} p^r q^w (1 - p - q)^{n-r-w}.$$

This is the *trinomial distribution*, with parameters n, p, and q. The 'multinomial distribution' is the obvious generalization of this distribution to the case of some number, say t, of possible outcomes. ●

(4) Poisson distribution. A *Poisson* variable is a random variable with the Poisson mass function

$$f(k) = \frac{\lambda^k}{k!}e^{-\lambda}, \qquad k = 0, 1, 2, \ldots$$

for some $\lambda > 0$. It can be obtained in practice in the following way. Let Y be a bin(n, p) variable, and suppose that n is very large and p is very small (an example might be the number Y of misprints on the front page of the *Grauniad*, where n is the total number of characters and p is the probability for each character that the typesetter has made an error). Now, let $n \to \infty$ and $p \to 0$ in such a way that $\mathbb{E}(Y) = np$ approaches a non-zero constant λ. Then, for $k = 0, 1, 2, \ldots$,

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{k!}\left(\frac{np}{1-p}\right)^k (1-p)^n \to \frac{\lambda^k}{k!}e^{-\lambda}.$$

Check that both the mean and the variance of this distribution are equal to λ. Now do Problem (2.7.7) again (*exercise*). ●

(5) Geometric distribution. A *geometric* variable is a random variable with the geometric mass function

$$f(k) = p(1-p)^{k-1}, \qquad k = 1, 2, \ldots$$

for some number p in $(0, 1)$. This distribution arises in the following way. Suppose that independent Bernoulli trials (parameter p) are performed at times $1, 2, \ldots$. Let W be the time which elapses before the first success; W is called a *waiting time*. Then $\mathbb{P}(W > k) = (1-p)^k$ and thus

$$\mathbb{P}(W = k) = \mathbb{P}(W > k-1) - \mathbb{P}(W > k) = p(1-p)^{k-1}.$$

The reader should check, preferably at this point, that the mean and variance are p^{-1} and $(1-p)p^{-2}$ respectively. ●

(6) Negative binomial distribution. More generally, in the previous example, let W_r be the waiting time for the rth success. Check that W_r has mass function

$$\mathbb{P}(W_r = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \qquad k = r, r+1, \ldots;$$

it is said to have the *negative binomial distribution* with parameters r and p. The random variable W_r is the sum of r independent geometric variables. To see this, let X_1 be the waiting time for the first success, X_2 the *further* waiting time for the second success, X_3 the *further* waiting time for the third success, and so on. Then $X_1, X_2, \ldots$ are independent and geometric, and

$$W_r = X_1 + X_2 + \cdots + X_r.$$

Apply Theorems (3.3.8) and (3.3.11) to find the mean and the variance of W_r. ●

Exercises for Section 3.5

1. De Moivre trials. Each trial may result in any of t given outcomes, the ith outcome having probability p_i. Let N_i be the number of occurrences of the ith outcome in n independent trials. Show that

$$\mathbb{P}(N_i = n_i \text{ for } 1 \leq i \leq t) = \frac{n!}{n_1! n_2! \cdots n_t!} p_1^{n_1} p_2^{n_2} \cdots p_t^{n_t}$$

for any collection $n_1, n_2, \ldots, n_t$ of non-negative integers with sum n. The vector N is said to have the *multinomial distribution.*

2. In your pocket is a random number N of coins, where N has the Poisson distribution with parameter λ. You toss each coin once, with heads showing with probability p each time. Show that the total number of heads has the Poisson distribution with parameter λp.

3. Let X be Poisson distributed where $\mathbb{P}(X = n) = p_n(\lambda) = \lambda^n e^{-\lambda}/n!$ for $n \geq 0$. Show that $\mathbb{P}(X \leq n) = 1 - \int_0^\lambda p_n(x)\, dx$.

4. Capture–recapture. A population of b animals has had a number a of its members captured, marked, and released. Let X be the number of animals it is necessary to recapture (without re-release) in order to obtain m marked animals. Show that

$$\mathbb{P}(X = n) = \frac{a}{b} \binom{a-1}{m-1} \binom{b-a}{n-m} \Big/ \binom{b-1}{n-1},$$

and find $\mathbb{E}X$. This distribution has been called *negative hypergeometric.*

3.6 Dependence

Probability theory is largely concerned with families of random variables; these families will not in general consist entirely of independent variables.

(1) Example. Suppose that we back three horses to win as an accumulator. If our stake is £1 and the starting prices are α, β, and γ, then our total profit is

$$W = (\alpha + 1)(\beta + 1)(\gamma + 1) I_1 I_2 I_3 - 1$$

where I_i denotes the indicator of a win in the ith race by our horse. (In checking this expression remember that a bet of £B on a horse with starting price α brings a return of £$B(\alpha + 1)$, should this horse win.) We lose £1 if some backed horse fails to win. It seems clear that the random variables W and I_1 are *not* independent. If the races are run independently, then

$$\mathbb{P}(W = -1) = \mathbb{P}(I_1 I_2 I_3 = 0),$$

but

$$\mathbb{P}(W = -1 \mid I_1 = 1) = \mathbb{P}(I_2 I_3 = 0)$$

which are different from each other unless the first backed horse is guaranteed victory. ●

We require a tool for studying collections of dependent variables. Knowledge of their individual mass functions is little help by itself. Just as the main tools for studying a random

variable is its distribution function, so the study of, say, a pair of random variables is based on its 'joint' distribution function and mass function.

(2) Definition. The **joint distribution function** $F : \mathbb{R}^2 \to [0, 1]$ of X and Y, where X and Y are discrete variables, is given by

$$F(x, y) = \mathbb{P}(X \leq x \text{ and } Y \leq y).$$

Their **joint mass function** $f : \mathbb{R}^2 \to [0, 1]$ is given by

$$f(x, y) = \mathbb{P}(X = x \text{ and } Y = y).$$

Joint distribution functions and joint mass functions of larger collections of variables are defined similarly. The functions F and f can be characterized in much the same way (Lemmas (2.1.6) and (3.1.2)) as the corresponding functions of a single variable.We omit the details. We write $F_{X,Y}$ and $f_{X,Y}$ when we need to stress the role of X and Y. You may think of the joint mass function in the following way. If $A_x = \{X = x\}$ and $B_y = \{Y = y\}$, then

$$f(x, y) = \mathbb{P}(A_x \cap B_y).$$

The definition of independence can now be reformulated in a lemma.

(3) Lemma. *The discrete random variables X and Y are independent if and only if*

(4) $$f_{X,Y}(x, y) = f_X(x) f_Y(y) \quad \textit{for all } x, y \in \mathbb{R}.$$

More generally, X and Y are independent if and only if $f_{X,Y}(x, y)$ can be factorized as the product $g(x)h(y)$ of a function of x alone and a function of y alone.

Proof. This is Problem (3.11.1). ■

Suppose that X and Y have joint mass function $f_{X,Y}$ and we wish to check whether or not (4) holds. First we need to calculate the *marginal mass functions* f_X and f_Y from our knowledge of $f_{X,Y}$. These are found in the following way:

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P}\Big(\bigcup_y (\{X = x\} \cap \{Y = y\})\Big)$$
$$= \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f_{X,Y}(x, y),$$

and similarly $f_Y(y) = \sum_x f_{X,Y}(x, y)$. Having found the marginals, it is a trivial matter to see whether (4) holds or not.

Remark. We stress that the factorization (4) must hold for *all* x and y in order that X and Y be independent.

(5) Example. Calculation of marginals. In Example (3.2.2) we encountered a pair X, Y of variables with a joint mass function

$$f(x, y) = \frac{\alpha^x \beta^y}{x!\, y!} e^{-\alpha-\beta} \quad \text{for} \quad x, y = 0, 1, 2, \ldots$$

where $\alpha, \beta > 0$. The marginal mass function of X is

$$f_X(x) = \sum_y f(x, y) = \frac{\alpha^x}{x!} e^{-\alpha} \sum_{y=0}^{\infty} \frac{\beta^y}{y!} e^{-\beta} = \frac{\alpha^x}{x!} e^{-\alpha}$$

and so X has the Poisson distribution with parameter α. Similarly Y has the Poisson distribution with parameter β. It is easy to check that (4) holds, whence X and Y are independent. ●

For any discrete pair X, Y, a real function $g(X, Y)$ is a random variable. We shall often need to find its expectation. To avoid explicit calculation of its mass function, we shall use the following more general form of the law of the unconscious statistician, Lemma (3.3.3).

(6) Lemma. $\mathbb{E}(g(X, Y)) = \sum_{x,y} g(x, y) f_{X,Y}(x, y)$.

Proof. As for Lemma (3.3.3). ■

For example, $\mathbb{E}(XY) = \sum_{x,y} xy f_{X,Y}(x, y)$. This formula is particularly useful to statisticians who may need to find simple ways of explaining dependence to laymen. For instance, suppose that the government wishes to announce that the dependence between defence spending and the cost of living is very small. It should *not* publish an estimate of the joint mass function unless its object is obfuscation alone. Most members of the public would prefer to find that this dependence can be represented in terms of a single number on a prescribed scale. Towards this end we make the following definition†.

(7) Definition. The **covariance** of X and Y is

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

The **correlation (coefficient)** of X and Y is

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$$

as long as the variances are non-zero.

Note that the concept of covariance generalizes that of variance in that $\operatorname{cov}(X, X) = \operatorname{var}(X)$. Expanding the covariance gives

$$\operatorname{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Remember, Definition (3.3.10), that X and Y are called *uncorrelated* if $\operatorname{cov}(X, Y) = 0$. Also, independent variables are always uncorrelated, although the converse is not true. Covariance itself is not a satisfactory measure of dependence because the scale of values which $\operatorname{cov}(X, Y)$ may take contains no points which are clearly interpretable in terms of the relationship between X and Y. The following lemma shows that this is not the case for correlations.

(8) Lemma. *The correlation coefficient ρ satisfies $|\rho(X, Y)| \leq 1$ with equality if and only if $\mathbb{P}(aX + bY = c) = 1$ for some $a, b, c \in \mathbb{R}$.*

†The concepts and terminology in this definition were formulated by Francis Galton in the late 1880s.

The proof is an application of the following important inequality.

(9) Theorem. Cauchy–Schwarz inequality. *For random variables X and Y,*

$$\{\mathbb{E}(XY)\}^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

with equality if and only if $\mathbb{P}(aX = bY) = 1$ for some real a and b, at least one of which is non-zero.

Proof. We can assume that $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2)$ are strictly positive, since otherwise the result follows immediately from Problem (3.11.2). For $a, b \in \mathbb{R}$, let $Z = aX - bY$. Then

$$0 \leq \mathbb{E}(Z^2) = a^2\mathbb{E}(X^2) - 2ab\mathbb{E}(XY) + b^2\mathbb{E}(Y^2).$$

Thus the right-hand side is a quadratic in the variable a with at most one real root. Its discriminant must be non-positive. That is to say, if $b \neq 0$,

$$\mathbb{E}(XY)^2 - \mathbb{E}(X^2)\mathbb{E}(Y^2) \leq 0.$$

The discriminant is zero if and only if the quadratic has a real root. This occurs if and only if

$$\mathbb{E}\big((aX - bY)^2\big) = 0 \quad \text{for some } a \text{ and } b,$$

which, by Problem (3.11.2), completes the proof. ■

Proof of (8). Apply (9) to the variables $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$. ■

A more careful treatment than this proof shows that $\rho = +1$ if and only if Y *increases* linearly with X and $\rho = -1$ if and only if Y *decreases* linearly as X increases.

(10) Example. Here is a tedious numerical example of the use of joint mass functions. Let X and Y take values in $\{1, 2, 3\}$ and $\{-1, 0, 2\}$ respectively, with joint mass function f where $f(x, y)$ is the appropriate entry in Table 3.1.

	$y = -1$	$y = 0$	$y = 2$	f_X
$x = 1$	$\frac{1}{18}$	$\frac{3}{18}$	$\frac{2}{18}$	$\frac{6}{18}$
$x = 2$	$\frac{2}{18}$	0	$\frac{3}{18}$	$\frac{5}{18}$
$x = 3$	0	$\frac{4}{18}$	$\frac{3}{18}$	$\frac{7}{18}$
f_Y	$\frac{3}{18}$	$\frac{7}{18}$	$\frac{8}{18}$	

Table 3.1. The joint mass function of the random variables X and Y. The indicated row and column sums are the marginal mass functions f_X and f_Y.

A quick calculation gives

$$\mathbb{E}(XY) = \sum_{x,y} xyf(x, y) = 29/18,$$

$$\mathbb{E}(X) = \sum_{x} xf_X(x) = 37/18, \quad \mathbb{E}(Y) = 13/18,$$

$$\operatorname{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 233/324, \quad \operatorname{var}(Y) = 461/324,$$

$$\operatorname{cov}(X, Y) = 41/324, \quad \rho(X, Y) = 41/\sqrt{107413}.$$

●

Exercises for Section 3.6

1. Show that the collection of random variables on a given probability space and having finite variance forms a vector space over the reals.

2. Find the marginal mass functions of the multinomial distribution of Exercise (3.5.1).

3. Let X and Y be discrete random variables with joint mass function

$$f(x, y) = \frac{C}{(x + y - 1)(x + y)(x + y + 1)}, \qquad x, y = 1, 2, 3, \ldots.$$

Find the marginal mass functions of X and Y, calculate C, and also the covariance of X and Y.

4. Let X and Y be discrete random variables with mean 0, variance 1, and covariance ρ. Show that $\mathbb{E}(\max\{X^2, Y^2\}) \leq 1 + \sqrt{1 - \rho^2}$.

5. Mutual information. Let X and Y be discrete random variables with joint mass function f.

(a) Show that $\mathbb{E}(\log f_X(X)) \geq \mathbb{E}(\log f_Y(X))$.

(b) Show that the *mutual information*

$$I = \mathbb{E}\left(\log\left\{\frac{f(X, Y)}{f_X(X) f_Y(Y)}\right\}\right)$$

satisfies $I \geq 0$, with equality if and only if X and Y are independent.

6. Voter paradox. Let X, Y, Z be discrete random variables with the property that their values are distinct with probability 1. Let $a = \mathbb{P}(X > Y)$, $b = \mathbb{P}(Y > Z)$, $c = \mathbb{P}(Z > X)$.

(a) Show that $\min\{a, b, c\} \leq \frac{2}{3}$, and give an example where this bound is attained.

(b) Show that, if X, Y, Z are independent and identically distributed, then $a = b = c = \frac{1}{2}$.

(c) Find $\min\{a, b, c\}$ and $\sup_p \min\{a, b, c\}$ when $\mathbb{P}(X = 0) = 1$, and Y, Z are independent with $\mathbb{P}(Z = 1) = \mathbb{P}(Y = -1) = p$, $\mathbb{P}(Z = -2) = \mathbb{P}(Y = 2) = 1 - p$. Here, $\sup_p$ denotes the supremum as p varies over $[0, 1]$.

[Part (a) is related to the observation that, in an election, it is possible for more than half of the voters to prefer candidate A to candidate B, more than half B to C, and more than half C to A.]

7. Benford's distribution, or the law of anomalous numbers. If one picks a numerical entry at random from an almanac, or the annual accounts of a corporation, the first two significant digits, X, Y, are found to have approximately the joint mass function

$$f(x, y) = \log_{10}\left(1 + \frac{1}{10x + y}\right), \quad 1 \leq x \leq 9,\ 0 \leq y \leq 9.$$

Find the mass function of X and an approximation to its mean. [A heuristic explanation for this phenomenon may be found in the second of Feller's volumes (1971).]

8. Let X and Y have joint mass function

$$f(j, k) = \frac{c(j + k)a^{j+k}}{j!\,k!}, \quad j, k \geq 0,$$

where a is a constant. Find c, $\mathbb{P}(X = j)$, $\mathbb{P}(X + Y = r)$, and $\mathbb{E}(X)$.

3.7 Conditional distributions and conditional expectation

In Section 1.4 we discussed the conditional probability $\mathbb{P}(B \mid A)$. This may be set in the more general context of the conditional distribution of one variable Y given the value of another variable X; this reduces to the definition of the conditional probabilities of events A and B if $X = I_A$ and $Y = I_B$.

Let X and Y be two discrete variables on $(\Omega, \mathcal{F}, \mathbb{P})$.

(1) Definition. The **conditional distribution function** of Y given $X = x$, written $F_{Y|X}(\cdot \mid x)$, is defined by

$$F_{Y|X}(y \mid x) = \mathbb{P}(Y \leq y \mid X = x)$$

for any x such that $\mathbb{P}(X = x) > 0$. The **conditional (probability) mass function** of Y given $X = x$, written $f_{Y|X}(\cdot \mid x)$, is defined by

$$f_{Y|X}(y \mid x) = \mathbb{P}(Y = y \mid X = x) \tag{2}$$

for any x such that $\mathbb{P}(X = x) > 0$.

Formula (2) is easy to remember as $f_{Y|X} = f_{X,Y}/f_X$. Conditional distributions and mass functions are undefined at values of x for which $\mathbb{P}(X = x) = 0$. Clearly X and Y are independent if and only if $f_{Y|X} = f_Y$.

Suppose we are told that $X = x$. Conditional upon this, the new distribution of Y has mass function $f_{Y|X}(y \mid x)$, which we think of as a function of y. The expected value of this distribution, $\sum_y y f_{Y|X}(y \mid x)$, is called the *conditional expectation* of Y given $X = x$ and is written $\psi(x) = \mathbb{E}(Y \mid X = x)$. Now, we observe that the conditional expectation depends on the value x taken by X, and can be thought of as a function $\psi(X)$ of X itself.

(3) Definition. Let $\psi(x) = \mathbb{E}(Y \mid X = x)$. Then $\psi(X)$ is called the **conditional expectation** of Y given X, written as $\mathbb{E}(Y \mid X)$.

Although 'conditional expectation' sounds like a number, it is actually a random variable. It has the following important property.

(4) Theorem. *The conditional expectation $\psi(X) = \mathbb{E}(Y \mid X)$ satisfies*

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

Proof. By Lemma (3.3.3),

$$\begin{aligned}\mathbb{E}(\psi(X)) &= \sum_x \psi(x) f_X(x) = \sum_{x,y} y f_{Y|X}(y \mid x) f_X(x) \\ &= \sum_{x,y} y f_{X,Y}(x, y) = \sum_y y f_Y(y) = \mathbb{E}(Y).\end{aligned}$$

■

This is an extremely useful theorem, to which we shall make repeated reference. It often provides a useful method for calculating $\mathbb{E}(Y)$, since it asserts that

$$\mathbb{E}(Y) = \sum_x \mathbb{E}(Y \mid X = x)\mathbb{P}(X = x).$$

(5) **Example.** A hen lays N eggs, where N has the Poisson distribution with parameter λ. Each egg hatches with probability p $(= 1 - q)$ independently of the other eggs. Let K be the number of chicks. Find $\mathbb{E}(K \mid N)$, $\mathbb{E}(K)$, and $\mathbb{E}(N \mid K)$.

Solution. We are given that

$$f_N(n) = \frac{\lambda^n}{n!}e^{-\lambda}, \qquad f_{K|N}(k \mid n) = \binom{n}{k}p^k(1-p)^{n-k}.$$

Therefore

$$\psi(n) = \mathbb{E}(K \mid N = n) = \sum_k k f_{K|N}(k \mid n) = pn.$$

Thus $\mathbb{E}(K \mid N) = \psi(N) = pN$ and

$$\mathbb{E}(K) = \mathbb{E}(\psi(N)) = p\mathbb{E}(N) = p\lambda.$$

To find $\mathbb{E}(N \mid K)$ we need to know the conditional mass function $f_{N|K}$ of N given K. However,

$$\begin{aligned} f_{N|K}(n \mid k) &= \mathbb{P}(N = n \mid K = k) \\ &= \frac{\mathbb{P}(K = k \mid N = n)\mathbb{P}(N = n)}{\mathbb{P}(K = k)} \\ &= \frac{\binom{n}{k}p^k(1-p)^{n-k}(\lambda^n/n!)e^{-\lambda}}{\sum_{m \geq k}\binom{m}{k}p^k(1-p)^{m-k}(\lambda^m/m!)e^{-\lambda}} \quad \text{if } n \geq k \\ &= \frac{(q\lambda)^{n-k}}{(n-k)!}e^{-q\lambda}. \end{aligned}$$

Hence

$$\mathbb{E}(N \mid K = k) = \sum_{n \geq k} n\frac{(q\lambda)^{n-k}}{(n-k)!}e^{-q\lambda} = k + q\lambda,$$

giving $\mathbb{E}(N \mid K) = K + q\lambda$. ●

There is a more general version of Theorem (4), and this will be of interest later.

(6) **Theorem.** *The conditional expectation $\psi(X) = \mathbb{E}(Y \mid X)$ satisfies*

$$\mathbb{E}\big(\psi(X)g(X)\big) = \mathbb{E}\big(Yg(X)\big) \tag{7}$$

for any function g for which both expectations exist.

Setting $g(x) = 1$ for all x, we obtain the result of (4). Whilst Theorem (6) is useful in its own right, we shall see later that its principal interest lies elsewhere. The conclusion of the theorem may be taken as a *definition* of conditional expectation—as a function $\psi(X)$ of X such that (7) holds for all suitable functions g. Such a definition is convenient when working with a notion of conditional expectation more general than that dealt with here.

Proof. As in the proof of (4),

$$\begin{aligned} \mathbb{E}\big(\psi(X)g(X)\big) &= \sum_x \psi(x)g(x)f_X(x) = \sum_{x,y} yg(x)f_{Y|X}(y \mid x)f_X(x) \\ &= \sum_{x,y} yg(x)f_{X,Y}(x, y) = \mathbb{E}\big(Yg(X)\big). \end{aligned}$$

■

Exercises for Section 3.7

1. Show the following:
(a) $\mathbb{E}(aY + bZ \mid X) = a\mathbb{E}(Y \mid X) + b\mathbb{E}(Z \mid X)$ for $a, b \in \mathbb{R}$,
(b) $\mathbb{E}(Y \mid X) \geq 0$ if $Y \geq 0$,
(c) $\mathbb{E}(1 \mid X) = 1$,
(d) if X and Y are independent then $\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$,
(e) ('pull-through property') $\mathbb{E}(Yg(X) \mid X) = g(X)\mathbb{E}(Y \mid X)$ for any suitable function g,
(f) ('tower property') $\mathbb{E}\{\mathbb{E}(Y \mid X, Z) \mid X\} = \mathbb{E}(Y \mid X) = \mathbb{E}\{\mathbb{E}(Y \mid X) \mid X, Z\}$.

2. Uniqueness of conditional expectation. Suppose that X and Y are discrete random variables, and that $\phi(X)$ and $\psi(X)$ are two functions of X satisfying

$$\mathbb{E}\big(\phi(X)g(X)\big) = \mathbb{E}\big(\psi(X)g(X)\big) = \mathbb{E}\big(Yg(X)\big)$$

for any function g for which all the expectations exist. Show that $\phi(X)$ and $\psi(X)$ are almost surely equal, in that $\mathbb{P}(\phi(X) = \psi(X)) = 1$.

3. Suppose that the conditional expectation of Y given X is defined as the (almost surely) unique function $\psi(X)$ such that $\mathbb{E}(\psi(X)g(X)) = \mathbb{E}(Yg(X))$ for all functions g for which the expectations exist. Show (a)–(f) of Exercise (1) above (with the occasional addition of the expression 'with probability 1').

4. How should we define $\operatorname{var}(Y \mid X)$, the conditional variance of Y given X? Show that $\operatorname{var}(Y) = \mathbb{E}(\operatorname{var}(Y \mid X)) + \operatorname{var}(\mathbb{E}(Y \mid X))$.

5. The lifetime of a machine (in days) is a random variable T with mass function f. Given that the machine is working after t days, what is the mean subsequent lifetime of the machine when:
(a) $f(x) = (N+1)^{-1}$ for $x \in \{0, 1, \ldots, N\}$,
(b) $f(x) = 2^{-x}$ for $x = 1, 2, \ldots$.
(The first part of Problem (3.11.13) may be useful.)

6. Let $X_1, X_2, \ldots$ be identically distributed random variables with mean μ, and let N be a random variable taking values in the non-negative integers and independent of the X_i. Let $S = X_1 + X_2 + \cdots + X_N$. Show that $\mathbb{E}(S \mid N) = \mu N$, and deduce that $\mathbb{E}(S) = \mu\mathbb{E}(N)$.

7. A factory has produced n robots, each of which is faulty with probability ϕ. To each robot a test is applied which detects the fault (if present) with probability δ. Let X be the number of faulty robots, and Y the number detected as faulty. Assuming the usual independence, show that

$$\mathbb{E}(X \mid Y) = \big\{n\phi(1-\delta) + (1-\phi)Y\big\}/(1-\phi\delta).$$

8. Families. Each child is equally likely to be male or female, independently of all other children.
(a) Show that, in a family of predetermined size, the expected number of boys equals the expected number of girls. Was the assumption of independence necessary?
(b) A randomly selected child is male; does the expected number of his brothers equal the expected number of his sisters? What happens if you do not require independence?

9. Let X and Y be independent with mean μ. Explain the error in the following equation:

$$\text{'}\mathbb{E}(X \mid X + Y = z) = \mathbb{E}(X \mid X = z - Y) = \mathbb{E}(z - Y) = z - \mu\text{'}.$$

10. A coin shows heads with probability p. Let X_n be the number of flips required to obtain a run of n consecutive heads. Show that $\mathbb{E}(X_n) = \sum_{k=1}^{n} p^{-k}$.

3.8 Sums of random variables

Much of the classical theory of probability concerns sums of random variables. We have seen already many such sums; the number of heads in n tosses of a coin is one of the simplest such examples, but we shall encounter many situations which are more complicated than this. One particular complication is when the summands are dependent. The first stage in developing a systematic technique is to find a formula for describing the mass function of the sum $Z = X + Y$ of two variables having joint mass function $f(x, y)$.

(1) **Theorem.** *We have that* $\mathbb{P}(X + Y = z) = \sum_x f(x, z - x)$.

Proof. The union

$$\{X + Y = z\} = \bigcup_x \big(\{X = x\} \cap \{Y = z - x\}\big)$$

is disjoint, and at most countably many of its contributions have non-zero probability. Therefore

$$\mathbb{P}(X + Y = z) = \sum_x \mathbb{P}(X = x,\ Y = z - x) = \sum_x f(x, z - x). \qquad \blacksquare$$

If X and Y are independent, then

$$\mathbb{P}(X + Y = z) = f_{X+Y}(z) = \sum_x f_X(x) f_Y(z - x) = \sum_y f_X(z - y) f_Y(y).$$

The mass function of $X + Y$ is called the *convolution* of the mass functions of X and Y, and is written

(2) $$f_{X+Y} = f_X * f_Y.$$

(3) **Example (3.5.6) revisited.** Let X_1 and X_2 be independent geometric variables with common mass function

$$f(k) = p(1 - p)^{k-1}, \qquad k = 1, 2, \ldots.$$

By (2), $Z = X_1 + X_2$ has mass function

$$\begin{aligned}
\mathbb{P}(Z = z) &= \sum_k \mathbb{P}(X_1 = k)\mathbb{P}(X_2 = z - k) \\
&= \sum_{k=1}^{z-1} p(1 - p)^{k-1} p(1 - p)^{z-k-1} \\
&= (z - 1)p^2(1 - p)^{z-2}, \qquad z = 2, 3, \ldots
\end{aligned}$$

in agreement with Example (3.5.6). The general formula for the sum of a number, r say, of geometric variables can easily be verified by induction. ●

Exercises for Section 3.8

1. Let X and Y be independent variables, X being equally likely to take any value in $\{0, 1, \ldots, m\}$, and Y similarly in $\{0, 1, \ldots, n\}$. Find the mass function of $Z = X + Y$. The random variable Z is said to have the *trapezoidal distribution.*

2. Let X and Y have the joint mass function

$$f(x, y) = \frac{C}{(x + y - 1)(x + y)(x + y + 1)}, \qquad x, y = 1, 2, 3, \ldots.$$

Find the mass functions of $U = X + Y$ and $V = X - Y$.

3. Let X and Y be independent geometric random variables with respective parameters α and β. Show that

$$\mathbb{P}(X + Y = z) = \frac{\alpha\beta}{\alpha - \beta}\{(1 - \beta)^{z-1} - (1 - \alpha)^{z-1}\}.$$

4. Let $\{X_r : 1 \leq r \leq n\}$ be independent geometric random variables with parameter p. Show that $Z = \sum_{r=1}^{n} X_r$ has a negative binomial distribution. [Hint: No calculations are necessary.]

5. Pepys's problem†. Sam rolls $6n$ dice once; he needs at least n sixes. Isaac rolls $6(n + 1)$ dice; he needs at least $n + 1$ sixes. Who is more likely to obtain the number of sixes he needs?

6. Let N be Poisson distributed with parameter λ. Show that, for any function g such that the expectations exist, $\mathbb{E}(Ng(N)) = \lambda\mathbb{E}g(N + 1)$. More generally, if $S = \sum_{r=1}^{N} X_r$, where $\{X_r : r \geq 0\}$ are independent identically distributed non-negative integer-valued random variables, show that

$$\mathbb{E}\big(Sg(S)\big) = \lambda\mathbb{E}\big(g(S + X_0)X_0\big).$$

3.9 Simple random walk

Until now we have dealt largely with general theory; the final two sections of this chapter may provide some lighter relief. One of the simplest random processes is so-called 'simple random walk'‡; this process arises in many ways, of which the following is traditional. A gambler G plays the following game at the casino. The croupier tosses a (possibly biased) coin repeatedly; each time heads appears, he gives G one franc, and each time tails appears he takes one franc from G. Writing S_n for G's fortune after n tosses of the coin, we have that $S_{n+1} = S_n + X_{n+1}$ where X_{n+1} is a random variable taking the value 1 with some fixed probability p and -1 otherwise; furthermore, X_{n+1} is assumed independent of the results of all previous tosses. Thus

$$S_n = S_0 + \sum_{i=1}^{n} X_i, \tag{1}$$

†Pepys put a simple version of this problem to Newton in 1693, but was reluctant to accept the correct reply he received.

‡Karl Pearson coined the term 'random walk' in 1906, and (using a result of Rayleigh) demonstrated the theorem that "the most likely place to find a drunken walker is somewhere near his starting point", empirical verification of which is not hard to find.

so that S_n is obtained from the initial fortune S_0 by the addition of n independent random variables. We are assuming here that there are no constraints on G's fortune imposed externally, such as that the game is terminated if his fortune is reduced to zero.

An alternative picture of 'simple random walk' involves the motion of a particle—a particle which inhabits the set of integers and which moves at each step either one step to the right, with probability p, or one step to the left, the directions of different steps being independent of each other. More complicated random walks arise when the steps of the particle are allowed to have some general distribution on the integers, or the reals, so that the position S_n at time n is given by (1) where the X_i are independent and identically distributed random variables having some specified distribution function. Even greater generality is obtained by assuming that the X_i take values in $\mathbb{R}^d$ for some $d \geq 1$, or even some vector space over the real numbers. Random walks may be used with some success in modelling various practical situations, such as the numbers of cars in a toll queue at 5 minute intervals, the position of a pollen grain suspended in fluid at 1 second intervals, or the value of the Dow–Jones index each Monday morning. In each case, it may not be too bad a guess that the $(n+1)$th reading differs from the nth by a random quantity which is independent of previous jumps but has the same probability distribution. The theory of random walks is a basic tool in the probabilist's kit, and we shall concern ourselves here with 'simple random walk' only.

At any instant of time a particle inhabits one of the integer points of the real line. At time 0 it starts from some specified point, and at each subsequent epoch of time $1, 2, \ldots$ it moves from its current position to a new position according to the following law. With probability p it moves one step to the right, and with probability $q = 1 - p$ it moves one step to the left; moves are independent of each other. The walk is called *symmetric* if $p = q = \frac{1}{2}$. Example (1.7.4) concerned a symmetric random walk with 'absorbing' barriers at the points 0 and N. In general, let S_n denote the position of the particle after n moves, and set $S_0 = a$. Then

$$S_n = a + \sum_{i=1}^{n} X_i \tag{2}$$

where $X_1, X_2, \ldots$ is a sequence of independent Bernoulli variables taking values $+1$ and -1 (rather than $+1$ and 0 as before) with probabilities p and q.

We record the motion of the particle as the sequence $\{(n, S_n) : n \geq 0\}$ of Cartesian coordinates of points in the plane. This collection of points, joined by solid lines between neighbours, is called the *path* of the particle. In the example shown in Figure 3.2, the particle has visited the points $0, 1, 0, -1, 0, 1, 2$ in succession. This representation has a confusing aspect in that the direction of the particle's steps is parallel to the y-axis, whereas we have previously been specifying the movement in the traditional way as to the right or to the left. In future, any reference to the x-axis or the y-axis will pertain to a diagram of its path as exemplified by Figure 3.2.

The sequence (2) of partial sums has three important properties.

(3) Lemma. *The simple random walk is* spatially homogeneous; *that is*

$$\mathbb{P}(S_n = j \mid S_0 = a) = \mathbb{P}(S_n = j + b \mid S_0 = a + b).$$

Proof. Both sides equal $\mathbb{P}\left(\sum_1^n X_i = j - a\right)$. ■

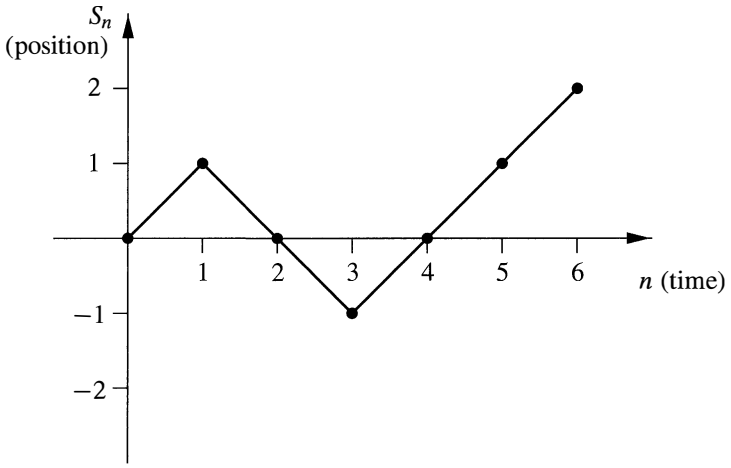

Figure 3.2. A random walk S_n.

(4) Lemma. *The simple random walk is* temporally homogeneous; *that is*

$$\mathbb{P}(S_n = j \mid S_0 = a) = \mathbb{P}(S_{m+n} = j \mid S_m = a).$$

Proof. The left- and right-hand sides satisfy

$$\text{LHS} = \mathbb{P}\left(\sum_1^n X_i = j - a\right) = \mathbb{P}\left(\sum_{m+1}^{m+n} X_i = j - a\right) = \text{RHS}. \qquad \blacksquare$$

(5) Lemma. *The simple random walk has the* Markov property; *that is*

$$\mathbb{P}(S_{m+n} = j \mid S_0, S_1, \ldots, S_m) = \mathbb{P}(S_{m+n} = j \mid S_m), \qquad n \geq 0.$$

Statements such as $\mathbb{P}(S = j \mid X, Y) = \mathbb{P}(S = j \mid X)$ are to be interpreted in the obvious way as meaning that

$$\mathbb{P}(S = j \mid X = x,\ Y = y) = \mathbb{P}(S = j \mid X = x) \quad \text{for all } x \text{ and } y;$$

this is a slight abuse of notation.

Proof. If one knows the value of S_m, then the distribution of S_{m+n} depends only on the jumps $X_{m+1}, \ldots, X_{m+n}$, and cannot depend on further information concerning the values of $S_0, S_1, \ldots, S_{m-1}$. $\blacksquare$

This 'Markov property' is often expressed informally by saying that, conditional upon knowing the value of the process at the mth step, its values after the mth step do not depend on its values before the mth step. More colloquially: conditional upon the present, the future does not depend on the past. We shall meet this property again later.

(6) Example. Absorbing barriers. Let us revisit Example (1.7.4) for general values of p. Equation (1.7.5) gives us the following difference equation for the probabilities $\{p_k\}$ where p_k is the probability of ultimate ruin starting from k:

$$p_k = p \cdot p_{k+1} + q \cdot p_{k-1} \quad \text{if} \quad 1 \le k \le N-1 \tag{7}$$

with boundary conditions $p_0 = 1$, $p_N = 0$. The solution of such a difference equation proceeds as follows. Look for a solution of the form $p_k = \theta^k$. Substitute this into (7) and cancel out the power θ^{k-1} to obtain $p\theta^2 - \theta + q = 0$, which has roots $\theta_1 = 1, \theta_2 = q/p$. If $p \neq \frac{1}{2}$ then these roots are distinct and the general solution of (7) is $p_k = A_1\theta_1^k + A_2\theta_2^k$ for arbitrary constants A_1 and A_2. Use the boundary conditions to obtain

$$p_k = \frac{(q/p)^k - (q/p)^N}{1 - (q/p)^N}.$$

If $p = \frac{1}{2}$ then $\theta_1 = \theta_2 = 1$ and the general solution to (7) is $p_k = A_1 + A_2 k$. Use the boundary conditions to obtain $p_k = 1 - (k/N)$.

A more complicated equation is obtained for the mean number D_k of steps before the particle hits one of the absorbing barriers, starting from k. In this case we use conditional expectations and (3.7.4) to find that

$$D_k = p(1 + D_{k+1}) + q(1 + D_{k-1}) \quad \text{if} \quad 1 \le k \le N-1 \tag{8}$$

with the boundary conditions $D_0 = D_N = 0$. Try solving this; you need to find a general solution and a particular solution, as in the solution of second-order linear differential equations. This answer is

$$D_k = \begin{cases} \dfrac{1}{q-p}\left[k - N\left(\dfrac{1-(q/p)^k}{1-(q/p)^N}\right)\right] & \text{if } p \neq \frac{1}{2}, \\ k(N-k) & \text{if } p = \frac{1}{2}. \end{cases} \tag{9}$$

●

(10) Example. Retaining barriers. In Example (1.7.4), suppose that the Jaguar buyer has a rich uncle who will guarantee all his losses. Then the random walk does not end when the particle hits zero, although it cannot visit a negative integer. Instead $\mathbb{P}(S_{n+1} = 0 \mid S_n = 0) = q$ and $\mathbb{P}(S_{n+1} = 1 \mid S_n = 0) = p$. The origin is said to have a 'retaining' barrier (sometimes called 'reflecting').

What now is the expected duration of the game? The mean duration F_k, starting from k, satisfies the same difference equation (8) as before but subject to different boundary conditions. We leave it as an *exercise* to show that the boundary conditions are $F_N = 0$, $pF_0 = 1 + pF_1$, and hence to find F_k. ●

In such examples the techniques of 'conditioning' are supremely useful. The idea is that in order to calculate a probability $\mathbb{P}(A)$ or expectation $\mathbb{E}(Y)$ we condition either on some partition of Ω (and use Lemma (1.4.4)) or on the outcome of some random variable (and use Theorem (3.7.4) or the forthcoming Theorem (4.6.5)). In this section this technique yielded the difference equations (7) and (8). In later sections the same idea will yield differential equations, integral equations, and functional equations, some of which can be solved.

Exercises for Section 3.9

1. Let T be the time which elapses before a simple random walk is absorbed at either of the absorbing barriers at 0 and N, having started at k where $0 \leq k \leq N$. Show that $\mathbb{P}(T < \infty) = 1$ and $\mathbb{E}(T^k) < \infty$ for all $k \geq 1$.

2. For simple random walk S with absorbing barriers at 0 and N, let W be the event that the particle is absorbed at 0 rather than at N, and let $p_k = \mathbb{P}(W \mid S_0 = k)$. Show that, if the particle starts at k where $0 < k < N$, the conditional probability that the first step is rightwards, given W, equals pp_{k+1}/p_k. Deduce that the mean duration J_k of the walk, conditional on W, satisfies the equation

$$pp_{k+1}J_{k+1} - p_kJ_k + (p_k - pp_{k+1})J_{k-1} = -p_k, \qquad \text{for } 0 < k < N.$$

Show that we may take as boundary condition $J_0 = 0$. Find J_k in the symmetric case, when $p = \frac{1}{2}$.

3. With the notation of Exercise (2), suppose further that at any step the particle may remain where it is with probability r where $p + q + r = 1$. Show that J_k satisfies

$$pp_{k+1}J_{k+1} - (1-r)p_kJ_k + qp_{k-1}J_{k-1} = -p_k$$

and that, when $\rho = q/p \neq 1$,

$$J_k = \frac{1}{p-q} \cdot \frac{1}{\rho^k - \rho^N} \left\{ k(\rho^k + \rho^N) - \frac{2N\rho^N(1-\rho^k)}{1-\rho^N} \right\}.$$

4. Problem of the points. A coin is tossed repeatedly, heads turning up with probability p on each toss. Player A wins the game if m heads appear before n tails have appeared, and player B wins otherwise. Let p_{mn} be the probability that A wins the game. Set up a difference equation for the p_{mn}. What are the boundary conditions?

5. Consider a simple random walk on the set $\{0, 1, 2, \ldots, N\}$ in which each step is to the right with probability p or to the left with probability $q = 1 - p$. Absorbing barriers are placed at 0 and N. Show that the number X of positive steps of the walk before absorption satisfies

$$\mathbb{E}(X) = \tfrac{1}{2}\{D_k - k + N(1 - p_k)\}$$

where D_k is the mean number of steps until absorption and p_k is the probability of absorption at 0.

6. (a) "Millionaires should always gamble, poor men never" [J. M. Keynes].
(b) "If I wanted to gamble, I would buy a casino" [P. Getty].
(c) "That the chance of gain is naturally overvalued, we may learn from the universal success of lotteries" [Adam Smith, 1776].
Discuss.

3.10 Random walk: counting sample paths

In the previous section, our principal technique was to condition on the first step of the walk and then solve the ensuing difference equation. Another primitive but useful technique is to count. Let $X_1, X_2, \ldots$ be independent variables, each taking the values -1 and 1 with probabilities $q = 1 - p$ and p, as before, and let

$$\textbf{(1)} \qquad S_n = a + \sum_{i=1}^{n} X_i$$

be the position of the corresponding random walker after n steps, having started at $S_0 = a$. The set of realizations of the walk is the set of vectors $\mathbf{s} = (s_0, s_1, \dots)$ with $s_0 = a$ and $s_{i+1} - s_i = \pm 1$, and any such vector may be thought of as a 'sample path' of the walk, drawn in the manner of Figure 3.2. The probability that the first n steps of the walk follow a given path $\mathbf{s} = (s_0, s_1, \dots, s_n)$ is $p^r q^l$ where r is the number of steps of s to the right and l is the number to the left†; that is to say, $r = |\{i : s_{i+1} - s_i = 1\}|$ and $l = |\{i : s_{i+1} - s_i = -1\}|$. Any event may be expressed in terms of an appropriate set of paths, and the probability of the event is the sum of the component probabilities. For example, $\mathbb{P}(S_n = b) = \sum_r M_n^r(a, b) p^r q^{n-r}$ where $M_n^r(a, b)$ is the number of paths $(s_0, s_1, \dots, s_n)$ with $s_0 = a$, $s_n = b$, and having exactly r rightward steps. It is easy to see that $r + l = n$, the total number of steps, and $r - l = b - a$, the aggregate rightward displacement, so that $r = \frac{1}{2}(n + b - a)$ and $l = \frac{1}{2}(n - b + a)$. Thus

$$\mathbf{(2)} \qquad \mathbb{P}(S_n = b) = \binom{n}{\frac{1}{2}(n+b-a)} p^{\frac{1}{2}(n+b-a)} q^{\frac{1}{2}(n-b+a)},$$

since there are exactly $\binom{n}{r}$ paths with length n having r rightward steps and $n - r$ leftward steps. Formula (2) is useful only if $\frac{1}{2}(n + b - a)$ is an integer lying in the range $0, 1, \dots, n$; otherwise, the probability in question equals 0.

Natural equations of interest for the walk include:

(a) when does the first visit of the random walk to a given point occur; and

(b) what is the furthest rightward point visited by the random walk by time n?

Such questions may be answered with the aid of certain elegant results and techniques for counting paths. The first of these is the 'reflection principle'. Here is some basic notation. As in Figure 3.2, we keep a record of the random walk S through its path $\{(n, S_n) : n \geq 0\}$.

Suppose we know that $S_0 = a$ and $S_n = b$. The random walk may or may not have visited the origin between times 0 and n. Let $N_n(a, b)$ be the number of possible paths from $(0, a)$ to (n, b), and let $N_n^0(a, b)$ be the number of such paths which contain some point $(k, 0)$ on the x-axis.

(3) Theorem. The reflection principle. *If $a, b > 0$ then $N_n^0(a, b) = N_n(-a, b)$.*

Proof. Each path from $(0, -a)$ to (n, b) intersects the x-axis at some earliest point $(k, 0)$. Reflect the segment of the path with $0 \leq x \leq k$ in the x-axis to obtain a path joining $(0, a)$ to (n, b) which intersects the x-axis (see Figure 3.3). This operation gives a one–one correspondence between the collections of such paths, and the theorem is proved. ■

We have, as before, a formula for $N_n(a, b)$.

(4) Lemma. $N_n(a, b) = \binom{n}{\frac{1}{2}(n+b-a)}$.

Proof. Choose a path from $(0, a)$ to (n, b) and let α and β be the numbers of positive and negative steps, respectively, in this path. Then $\alpha + \beta = n$ and $\alpha - \beta = b - a$, so that $\alpha = \frac{1}{2}(n + b - a)$. The number of such paths is the number of ways of picking α positive steps from the n available. That is

$$\mathbf{(5)} \qquad N_n(a, b) = \binom{n}{\alpha} = \binom{n}{\frac{1}{2}(n+b-a)}. \qquad ■$$

†The words 'right' and 'left' are to be interpreted as meaning in the positive and negative directions respectively, plotted along the y-axis as in Figure 3.2.

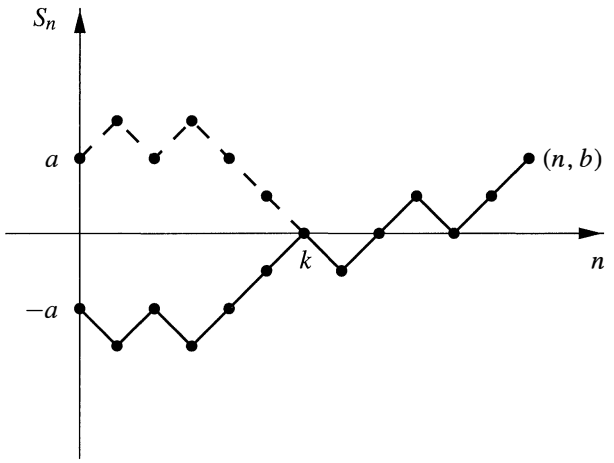

Figure 3.3. A random walk; the dashed line is the reflection of the first segment of the walk.

The famous 'ballot theorem' is a consequence of these elementary results; it was proved first by W. A. Whitworth in 1878.

(6) Corollary†. Ballot theorem. *If $b > 0$ then the number of paths from $(0, 0)$ to (n, b) which do not revisit the x-axis equals $(b/n)N_n(0, b)$.*

Proof. The first step of all such paths is to $(1, 1)$, and so the number of such path is

$$N_{n-1}(1, b) - N_{n-1}^0(1, b) = N_{n-1}(1, b) - N_{n-1}(-1, b)$$

by the reflection principle. We now use (4) and an elementary calculation to obtain the required result. ■

As an application, and an explanation of the title of the theorem, we may easily answer the following amusing question. Suppose that, in a ballot, candidate A scores α votes and candidate B scores β votes where $\alpha > \beta$. What is the probability that, during the ballot, A was always ahead of B? Let X_i equal 1 if the ith vote was cast for A, and -1 otherwise. Assuming that each possible combination of α votes for A and β votes for B is equally likely, we have that the probability is question is the proportion of paths from $(0, 0)$ to $(\alpha + \beta, \alpha - \beta)$ which do not revisit the x-axis. Using the ballot theorem, we obtain the answer $(\alpha - \beta)/(\alpha + \beta)$.

Here are some applications of the reflection principle to random walks. First, what is the probability that the walk does not revisit its starting point in the first n steps? We may as well assume that $S_0 = 0$, so that $S_1 \neq 0, \ldots, S_n \neq 0$ if and only if $S_1 S_2 \cdots S_n \neq 0$.

(7) Theorem. *If $S_0 = 0$ then, for $n \geq 1$,*

$$\mathbf{(8)} \qquad \mathbb{P}(S_1 S_2 \cdots S_n \neq 0,\ S_n = b) = \frac{|b|}{n}\mathbb{P}(S_n = b),$$

and therefore

$$\mathbf{(9)} \qquad \mathbb{P}(S_1 S_2 \cdots S_n \neq 0) = \frac{1}{n}\mathbb{E}|S_n|.$$

†Derived from the Latin word 'corollarium' meaning 'money paid for a garland' or 'tip'.

Proof. Suppose that $S_0 = 0$ and $S_n = b$ (> 0). The event in question occurs if and only if the path of the random walk does not visit the x-axis in the time interval $[1, n]$. The number of such paths is, by the ballot theorem, $(b/n)N_n(0, b)$, and each such path has $\frac{1}{2}(n+b)$ rightward steps and $\frac{1}{2}(n-b)$ leftward steps. Therefore

$$\mathbb{P}(S_1 S_2 \cdots S_n \neq 0,\ S_n = b) = \frac{b}{n} N_n(0, b) p^{\frac{1}{2}(n+b)} q^{\frac{1}{2}(n-b)} = \frac{b}{n} \mathbb{P}(S_n = b)$$

as required. A similar calculation is valid if $b < 0$. ■

Another feature of interest is the maximum value attained by the random walk. We write $M_n = \max\{S_i : 0 \leq i \leq n\}$ for the maximum value up to time n, and shall suppose that $S_0 = 0$, so that $M_n \geq 0$. Clearly $M_n \geq S_n$, and the first part of the next theorem is therefore trivial.

(10) Theorem. *Suppose that $S_0 = 0$. Then, for $r \geq 1$,*

$$\mathbb{P}(M_n \geq r,\ S_n = b) = \begin{cases} \mathbb{P}(S_n = b) & \text{if } b \geq r, \\ (q/p)^{r-b} \mathbb{P}(S_n = 2r - b) & \text{if } b < r. \end{cases} \tag{11}$$

It follows that, for $r \geq 1$,

$$\begin{aligned} \mathbb{P}(M_n \geq r) &= \mathbb{P}(S_n \geq r) + \sum_{b=-\infty}^{r-1} (q/p)^{r-b} \mathbb{P}(S_n = 2r - b) \\ &= \mathbb{P}(S_n = r) + \sum_{c=r+1}^{\infty} \left[1 + (q/p)^{c-r}\right] \mathbb{P}(S_n = c), \end{aligned} \tag{12}$$

yielding in the symmetric case when $p = q = \frac{1}{2}$ that

$$\mathbb{P}(M_n \geq r) = 2\mathbb{P}(S_n \geq r + 1) + \mathbb{P}(S_n = r), \tag{13}$$

which is easily expressed in terms of the binomial distribution.

Proof of (10). We may assume that $r \geq 1$ and $b < r$. Let $N_n^r(0, b)$ be the number of paths from $(0, 0)$ to (n, b) which include some point having height r, which is to say some point (i, r) with $0 < i < n$; for such a path π, let (i_π, r) be the earliest such point. We may reflect the segment of the path with $i_\pi \leq x \leq n$ in the line $y = r$ to obtain a path π' joining $(0, 0)$ to $(n, 2r - b)$. Any such path π' is obtained thus from a unique path π, and therefore $N_n^r(0, b) = N_n(0, 2r - b)$. It follows as required that

$$\begin{aligned} \mathbb{P}(M_n \geq r,\ S_n = b) &= N_n^r(0, b) p^{\frac{1}{2}(n+b)} q^{\frac{1}{2}(n-b)} \\ &= (q/p)^{r-b} N_n(0, 2r - b) p^{\frac{1}{2}(n+2r-b)} q^{\frac{1}{2}(n-2r+b)} \\ &= (q/p)^{r-b} \mathbb{P}(S_n = 2r - b). \end{aligned}$$

■

What is the chance that the walk reaches a new maximum at a particular time? More precisely, what is the probability that the walk, starting from 0, reaches the point b (> 0) for the time time at the nth step? Writing $f_b(n)$ for this probability, we have that

$$\begin{aligned} f_b(n) &= \mathbb{P}(M_{n-1} = S_{n-1} = b-1,\ S_n = b) \\ &= p\Big[\mathbb{P}(M_{n-1} \geq b-1,\ S_{n-1} = b-1) - \mathbb{P}(M_{n-1} \geq b,\ S_{n-1} = b-1)\Big] \\ &= p\big[\mathbb{P}(S_{n-1} = b-1) - (q/p)\mathbb{P}(S_{n-1} = b+1)\big] \quad \text{by (11)} \\ &= \frac{b}{n}\mathbb{P}(S_n = b) \end{aligned}$$

by a simple calculation using (2). A similar conclusion may be reached if $b < 0$, and we arrive at the following.

(14) Hitting time theorem. *The probability $f_b(n)$ that a random walk S hits the point b for the first time at the nth step, having started from 0, satisfies*

$$f_b(n) = \frac{|b|}{n}\mathbb{P}(S_n = b) \quad \text{if} \quad n \geq 1. \tag{15}$$

The conclusion here has a close resemblance to that of the ballot theorem, and particularly Theorem (7). This is no coincidence: a closer examination of the two results leads to another technique for random walks, the technique of 'reversal'. If the first n steps of the original random walk are

$$\{0, S_1, S_2, \ldots, S_n\} = \left\{0, X_1, X_1 + X_2, \ldots, \sum_1^n X_i\right\}$$

then the steps of the *reversed* walk, denoted by $0, T_1, \ldots, T_n$, are given by

$$\{0, T_1, T_2, \ldots, T_n\} = \left\{0, X_n, X_n + X_{n-1}, \ldots, \sum_1^n X_i\right\}.$$

Draw a diagram to see how the two walks correspond to each other. The X_i are independent and identically distributed, and it follows that the two walks have identical distributions even if $p \neq \frac{1}{2}$. Notice that the addition of an extra step to the original walk may change *every* step of the reversed walk.

Now, the original walk satisfies $S_n = b$ (> 0) and $S_1 S_2 \cdots S_n \neq 0$ if and only if the reversed walk satisfied $T_n = b$ and $T_n - T_{n-i} = X_1 + \cdots + X_i > 0$ for all $i \geq 1$, which is to say that the first visit of the reversed walk to the point b takes place at time n. Therefore

$$\mathbb{P}(S_1 S_2 \cdots S_n \neq 0,\ S_n = b) = f_b(n) \quad \text{if} \quad b > 0. \tag{16}$$

This is the 'coincidence' remarked above; a similar argument is valid if $b < 0$. The technique of reversal has other applications. For example, let μ_b be the mean number of visits of the walk to the point b before it returns to its starting point. If $S_0 = 0$ then, by (16),

$$\mu_b = \sum_{n=1}^{\infty} \mathbb{P}(S_1 S_2 \cdots S_n \neq 0,\ S_n = b) = \sum_{n=1}^{\infty} f_b(n) = \mathbb{P}(S_n = b \text{ for some } n), \tag{17}$$

the probability of ultimately visiting b. This leads to the following result.

(18) Theorem. *If $p = \frac{1}{2}$ and $S_0 = 0$, for any b ($\neq 0$) the mean number μ_b of visits of the walk to the point b before returning to the origin equals* 1.

Proof. Let $f_b = \mathbb{P}(S_n = b \text{ for some } n \geq 0)$. We have, by conditioning on the value of S_1, that $f_b = \frac{1}{2}(f_{b+1} + f_{b-1})$ for $b > 0$, with boundary condition $f_0 = 1$. The solution of this difference equation is $f_b = Ab + B$ for constants A and B. The unique such solution lying in $[0, 1]$ with $f_0 = 1$ is given by $f_b = 1$ for all $b \geq 0$. By symmetry, $f_b = 1$ for $b \leq 0$. However, $f_b = \mu_b$ for $b \neq 0$, and the claim follows. ■

'The truly amazing implications of this result appear best in the language of fair games. A perfect coin is tossed until the first equalization of the accumulated numbers of heads and tails. The gambler receives one penny for every time that the accumulated number of heads exceeds the accumulated number of tails by m. *The "fair entrance fee" equals* 1 *independently of* m.' (Feller 1968, p. 367).

We conclude with two celebrated properties of the symmetric random walk.

(19) Theorem. Arc sine law for last visit to the origin. *Suppose that $p = \frac{1}{2}$ and $S_0 = 0$. The probability that the last visit to* 0 *up to time* $2n$ *occurred at time* $2k$ *is* $\mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2n-2k} = 0)$.

In advance of proving this, we note some consequences. Writing $\alpha_{2n}(2k)$ for the probability referred to in the theorem, it follows from the theorem that $\alpha_{2n}(2k) = u_{2k}u_{2n-2k}$ where

$$u_{2k} = \mathbb{P}(S_{2k} = 0) = \binom{2k}{k} 2^{-2k}.$$

In order to understand the behaviour of u_{2k} for large values of k, we use Stirling's formula:

$$n! \sim n^n e^{-n}\sqrt{2\pi n} \quad \text{as} \quad n \to \infty, \tag{20}$$

which is to say that the ratio of the left-hand side to the right-hand side tends to 1 as $n \to \infty$. Applying this formula, we obtain that $u_{2k} \sim 1/\sqrt{\pi k}$ as $k \to \infty$. This gives rise to the approximation

$$\alpha_{2n}(2k) \simeq \frac{1}{\pi\sqrt{k(n-k)}},$$

valid for values of k which are close to neither 0 nor n. With T_{2n} denoting the time of the last visit to 0 up to time $2n$, it follows that

$$\mathbb{P}(T_{2n} \leq 2xn) \simeq \sum_{k \leq xn} \frac{1}{\pi\sqrt{k(n-k)}} \sim \int_{u=0}^{xn} \frac{1}{\pi\sqrt{u(n-u)}}\, du = \frac{2}{\pi} \sin^{-1}\sqrt{x},$$

which is to say that $T_{2n}/(2n)$ has a distribution function which is approximately $(2/\pi)\sin^{-1}\sqrt{x}$ when n is sufficiently large. We have proved a limit theorem.

The arc sine law is rather surprising. One may think that, in a long run of $2n$ tosses of a fair coin, the epochs of time at which there have appeared equal numbers of heads and tails should appear rather frequently. On the contrary, there is for example probability $\frac{1}{2}$ that no such epoch arrived in the final n tosses, and indeed probability approximately $\frac{1}{5}$ that no such epoch occurred after the first $\frac{1}{5}n$ tosses. One may think that, in a long run of $2n$ tosses of a

fair coin, the last time at which the numbers of heads and tails were equal tends to be close to the end. On the contrary, the distribution of this time is symmetric around the midpoint.

How much time does a symmetric random walk spend to the right of the origin? More precisely, for how many values of k satisfying $0 \le k \le 2n$ is it the case that $S_k > 0$? Intuitively, one might expect the answer to be around n with large probability, but the truth is quite different. With large probability, the proportion of time spent to the right (or to the left) of the origin is near to 0 or to 1, but not near to $\frac{1}{2}$. That is to say, in a long sequence of tosses of a fair coin, there is large probability that one face (either heads or tails) will lead the other for a disproportionate amount of time.

(21) Theorem. Arc sine law for sojourn times. *Suppose that $p = \frac{1}{2}$ and $S_0 = 0$. The probability that the walk spends exactly $2k$ intervals of time, up to time $2n$, to the right of the origin equals $\mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2n-2k} = 0)$.*

We say that the interval $(k, k+1)$ is spent to the right of the origin if either $S_k > 0$ or $S_{k+1} > 0$. It is clear that the number of such intervals is even if the total number of steps is even. The conclusion of this theorem is most striking. First, the answer is the same as that of Theorem (19). Secondly, by the calculations following (19) we have that the probability that the walk spends $2xn$ units of time or less to the right of the origin is approximately $(2/\pi)\sin^{-1}\sqrt{x}$.

Proof of (19). The probability in question is

$$\begin{aligned}\alpha_{2n}(2k) &= \mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2k+1}S_{2k+2}\cdots S_{2n} \neq 0 \mid S_{2k} = 0)\\ &= \mathbb{P}(S_{2k} = 0)\mathbb{P}(S_1 S_2 \cdots S_{2n-2k} \neq 0).\end{aligned}$$

Now, setting $m = n - k$, we have by (8) that

$$\begin{aligned}\textbf{(22)}\qquad \mathbb{P}(S_1 S_2 \cdots S_{2m} \neq 0) &= 2\sum_{k=1}^{m} \frac{2k}{2m}\mathbb{P}(S_{2m} = 2k) = 2\sum_{k=1}^{m} \frac{2k}{2m}\binom{2m}{m+k}\left(\frac{1}{2}\right)^{2m}\\ &= 2\left(\frac{1}{2}\right)^{2m}\sum_{k=1}^{m}\left[\binom{2m-1}{m+k-1} - \binom{2m-1}{m+k}\right]\\ &= 2\left(\frac{1}{2}\right)^{2m}\binom{2m-1}{m}\\ &= \binom{2m}{m}\left(\frac{1}{2}\right)^{2m} = \mathbb{P}(S_{2m} = 0).\end{aligned}$$

■

In passing, note the proof in (22) that

$$\textbf{(23)}\qquad \mathbb{P}(S_1 S_2 \cdots S_{2m} \neq 0) = \mathbb{P}(S_{2m} = 0)$$

for the simple symmetric random walk.

Proof of (21). Let $\beta_{2n}(2k)$ be the probability in question, and write $u_{2m} = \mathbb{P}(S_{2m} = 0)$ as before. We are claiming that, for all $m \ge 1$,

$$\textbf{(24)}\qquad \beta_{2m}(2k) = u_{2k}u_{2m-2k} \quad \text{if} \quad 0 \le k \le m.$$

First,

$$\begin{aligned}\mathbb{P}(S_1S_2\cdots S_{2m}>0) &= \mathbb{P}(S_1=1,\ S_2\geq 1,\ldots,\ S_{2m}\geq 1)\\ &= \tfrac{1}{2}\mathbb{P}(S_1\geq 0,\ S_2\geq 0,\ldots,\ S_{2m-1}\geq 0),\end{aligned}$$

where the second line follows by considering the walk $S_1-1, S_2-1,\ldots,S_{2m}-1$. Now S_{2m-1} is an odd number, so that $S_{2m-1}\geq 0$ implies that $S_{2m}\geq 0$ also. Thus

$$\mathbb{P}(S_1S_2\cdots S_{2m}>0)=\tfrac{1}{2}\mathbb{P}(S_1\geq 0, S_2\geq 0,\ldots,S_{2m}\geq 0),$$

yielding by (23) that

$$\tfrac{1}{2}u_{2m}=\mathbb{P}(S_1S_2\cdots S_{2m}>0)=\tfrac{1}{2}\beta_{2m}(2m),$$

and (24) follows for $k=m$, and therefore for $k=0$ also by symmetry.

Let n be a positive integer, and let T be the time of the first return of the walk to the origin. If $S_{2n}=0$ then $T\leq 2n$; the probability mass function $f_{2r}=\mathbb{P}(T=2r)$ satisfies

$$\mathbb{P}(S_{2n}=0)=\sum_{r=1}^{n}\mathbb{P}(S_{2n}=0\mid T=2r)\mathbb{P}(T=2r)=\sum_{r=1}^{n}\mathbb{P}(S_{2n-2r}=0)\mathbb{P}(T=2r),$$

which is to say that

$$u_{2n}=\sum_{r=1}^{n}u_{2n-2r}f_{2r}. \tag{25}$$

Let $1\leq k\leq n-1$, and consider $\beta_{2n}(2k)$. The corresponding event entails that $T=2r$ for some r satisfying $1\leq r<n$. The time interval $(0,T)$ is spent entirely either to the right or the left of the origin, and each possibility has probability $\frac{1}{2}$. Therefore,

$$\beta_{2n}(2k)=\sum_{r=1}^{k}\tfrac{1}{2}\mathbb{P}(T=2r)\beta_{2n-2r}(2k-2r)+\sum_{r=1}^{n-k}\tfrac{1}{2}\mathbb{P}(T=2r)\beta_{2n-2r}(2k). \tag{26}$$

We conclude the proof by using induction. Certainly (24) is valid for all k if $m=1$. Assume (24) is valid for all k and all $m<n$.

From (26),

$$\begin{aligned}\beta_{2n}(2k) &= \tfrac{1}{2}\sum_{r=1}^{k}f_{2r}u_{2k-2r}u_{2n-2k}+\tfrac{1}{2}\sum_{r=1}^{n-k}f_{2r}u_{2k}u_{2n-2k-2r}\\ &= \tfrac{1}{2}u_{2n-2k}u_{2k}+\tfrac{1}{2}u_{2k}u_{2n-2k}=u_{2k}u_{2n-2k}\end{aligned}$$

by (25), as required. ■

Exercises for Section 3.10

1. Consider a symmetric simple random walk S with $S_0 = 0$. Let $T = \min\{n \geq 1 : S_n = 0\}$ be the time of the first return of the walk to its starting point. Show that

$$\mathbb{P}(T = 2n) = \frac{1}{2n-1}\binom{2n}{n}2^{-2n},$$

and deduce that $\mathbb{E}(T^\alpha) < \infty$ if and only if $\alpha < \frac{1}{2}$. You may need Stirling's formula: $n! \sim n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}$.

2. For a symmetric simple random walk starting at 0, show that the mass function of the maximum satisfies $\mathbb{P}(M_n = r) = \mathbb{P}(S_n = r) + \mathbb{P}(S_n = r+1)$ for $r \geq 0$.

3. For a symmetric simple random walk starting at 0, show that the probability that the first visit to S_{2n} takes place at time $2k$ equals the product $\mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2n-2k} = 0)$, for $0 \leq k \leq n$.

3.11 Problems

1. (a) Let X and Y be independent discrete random variables, and let $g, h : \mathbb{R} \to \mathbb{R}$. Show that $g(X)$ and $h(Y)$ are independent.

(b) Show that two discrete random variables X and Y are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.

(c) More generally, show that X and Y are independent if and only if $f_{X,Y}(x, y)$ can be factorized as the product $g(x)h(y)$ of a function of x alone and a function of y alone.

2. Show that if $\text{var}(X) = 0$ then X is almost surely constant; that is, there exists $a \in \mathbb{R}$ such that $\mathbb{P}(X = a) = 1$. (First show that if $\mathbb{E}(X^2) = 0$ then $\mathbb{P}(X = 0) = 1$.)

3. (a) Let X be a discrete random variable and let $g : \mathbb{R} \to \mathbb{R}$. Show that, when the sum is absolutely convergent,

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

(b) If X and Y are independent and $g, h : \mathbb{R} \to \mathbb{R}$, show that $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$ whenever these expectations exist.

4. Let $\Omega = \{\omega_1, \omega_2, \omega_3\}$, with $\mathbb{P}(\omega_1) = \mathbb{P}(\omega_2) = \mathbb{P}(\omega_3) = \frac{1}{3}$. Define $X, Y, Z : \Omega \to \mathbb{R}$ by

$$\begin{aligned} X(\omega_1) &= 1, \quad X(\omega_2) = 2, \quad X(\omega_3) = 3, \\ Y(\omega_1) &= 2, \quad Y(\omega_2) = 3, \quad Y(\omega_3) = 1, \\ Z(\omega_1) &= 2, \quad Z(\omega_2) = 2, \quad Z(\omega_3) = 1. \end{aligned}$$

Show that X and Y have the same mass functions. Find the mass functions of $X + Y$, XY, and X/Y. Find the conditional mass functions $f_{Y|Z}$ and $f_{Z|Y}$.

5. For what values of k and α is f a mass function, where:

(a) $f(n) = k/\{n(n+1)\}$, $n = 1, 2, \ldots$,

(b) $f(n) = kn^\alpha$, $n = 1, 2, \ldots$ (*zeta* or *Zipf distribution*)?

6. Let X and Y be independent Poisson variables with respective parameters λ and μ. Show that:
(a) $X + Y$ is Poisson, parameter $\lambda + \mu$,
(b) the conditional distribution of X, given $X + Y = n$, is binomial, and find its parameters.

7. If X is geometric, show that $\mathbb{P}(X = n + k \mid X > n) = \mathbb{P}(X = k)$ for $k, n \geq 1$. Why do you think that this is called the 'lack of memory' property? Does any other distribution on the positive integers have this property?

8. Show that the sum of two independent binomial variables, bin(m, p) and bin(n, p) respectively, is bin$(m + n, p)$.

9. Let N be the number of heads occurring in n tosses of a biased coin. Write down the mass function of N in terms of the probability p of heads turning up on each toss. Prove and utilize the identity

$$\sum_i \binom{n}{2i} x^{2i} y^{n-2i} = \tfrac{1}{2}\{(x+y)^n + (y-x)^n\}$$

in order to calculate the probability p_n that N is even. Compare with Problem (1.8.20).

10. An urn contains N balls, b of which are blue and r $(= N - b)$ of which are red. A random sample of n balls is withdrawn without replacement from the urn. Show that the number B of blue balls in this sample has the mass function

$$\mathbb{P}(B = k) = \binom{b}{k}\binom{N-b}{n-k} \Big/ \binom{N}{n}.$$

This is called the *hypergeometric distribution* with parameters N, b, and n. Show further that if N, b, and r approach ∞ in such a way that $b/N \to p$ and $r/N \to 1 - p$, then

$$\mathbb{P}(B = k) \to \binom{n}{k} p^k (1-p)^{n-k}.$$

You have shown that, for small n and large N, the distribution of B barely depends on whether or not the balls are replaced in the urn immediately after their withdrawal.

11. Let X and Y be independent bin(n, p) variables, and let $Z = X + Y$. Show that the conditional distribution of X given $Z = N$ is the hypergeometric distribution of Problem (3.11.10).

12. Suppose X and Y take values in $\{0, 1\}$, with joint mass function $f(x, y)$. Write $f(0,0) = a$, $f(0,1) = b$, $f(1,0) = c$, $f(1,1) = d$, and find necessary and sufficient conditions for X and Y to be: (a) uncorrelated, (b) independent.

13. (a) If X takes non-negative integer values show that

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} \mathbb{P}(X > n).$$

(b) An urn contains b blue and r red balls. Balls are removed at random until the first blue ball is drawn. Show that the expected number drawn is $(b + r + 1)/(b + 1)$.
(c) The balls are replaced and then removed at random until all the remaining balls are of the same colour. Find the expected number remaining in the urn.

14. Let $X_1, X_2, \ldots, X_n$ be independent random variables, and suppose that X_k is Bernoulli with parameter p_k. Show that $Y = X_1 + X_2 + \cdots + X_n$ has mean and variance given by

$$\mathbb{E}(Y) = \sum_1^n p_k, \quad \text{var}(Y) = \sum_1^n p_k(1 - p_k).$$

Show that, for $\mathbb{E}(Y)$ fixed, $\text{var}(Y)$ is a maximum when $p_1 = p_2 = \cdots = p_n$. That is to say, the variation in the sum is greatest when individuals are most alike. Is this contrary to intuition?

15. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a vector of random variables. The *covariance matrix* $\mathbf{V}(\mathbf{X})$ of $\mathbf{X}$ is defined to be the symmetric n by n matrix with entries $(v_{ij} : 1 \le i, j \le n)$ given by $v_{ij} = \text{cov}(X_i, X_j)$. Show that $|\mathbf{V}(\mathbf{X})| = 0$ if and only if the X_i are linearly dependent with probability one, in that $\mathbb{P}(a_1X_1 + a_2X_2 + \cdots + a_nX_n = b) = 1$ for some $\mathbf{a}$ and b. ($|\mathbf{V}|$ denotes the determinant of $\mathbf{V}$.)

16. Let X and Y be independent Bernoulli random variables with parameter $\frac{1}{2}$. Show that $X + Y$ and $|X - Y|$ are dependent though uncorrelated.

17. A secretary drops n matching pairs of letters and envelopes down the stairs, and then places the letters into the envelopes in a random order. Use indicators to show that the number X of correctly matched pairs has mean and variance 1 for all $n \ge 2$. Show that the mass function of X converges to a Poisson mass function as $n \to \infty$.

18. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a vector of independent random variables each having the Bernoulli distribution with parameter p. Let $f : \{0, 1\}^n \to \mathbb{R}$ be *increasing*, which is to say that $f(\mathbf{x}) \le f(\mathbf{y})$ whenever $x_i \le y_i$ for each i.

(a) Let $e(p) = \mathbb{E}(f(\mathbf{X}))$. Show that $e(p_1) \le e(p_2)$ if $p_1 \le p_2$.

(b) **FKG inequality**†. Let f and g be increasing functions from $\{0, 1\}^n$ into $\mathbb{R}$. Show by induction on n that $\text{cov}(f(\mathbf{X}), g(\mathbf{X})) \ge 0$.

19. Let $R(p)$ be the reliability function of a network G with a given source and sink, each edge of which is working with probability p, and let A be the event that there exists a working connection from source to sink. Show that

$$R(p) = \sum_\omega I_A(\omega) p^{N(\omega)} (1 - p)^{m - N(\omega)}$$

where ω is a typical realization (i.e., outcome) of the network, $N(\omega)$ is the number of working edges of ω, and m is the total number of edges of G.

Deduce that $R'(p) = \text{cov}(I_A, N)/\{p(1 - p)\}$, and hence that

$$\frac{R(p)(1 - R(p))}{p(1 - p)} \le R'(p) \le \sqrt{\frac{mR(p)(1 - R(p))}{p(1 - p)}}.$$

20. Let $R(p)$ be the reliability function of a network G, each edge of which is working with probability p.

(a) Show that $R(p_1 p_2) \le R(p_1)R(p_2)$ if $0 \le p_1, p_2 \le 1$.

(b) Show that $R(p^\gamma) \le R(p)^\gamma$ for all $0 \le p \le 1$ and $\gamma \ge 1$.

21. DNA fingerprinting. In a certain style of detective fiction, the sleuth is required to declare "the criminal has the unusual characteristics ...; find this person and you have your man". Assume that any given individual has these unusual characteristics with probability 10^{-7} independently of all other individuals, and that the city in question contains 10^7 inhabitants. Calculate the expected number of such people in the city.

---

†Named after C. Fortuin, P. Kasteleyn, and J. Ginibre (1971), but due in this form to T. E. Harris (1960).

(a) Given that the police inspector finds such a person, what is the probability that there is at least one other?
(b) If the inspector finds two such people, what is the probability that there is at least one more?
(c) How many such people need be found before the inspector can be reasonably confident that he has found them all?
(d) For the given population, how improbable should the characteristics of the criminal be, in order that he (or she) be specified uniquely?

22. In 1710, J. Arbuthnot observed that male births had exceeded female births in London for 82 successive years. Arguing that the two sexes are equally likely, and 2^{-82} is very small, he attributed this run of masculinity to Divine Providence. Let us assume that each birth results in a girl with probability $p = 0.485$, and that the outcomes of different confinements are independent of each other. Ignoring the possibility of twins (and so on), show that the probability that girls outnumber boys in $2n$ live births is no greater than $\binom{2n}{n} p^n q^n \{q/(q-p)\}$, where $q = 1 - p$. Suppose that 20,000 children are born in each of 82 successive years. Show that the probability that boys outnumber girls every year is at least 0.99. You may need Stirling's formula.

23. Consider a symmetric random walk with an absorbing barrier at N and a reflecting barrier at 0 (so that, when the particle is at 0, it moves to 1 at the next step). Let $\alpha_k(j)$ be the probability that the particle, having started at k, visits 0 exactly j times before being absorbed at N. We make the convention that, if $k = 0$, then the starting point counts as one visit. Show that

$$\alpha_k(j) = \frac{N-k}{N^2}\left(1 - \frac{1}{N}\right)^{j-1}, \qquad j \geq 1,\ 0 \leq k \leq N.$$

24. Problem of the points (3.9.4). A coin is tossed repeatedly, heads turning up with probability p on each toss. Player A wins the game if heads appears at least m times before tails has appeared n times; otherwise player B wins the game. Find the probability that A wins the game.

25. A coin is tossed repeatedly, heads appearing on each toss with probability p. A gambler starts with initial fortune k (where $0 < k < N$); he wins one point for each head and loses one point for each tail. If his fortune is ever 0 he is bankrupted, whilst if it ever reaches N he stops gambling to buy a Jaguar. Suppose that $p < \frac{1}{2}$. Show that the gambler can increase his chance of winning by doubling the stakes. You may assume that k and N are even.

What is the corresponding strategy if $p \geq \frac{1}{2}$?

26. A compulsive gambler is never satisfied. At each stage he wins £1 with probability p and loses £1 otherwise. Find the probability that he is ultimately bankrupted, having started with an initial fortune of £k.

27. Range of random walk. Let $\{X_n : n \geq 1\}$ be independent, identically distributed random variables taking integer values. Let $S_0 = 0$, $S_n = \sum_{i=1}^{n} X_i$. The *range* R_n of $S_0, S_1, \ldots, S_n$ is the number of distinct values taken by the sequence. Show that $\mathbb{P}(R_n = R_{n-1} + 1) = \mathbb{P}(S_1 S_2 \cdots S_n \neq 0)$, and deduce that, as $n \to \infty$,

$$\frac{1}{n}\mathbb{E}(R_n) \to \mathbb{P}(S_k \neq 0 \text{ for all } k \geq 1).$$

Hence show that, for the simple random walk, $n^{-1}\mathbb{E}(R_n) \to |p - q|$ as $n \to \infty$.

28. Arc sine law for maxima. Consider a symmetric random walk S starting from the origin, and let $M_n = \max\{S_i : 0 \leq i \leq n\}$. Show that, for $i = 2k, 2k+1$, the probability that the walk reaches M_{2n} for the first time at time i equals $\frac{1}{2}\mathbb{P}(S_{2k} = 0)\mathbb{P}(S_{2n-2k} = 0)$.

29. Let S be a symmetric random walk with $S_0 = 0$, and let N_n be the number of points that have been visited by S exactly once up to time n. Show that $\mathbb{E}(N_n) = 2$.

30. Family planning. Consider the following fragment of verse entitled 'Note for the scientist'.

People who have three daughters try for more,
And then its fifty–fifty they'll have four,
Those with a son or sons will let things be,
Hence all these surplus women, QED.

(a) What do you think of the argument?

(b) Show that the mean number of children of either sex in a family whose fertile parents have followed this policy equals 1. (You should assume that each delivery yields exactly one child whose sex is equally likely to be male or female.) Discuss.

31. Let $\beta > 1$, let $p_1, p_2, \ldots$ denote the prime numbers, and let $N(1), N(2), \ldots$ be independent random variables, $N(i)$ having mass function $\mathbb{P}(N(i) = k) = (1 - \gamma_i)\gamma_i^k$ for $k \geq 0$, where $\gamma_i = p_i^{-\beta}$ for all i. Show that $M = \prod_{i=1}^{\infty} p_i^{N(i)}$ is a random integer with mass function $\mathbb{P}(M = m) = Cm^{-\beta}$ for $m \geq 1$ (this may be called the *Dirichlet distribution*), where C is a constant satisfying

$$C = \prod_{i=1}^{\infty} \left(1 - \frac{1}{p_i^{\beta}}\right) = \left(\sum_{m=1}^{\infty} \frac{1}{m^{\beta}}\right)^{-1}.$$

32. $N + 1$ plates are laid out around a circular dining table, and a hot cake is passed between them in the manner of a symmetric random walk: each time it arrives on a plate, it is tossed to one of the two neighbouring plates, each possibility having probability $\frac{1}{2}$. The game stops at the moment when the cake has visited every plate at least once. Show that, with the exception of the plate where the cake began, each plate has probability $1/N$ of being the last plate visited by the cake.

33. Simplex algorithm. There are $\binom{n}{m}$ points ranked in order of merit with no matches. You seek to reach the best, B. If you are at the jth best, you step to any one of the $j - 1$ better points, with equal probability of stepping to each. Let r_j be the expected number of steps to reach B from the jth best vertex. Show that $r_j = \sum_{k=1}^{j-1} k^{-1}$. Give an asymptotic expression for the expected time to reach B from the worst vertex, for large m, n.

34. Dimer problem. There are n unstable molecules in a row, $m_1, m_2, \ldots, m_n$. One of the $n - 1$ pairs of neighbours, chosen at random, combines to form a stable dimer; this process continues until there remain U_n isolated molecules no two of which are adjacent. Show that the probability that m_1 remains isolated is $\sum_{r=0}^{n-1}(-1)^r/r! \to e^{-1}$ as $n \to \infty$. Deduce that $\lim_{n\to\infty} n^{-1}\mathbb{E}U_n = e^{-2}$.

35. Poisson approximation. Let $\{I_r : 1 \leq r \leq n\}$ be independent Bernoulli random variables with respective parameters $\{p_r : 1 \leq r \leq n\}$ satisfying $p_r \leq c < 1$ for all r and some c. Let $\lambda = \sum_{r=1}^{n} p_r$ and $X = \sum_{r=1}^{n} X_r$. Show that

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}\left\{1 + \mathrm{O}\left(\lambda \max_r p_r + \frac{k^2}{\lambda} \max p_r\right)\right\}.$$

36. Sampling. The length of the tail of the rth member of a troop of N chimeras is x_r. A random sample of n chimeras is taken (without replacement) and their tails measured. Let I_r be the indicator of the event that the rth chimera is in the sample. Set

$$X_r = x_r I_r, \quad \overline{Y} = \frac{1}{n}\sum_{r=1}^{N} X_r, \quad \overline{x} = \frac{1}{N}\sum_{r=1}^{N} x_r, \quad \sigma^2 = \frac{1}{N}\sum_{r=1}^{N}(x_r - \overline{x})^2.$$

Show that $\mathbb{E}(\overline{Y}) = \mu$, and $\mathrm{var}(\overline{Y}) = (N - n)\sigma^2/\{n(N - 1)\}$.

37. Berkson's fallacy. Any individual in a group G contracts a certain disease C with probability γ; such individuals are hospitalized with probability c. Independently of this, anyone in G may be in hospital with probability a, for some other reason. Let X be the number in hospital, and Y the number in hospital who have C (including those with C admitted for any other reason). Show that the correlation between X and Y is

$$\rho(X, Y) = \sqrt{\frac{\gamma p}{1-\gamma p} \cdot \frac{(1-a)(1-\gamma c)}{a+\gamma c - a\gamma c}},$$

where $p = a + c - ac$.

It has been stated erroneously that, when $\rho(X, Y)$ is near unity, this is evidence for a causal relation between being in G and contracting C.

38. A telephone sales company attempts repeatedly to sell new kitchens to each of the N families in a village. Family i agrees to buy a new kitchen after it has been solicited K_i times, where the K_i are independent identically distributed random variables with mass function $f(n) = \mathbb{P}(K_i = n)$. The value ∞ is allowed, so that $f(\infty) \geq 0$. Let X_n be the number of kitchens sold at the nth round of solicitations, so that $X_n = \sum_{i=1}^{N} I_{\{K_i = n\}}$. Suppose that N is a random variable with the Poisson distribution with parameter ν.

(a) Show that the X_n are independent random variables, X_r having the Poisson distribution with parameter $\nu f(r)$.

(b) The company loses heart after the Tth round of calls, where $T = \inf\{n : X_n = 0\}$. Let $S = X_1 + X_2 + \cdots + X_T$ be the number of solicitations made up to time T. Show further that $\mathbb{E}(S) = \nu\mathbb{E}(F(T))$ where $F(k) = f(1) + f(2) + \cdots + f(k)$.

39. A particle performs a random walk on the non-negative integers as follows. When at the point n (> 0) its next position is uniformly distributed on the set $\{0, 1, 2, \ldots, n+1\}$. When it hits 0 for the first time, it is absorbed. Suppose it starts at the point a.

(a) Find the probability that its position never exceeds a, and prove that, with probability 1, it is absorbed ultimately.

(b) Find the probability that the final step of the walk is from 1 to 0 when $a = 1$.

(c) Find the expected number of steps taken before absorption when $a = 1$.

40. Let G be a finite graph with neither loops nor multiple edges, and write d_v for the degree of the vertex v. An *independent set* is a set of vertices no pair of which is joined by an edge. Let $\alpha(G)$ be the size of the largest independent set of G. Use the probabilistic method to show that $\alpha(G) \geq \sum_v 1/(d_v + 1)$. [This conclusion is sometimes referred to as *Turán's theorem.*]